Knowledge discovery from patient forums: gaining novel medical insights from patient experiences

Dirkson, A.R.

KNOWLEDGE DISCOVERY FROM PATIENT FORUMS

GAINING NOVEL MEDICAL INSIGHTS FROM PATIENT EXPERIENCES

KNOWLEDGE DISCOVERY FROM PATIENT FORUMS

GAINING NOVEL MEDICAL INSIGHTS FROM PATIENT EXPERIENCES

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 6 december 2022
klokke 11:15 uur

door

Anne Dirkson

geboren te Gouda, Nederland
in 1993

Promotores:
Prof. dr. ir. W. Kraaij
Prof. dr. A.J. Gelderblom

Co-promotor:
Dr. S. Verberne

Promotiecommissie
Prof. dr. A. Plaat (Leiden Universiteit)
Prof. dr. M. Spruit (Leiden Universiteit)
Prof. dr. K. Verspoor (RMIT University)
Prof. dr. M. Hoogendoorn (Vrije Universiteit)
Prof. dr. L. van de Poll (Tilburg University)

Printed by: Ridderprint

Front & Back: Gaby de Jong-Verwegen

This thesis was the winner of the 2022 Krijn Rietveld Memorial Innovation Award

This publication was supported by the SIDN fonds

SIKS Dissertation Series No. 2022-26
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

CONTENTS

1 Introduction **1**
- 1.1 Patient forums as a source of collective experiential knowledge . . . 4
- 1.2 Using text mining to harvest collective experiential knowledge . . . 5
- 1.3 Potential for rare diseases . . . 6
- 1.4 Previous work focused on pharmacovigilance . . . 7
- 1.5 Methodological tasks and challenges . . . 8
  - 1.5.1 ADE extraction . . . 9
  - 1.5.2 ADE normalization . . . 11
  - 1.5.3 Determining ADE–drug relations . . . 12
- 1.6 Research questions and thesis outline . . . 13
  - 1.6.1 Part 1: Reducing noise in medical social media text . . . 14
  - 1.6.2 Part 2: Extracting adverse drug events . . . 15
  - 1.6.3 Part 3: Extracting recommended coping strategies . . . 15
  - 1.6.4 Part 4: Complementary value of disease-specific patient forums as a source of patient-reported outcomes . . . 16
- 1.7 Scientific contributions . . . 16
- 1.8 Guide for the reader . . . 19

I Preprocessing medical social media text **21**

2 Data-driven spelling correction **23**
- 2.1 Introduction . . . 24
- 2.2 Related work . . . 25
  - 2.2.1 Challenges in correcting spelling errors in medical social media . . . 25
  - 2.2.2 Lexical normalization of social media . . . 26
  - 2.2.3 Lexical normalization of clinical records . . . 27
- 2.3 Materials and methods . . . 28
  - 2.3.1 Data . . . 28
  - 2.3.2 Methods . . . 29
- 2.4 Results . . . 35
  - 2.4.1 Error distribution . . . 35
  - 2.4.2 Spelling correction . . . 35
  - 2.4.3 Spelling mistake detection . . . 39
  - 2.4.4 Impact of corpus size . . . 41
  - 2.4.5 Impact of the degree of noisiness of the data . . . 42
  - 2.4.6 Effect on OOV rate . . . 45
  - 2.4.7 External validation . . . 46
- 2.5 Discussion . . . 48
- 2.6 Conclusions and future work . . . 50

3 Detecting personal experiences **51**
3.1 Introduction . . . 52
3.2 Related work . . . 52
3.3 Methods . . . 53
3.3.1 Data . . . 53
3.3.2 Preprocessing . . . 53
3.3.3 Supervised classification . . . 53
3.3.4 Topic modeling of the whole data set . . . 54
3.4 Results . . . 55
3.4.1 Annotated data . . . 55
3.4.2 Classifier evaluation . . . 55
3.4.3 Influential features . . . 55
3.4.4 Error analysis for the best performing classifier . . . 56
3.4.5 Frequency and content of patient narratives . . . 57
3.5 Discussion . . . 58
3.6 Conclusion . . . 60

4 Conversation-aware filtering of relevant messages **61**
4.1 Introduction . . . 62
4.2 Related work . . . 63
4.3 Methods . . . 63
4.3.1 Models . . . 63
4.3.2 Feature analysis . . . 64
4.3.3 Model comparison . . . 64
4.4 Data . . . 64
4.5 Results . . . 66
4.5.1 Distribution of the target class in the discussion threads . . . 66
4.5.2 Model comparison . . . 66
4.5.3 Analysis of selected features . . . 68
4.6 Discussion . . . 68

II Extracting adverse drug effects (ADEs) **71**

5 Transfer learning for ADE extraction from Twitter **73**
5.1 Introduction . . . 74
5.2 Task descriptions . . . 74
5.3 Our approach . . . 74
5.3.1 Preprocessing . . . 74
5.3.2 Additional Data . . . 74
5.3.3 Text Classification (S1 and S4) . . . 75
5.3.4 Named Entity Recognition (S2) . . . 75
5.3.5 Concept normalization (S3) . . . 76
5.4 Results . . . 76
5.5 Conclusions . . . 78

6 Vulnerabilities of BERT for Named Entity Recognition 79
6.1 Introduction . . . 80
6.2 Related work . . . 81
6.3 Methods . . . 82
6.3.1 Aim of the attacks . . . 82
6.4 Experiments . . . 84
6.4.1 Data . . . 84
6.4.2 Target models . . . 85
6.4.3 Evaluation of adversarial attacks . . . 86
6.5 Results . . . 86
6.5.1 Entity Context Attack . . . 86
6.5.2 Evaluating the necessity of importance ranking . . . 87
6.5.3 Entity Attack . . . 87
6.5.4 Results of human evaluation . . . 88
6.6 Discussion and limitations . . . 89
6.7 Conclusions. . . . 90

7 Fuzzy representation of discontinuous entities 91
7.1 Introduction . . . 92
7.2 Related Work . . . 93
7.3 Methods . . . 94
7.3.1 The FuzzyBIO representation scheme . . . 94
7.3.2 Named entity recognition of ADEs . . . 94
7.3.3 Concept normalization of ADEs . . . 94
7.3.4 Evaluation . . . 94
7.4 Data. . . . 95
7.5 Results . . . 95
7.5.1 Intrinsic evaluation . . . 95
7.5.2 Extrinsic evaluation . . . 96
7.6 Discussion . . . 96
7.7 Conclusion . . . 97

III Extracting coping strategies for adverse drug effects 99

8 The discovery of recommended coping mechanisms 101
8.1 Related work . . . 103
8.2 Data . . . 105
8.2.1 Data collection . . . 105
8.2.2 Data annotation . . . 106
8.2.3 Coping Strategy Ontology . . . 108
8.2.4 Adding negative examples . . . 109
8.3 Methods . . . 111
8.3.1 Data preprocessing . . . 111
8.3.2 ADE extraction and data selection . . . 111
8.3.3 Coping Strategy extraction . . . 112
8.3.4 Negation detection. . . . 113

8.3.5 Relation extraction . . . 114
8.3.6 Data post-processing . . . 114
8.4 Results . . . 114
8.4.1 Data description . . . 114
8.4.2 Named entity recognition . . . 115
8.4.3 End-to-end extraction . . . 116
8.4.4 Case study on GIST ADE coping . . . 117
8.5 Discussion . . . 121
8.5.1 Comparison of approaches . . . 121
8.5.2 Relevance of our findings . . . 121
8.5.3 Potential application settings . . . 124
8.5.4 Limitations. . . . 125
8.5.5 Future work . . . 125
8.6 Conclusion . . . 126

IV GIST as a case study **127**

9 Patient forums as a complementary data source **129**
9.1 Introduction . . . 130
9.2 Materials and methods . . . 131
9.2.1 Data collection. . . . 131
9.2.2 Machine learning pipeline . . . 131
9.2.3 Data analysis. . . . 133
9.3 Results . . . 134
9.4 Discussion . . . 138
9.5 Conclusion . . . 140

10 Comparing questionnaire and forum data **141**
10.1 Introduction . . . 143
10.2 Methods . . . 144
10.2.1 Study design and participants . . . 144
10.2.2 Recruitment and data collection . . . 144
10.2.3 Study measures . . . 145
10.2.4 Statistical analysis . . . 145
10.3 Results . . . 146
10.3.1 Participants . . . 146
10.3.2 Prevalence scores . . . 146
10.3.3 Relation between questionnaire and forum symptoms . . . 146
10.4 Discussion . . . 149

11 Assessing sample bias **153**
11.1 Introduction . . . 154
11.2 Methods . . . 156
11.2.1 Study design & participants . . . 156
11.2.2 Survey . . . 156
11.2.3 Data analysis. . . . 158

11.3 Results . . . 158
11.3.1 Participants . . . 158
11.3.2 Social media usage . . . 160
11.3.3 Reasons for abstaining from online communication with peers . . . 160
11.3.4 Reasons for engaging with patient forums . . . 161
11.3.5 Characteristics of the patient forum users . . . 161
11.4 Discussion . . . 165
11.4.1 Summary of findings . . . 165
11.4.2 Comparison with existing literature . . . 165
11.4.3 Limitations . . . 166
11.4.4 Future work and recommendations . . . 167
11.5 Conclusion . . . 168

V Discussion **169**

12 Discussion **171**
12.1 Main findings . . . 171
12.2 Answer to main research question . . . 181
12.3 Future research . . . 182
12.3.1 Mining experiential knowledge from social media . . . 182
12.3.2 Ontology mapping and interoperability . . . 183
12.3.3 Dealing with real-world data . . . 185
12.4 Recommendations . . . 187
12.4.1 Knowledge discovery from social media . . . 187
12.4.2 Privacy and adopting FAIR metadata standards . . . 188
12.4.3 Developing annotation guidelines . . . 189
12.4.4 Long-term integration into healthcare . . . 191

References **193**

Summary **229**

Samenvatting **231**

Acknowledgements **233**

Appendices **235**

A Technical details of ADE extraction **237**
A.0.1 Data preprocessing . . . 237
A.0.2 Extracting ADEs from text . . . 238
A.0.3 ADE normalization . . . 239
A.0.4 Linking ADEs to medication . . . 240

B Supplementary files for Chapter 8 **249**

C Supplementary Tables for Chapter 10 **261**

Curriculum Vitæ **267**

List of Publications **269**

SIKS Dissertation Series **271**

1

INTRODUCTION

Day in, day out, patients live with their disease. They must balance the pressures of daily life with their medical care to create an acceptable life for themselves. As a patient, they may need to deal with symptoms like pain, side effects like fatigue, and social stigma. Yet, they are not only a patient, but may also simultaneously be a parent, son or daughter, friend, partner, or employee. These other roles may conflict with optimal medical care. For example, COPD patients indicate that they may neglect their bodies because of concerns for their children; financial worries or because they simply have too much to do [238]. While professionals often approach patients from a primarily medical point of view, patients need to weigh different values of which 'taking good care of one's body' is but one [49, 56, 238]. Although the tension between values is probably the largest for patients with chronic conditions, patients with more acute conditions may also face conflicting roles and values albeit for a shorter period of time.

By living with their disease, patients accrue experiences and thereby knowledge by acquaintance; "knowing" as in being familiar with, such as knowing what it is like to have an asthma attack [49]. Experiential knowledge arises when experiences lead to personal insight that allows a patient to cope with their illness [56]. Sociologist Borkman [38] was the first to theoretically define this term. Experiential knowledge is "truth learned from personal experience with a phenomenon rather than truth acquired by discursive reasoning, observation, or reflection on information provided by others". Experiential knowledge is mostly implicit and is often compared to cycling; one can put it into practice but it is difficult to describe and explain to someone else [56]. A patient attains "experiential expertise" when they are able to make these coping skills explicit and transfer their knowledge to others [38]. When patients share their experiential knowledge in person or online, the communal body of knowledge exceeds the limits of individual experiences and becomes "collective experiential knowledge" [56].

It is this collective experiential knowledge that is currently underutilized by medical research even though it could both direct research priorities and provide a complementary data source for new medical hypotheses. In Figure 1.1, we depict the current state of knowledge transfer between medical professionals, patients, and researchers (indicated

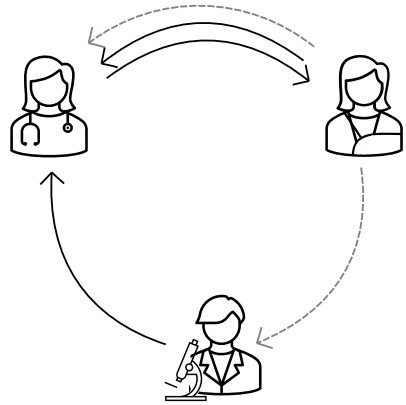


Figure 1.1: The knowledge cycle between medical professional, patient and researcher. The black lines indicate the current knowledge transfer; Researchers inform medical professionals who in turn share their knowledge with patients. Individual patients also share their knowledge with individual medical professionals. The gray dotted lines indicate the addition we want to make to this cycle by sharing the collective experiential knowledge of the patients with both researchers and medical professionals.

with solid black lines). We propose to utilize the collective experiential knowledge of the patient to improve medical research through knowledge transfer from patient to researcher and from patient to medical professional (as indicated by the blue dotted lines).

The first way in which collective experiential knowledge could benefit medical research is by directing research priorities into directions that match the needs of patients. Collective experiential knowledge can provide insight into which issues have been neglected or are considered most pressing in terms of their impact on quality of life. These priorities may not align with those of professionals. In the words of a patient:

"Something that objectively seems a "mosquito" can be subjectively be experienced as an "elephant" throughout your life." [56]

Since the 1990s, there is an increasing recognition that the patient perspective is valuable and thus patients should be involved in decisions about the research agenda. The role of patients has shifted from passive subjects to active partners through participation in advisory panels [45, 49]. There are two main reasons for this shift. The first is moral: Medical research is largely a public good and consequently, decision making about the direction of research should be a collective process involving all relevant actors, including patients. People have a moral as well as legal[1] right to participate in decisions that directly affect them. The second reason is more practical: Involving patients leads to research that is more relevant and of higher quality. The experiential knowledge of patients can provide a wider perspective and range of options [240]. Moreover, an empirical analysis of cases of patient participation in biomedical research by Caron-Flinterman et al. [49] suggests that experiential knowledge can contribute to the relevance and quality of biomedical research when translated into explicit demands or ideas.

[1] Wet op de geneeskundige behandelingsovereenkomst (WGBO)

The value of health care for the patient is the central tenet of value-based healthcare, a framework advocated by American Economist Michael E. Porter [205]. He postulates that high value for patients should be the overarching goal of health care. In turn, "value in health care is measured by the outcomes achieved, not the volume of services delivered" (p. 2477). Michael E. Porter [205] stresses that in order to gear health-care towards more value from the patient it is essential to measure and report outcomes systematically and over a long period of time. These outcomes should include health circumstances that are most relevant to the patient. Other researchers have underscored these ideas: Castro et al. [56] and Kickbusch and Gleicher [155] argue that the inclusion of patient knowledge can lead to health care that better meets the needs of patients and thereby ultimately leads to both better quality of care and quality of life.

The second way in which collective experiential knowledge could benefit medical research is by complementing professional medical knowledge and providing novel medical hypotheses. Previously, experiential knowledge has often been neglected because it is considered subjective and therefore not evidence-based nor scientific [56]. This is a classic logical positivist stance: any knowledge that is not objective, verifiable and rational is not valid [49]. However, this epistemological[2] view is contested, as even in science, pure objectivity is not possible and norms such as rationality are dependent on conceptual frameworks (or paradigms). As such, there cannot be one absolute truth.

Moreover, even if we accept that experiential knowledge is not objective or scientific, this does not undermine its validity to science. In the scientific method, hypotheses are often inductive, i.e., based on a finite set of observations. Experiential knowledge is well suited for providing such inductive evidence. Needless to say, these hypotheses still need to be validated by scientific research and are but part of the scientific process. As [243] rightly notes experiential knowledge alone cannot uncover causes or underlying mechanisms.

Experiential knowledge should not replace but can complement professional or academic knowledge as a source for hypothesis generation. In contrast to professional knowledge which is empirical and scientific but segmented [38], experiential knowledge is pragmatic, and holistic. Moreover, patients do have knowledge that professionals do not, mainly concerning how to cope with disabilities and situations. For example, patient reports of adverse drug events show patients can offer an independent, complementary perspective from that of health professionals [9, 77].

Aside from benefiting medical research, harvesting collective experiential knowledge can also benefit patients more directly by empowering them individually. It can, for instance, provide them aggregate insights into coping methods employed by other patients. Various previous qualitative studies have indicated that patients rely on the experiences of other patients for informational support (i.e., advice or guidance) [45, 157, 277, 324, 329].

In this thesis, we will harvest collective experiential knowledge from patient forums automatically using text mining techniques. In the following section 1.1, we will discuss why patient forums are a good data source for obtaining collective experiential knowledge of a specific patient population. In section 1.2, we will then discuss why text mining will be used to harvest the collective experiential knowledge from these forums. We focus our

[2]Epistemology is the philosophical study of the nature, origin, and limits of human knowledge

research on rare disorders, as collective experiential knowledge is especially promising for them. We discuss why this is the case in Section 1.3.

The automatic extraction of collective experiential knowledge from patient forums has not been researched previously. However, there has been some work into the extraction of adverse drug events specifically from patient forums. We will build upon this work and therefore we discuss previous work on adverse drug event detection from social media in Section 1.4. Finally, in Section 1.6 we outline the rest of the thesis and the research questions that have been explored.

1.1. PATIENT FORUMS AS A SOURCE OF COLLECTIVE EXPERIENTIAL KNOWLEDGE

Patient forums are forums centered around patient communities. Previous qualitative work has shown that patients gather on patient forums to exchange information and experiences; and support each other emotionally [157, 202, 277, 319, 324]. A patient forum can be a group on a general social media platform such as Facebook or on a specialized platform like PatientsLikeMe. Patient forums are a type of domain-specific or specialized social media, in contrast to generic social media like Twitter.

Social media in general has a number of distinct advantages compared to other potential information sources of collective experiential knowledge such as surveys or focus groups. The first advantage is the large volume of information that is not easily obtainable by other means [262]. A second advantage is that the information posted on social media is uncensored and thus unbiased by intermediaries. Previous studies [98, 128] have shown that the attitudes of medical professionals can bias the official reporting of outcomes reported by patients. Medical professionals may not report an adverse drug event (ADE) reported by a patient for various reasons including lack of time, uncertainty about whether the drug causes the ADE or because the ADE is either trivial or well-known [98, 128]. A third advantage is that patient-reported experiences on social media are unsolicited [128]. It has previously been found that patients share more information in unsolicited peer-to-peer interactions than with their physicians or at clinical trials [75]. Two other advantages of social media are that in contrast to surveys and focus groups no additional burden is placed on the patient and that it allows for real-time monitoring and early signal detection by providing near-instantaneous information [276].

Patient forums are particularly valuable as an information source, as opposed to, for instance Twitter, because they contain more posts where patients share information (i.e., give informational support) [116]. Previous research into medical knowledge discovery from social media has focused on the detection of adverse drug events. A systematic literature review by Golder et al. [114] revealed that 12 to 62% of posts concerning the illness of interest on disease-specific forums contained an adverse event, whereas only 0.02 to 11.5% of such posts on Twitter contained an adverse event. Moreover, forum posts unlike tweets are not restricted in their length[3], allowing for more elaborate discussions. Nonetheless, most research on medical knowledge from social media at present has focused on generic social media instead of patient forums [171, 262].

[3]Tweets are restricted to 280 characters. In our data, the median number of characters for forum post is 89 but can range up to 12.098: so 43 times the maximum length of a tweet.

Patient forums can be differentiated into disease-specific and medicine focused forums [228]. An example of a disease-specific forum is a patient forum for a specific patient population such as patients with breast cancer. In contrast, medicine-focused forums are forums where patients with a range of different diseases leave reviews for the medication they are taking, e.g, reviews for the drug paclitaxel on Askapatient.com. Our work centers around disease-specific patient forums, which we can use to harvest the collective experiential knowledge for a specific patient population.

Researchers have also posited some concerns with the use of social media: data may be unreliable, the privacy of the patient may not be sufficiently protected, and the forum users may not be representative of the general patient population [40, 58]. The first concern appears unfounded at least for reports of adverse drug events, which is the only type of experience that has been researched previously; They were shown to be of similar quality compared to those of health professionals [37, 322]. Furthermore, this concern can be mitigated by considering only the collective (i.e., the "wisdom of the crowd") and not individual experiential knowledge and through correct interpretation of results as hypotheses that require further validation.

Both the second and third concerns are valid points that should be taken into consideration. In our work, we protect patient privacy in the following ways: by only working with publicly available data and forum groups; by complying with the GDPR; adhering to data minimization principles (i.e., do not collect and store data you do not need); by restricting access to data to protect patient privacy, and by using private servers for data labeling. The representativeness of forum users is difficult to determine as this user information is either not available or not collected to protect the user's privacy. However, sample bias is not unique to social media as a data source. Clinical trials for example also suffer from sample bias, as they mostly include relatively healthy patients and exclude the elderly, patients with comorbidities, pregnant women, and children [274, 289]. Nevertheless, it is essential to take this into consideration when interpreting the results of automatic knowledge discovery from social media.

1.2. USING TEXT MINING TO HARVEST COLLECTIVE EXPERIENTIAL KNOWLEDGE

Text mining encompasses techniques that allow software to extract useful information from text, for instance, from social media messages or academic articles [107]. This includes methods that extract information, disambiguate[4] words, translate text or automatically summarize text. The first text mining techniques were rule-based, but over the past decades, text mining has been dominated by machine learning techniques. Machine learning methods are methods that teach computers (i.e., the machine) to learn and improve from experience without being explicitly programmed. Computers may gain experience through training examples that are provided for a certain task like entity extraction, just like humans learn how to complete a task through practice. These examples have often been labeled by humans. The research field that deals with how computers can be used to understand and manipulate natural language is called Natural

[4]Disambiguation is the determination of the sense or meaning of a word e.g., 'bank' as in sofa or 'bank' as in place to deposit money

Language Processing (NLP) [63].

Using text mining techniques, we are able to process and extract information from the large volume of messages on a patient forum. Automatic extraction does also introduce errors into the data as automatic methods cannot attain perfect performance, i.e., computers cannot understand language like humans can. For example, information may be missed (i.e., false negatives), or information might be found that is not there (i.e., false positives). Consequently, the information extracted from patient forums may be more noisy than information obtained from other sources such as surveys or clinical trials. Automatically extracted experiential knowledge should be interpreted in this light and only seen on an aggregate level. Further clinical research or surveys are then necessary to validate the hypotheses that result from automatically extracted collective experiential knowledge.

1.3. POTENTIAL FOR RARE DISEASES

Although hypothesis generation from forum data could benefit all patient communities, it is most valuable for patients with a rare disease [15]. These diseases are largely neglected by the research community: their rarity obstructs collecting large samples of data and the for-profit industry considers R&D for these diseases too costly [305]. New orphan drug legislation in the US and Europe have managed to improve the financial stimulus for research into rare diseases [118], but this is insufficient to incentivise adequate research for diseases with a very low prevalence [131]. Online forums could enable the coordinated, trans-geographic effort that is necessary to attain progress [15] in this research field. The general necessity to use trans-geographic research for dispersed groups through greater involvement of citizen (data) in scientific research is increasingly being recognized as essential [92].

What qualifies as a rare disease differs amongst different jurisdictions but on average, a rare disease lies between 40 and 50 cases per 100.000 individuals [249]. This translates to roughly 5000 to 8000 rare diseases in total affecting 27 to 36 million people in the EU [95] and between 25-30 million [306] in the United States.

Forums of patient communities with rare diseases are relatively active and focused, providing each other with useful information (i.e., informational support), due to the lack of research and other resources. Patients with rare diseases indicate that they find better information in online support groups than by talking to their physicians for many aspects of their medical care [109]. Furthermore, there are various cases of patient communities of rare diseases who have responded to the lack of medical provisions by mobilizing into grassroots organizations. These organizations proceed to analyze and aggregate their own patient-to-patient data to help others cope and to attempt to drive research and move closer to effective treatment [49, 108, 237].

For instance, a retrospective, observational study on the registry of patients with Gastrointestinal Stromal Tumor (GIST), a rare oncological condition, elucidated prognostic factors for subtypes of GIST and the impact on survival for different age groups [47]. Occasionally, patients are prompted to self-test based on the gathered anecdotal evidence and this data is also analyzed. An example is the use of indole-3-carbinol/3,30-diindolylmethane (found in cabbage) for Recurrent Respiratory Papillomatosis (RRP): one patient reported in the community's newsletter that it helped his daughter tremendously,

and subsequently patients started self-testing. The response rate was > 50% and the remission was 20%. A pilot study revealed similar results [68], but no formal Randomized Clinical Trial was conducted due to funding shortages and lack of professional interest [237]. Unfortunately, this reflects the general outcome of such efforts, as researchers often could not be persuaded to further the research done by patients [237].

To conclude, we will focus our work on rare disorders as their need is pressing, and their patient forums are both active and rich in experiential knowledge. Specifically, we will perform a case study of a large patient forum of patients with the rare oncological condition Gastrointestinal Stromal Tumor (GIST) in collaboration with a Dutch GIST patient organization and the Leiden University Medical Center. GIST has around 10-15 new cases per million each year [285] and it is the most common of the sarcomas; A group of mesenchymal tumor types that originates from the bone or soft tissue[5] of the body.

1.4. PREVIOUS WORK FOCUSED ON PHARMACOVIGILANCE

Previous work on extracting patient experiences from patient forums has focused solely on extracting adverse drug events (ADEs). The term adverse drug event is used to refer to "any untoward (i.e, unexpected and negative) medical occurrence that may appear during treatment with a pharmaceutical product but which does not necessarily have a causal relationship with the treatment" [102, 341]. Although often used interchangeably, the term adverse drug reaction (ADR) infers a causality relation between drug and effect, according to the World Health Organization [342]. Adverse drug reactions are generally established during clinical trials before the drug is approved and released unto the market. Afterwards, ADRs are monitored through post-market surveillance systems where doctors can report ADRs they diagnose in their patients.

Although a causal relationship is difficult to infer from nonclinical data like social media, the extraction of patient-reported adverse drug events has attracted attention due to its potential value for the post-market monitoring of drugs (also called *pharmacovigilance*). Current post-market systems lead to severe under-reporting of ADEs: on average only 10% of ADEs are discovered [130]. There is an increased recognition that information sources more representative of the everyday 'real world' are necessary [160, 236, 244]. Social media data is seen as one promising resource for the discovery of ADEs [13, 228].

Nevertheless, empirical research into the extent to which automatic extraction of ADEs from patient forums can benefit pharmacovigilance is limited to three studies [30, 321, 346]. Only the study by Benton et al. [30] focused on a specific disease, namely breast cancer, instead of a subset of drugs. They found that 20-25% of the automatically extracted ADEs were novel (i.e., not mentioned on the official drug label). Yeleswarapu et al. [346] similarly found unreported ADEs in user posts for 12 drugs of interest. In contrast, van Stekelenborg et al. [321] conclude that social media has no additional value as it is not able to detect signals before official post-market systems do. Their automatic analysis of 75 drugs in over 6 million posts from a mix of social media sources including Twitter, Facebook and patient forums was the first large-scale study [55] into the value of ADEs from social media for pharmacovigilance. Yet, van Stekelenborg et al. [321] also

[5] Soft tissue includes cartilage, fat, muscle, blood vessels, fibrous tissue and other connective tissues.

posit that if the extraction of ADEs could be improved, their conclusions would need to be revisited, as to a large extent, the value of patient forums for pharmacovigilance will depend on the quality of the automated extraction process. Since their methods appear to be outdated (state-of-the-art methods are further detailed in Section 1.5), we agree that their conclusions are preliminary.

Previous research on Twitter data reveals that social media data may complement traditional data sources by revealing under-reported mild ADEs. Overall, previous work found a high level of overlap between the ADEs found on social media and through traditional reporting channels. However comparatively more mild and symptomatic ADEs can be found on social media, whereas serious ADEs are underrepresented [114, 115, 280]. Directly comparing prevalence of adverse drug events is challenging, however, because prevalence is measured differently in different data sources [280]. Another way in which social media may complement post-market systems is by providing more information on the impact of ADEs on daily life. Although registration of ADEs through clinical trials or medical professionals often includes an assessment of the severity of the ADE according to the Common Terminology of Adverse Drug Events [307], the impact on daily life is not reported. Patient reports on social media may be able to provide insight.

Our work differs from previous work in a number of important regards, allowing us to shed new light on this discussion but also on the broader question of the value of experiential knowledge for medical research. The most important difference is that we go beyond the extraction of ADEs and also extract the coping strategies patients recommend for dealing with them. Our work also has two major differences regarding automatic ADE extraction. Firstly, with the exception of work by Benton et al. [30], previous studies use a list of drugs as a starting point. Since drugs are often prescribed for various diseases, they thus assess ADEs from various patient populations for a particular drug. Thereby they neglect that ADEs may differ for different patient populations for the same drug. In contrast, we focus on assessing ADEs experienced by a particular *patient population* for the drugs they take. Drugs need to be approved for each disorder separately through clinical trials. During these clinical trials, ADEs are monitored and collected. Our results can thus be compared to those of registration trials specific to the patient population at hand to understand which ADEs are novel. We believe that this approach is more promising than grouping various patient populations, as it allows for a more detailed investigation of ADE specifically for patients with rare diseases. Secondly, previous work relies on traditional machine learning methods (e.g., rule-based) that are no longer considered state-of-the-art in the field [337], while we aim to employ state-of-the-art methods. The integration of state-of-the-art methods such as BERT-based models for various components of the extraction process into one pipeline involves methodological challenges that we will discuss next.

1.5. METHODOLOGICAL TASKS AND CHALLENGES

The extraction of ADEs from social media consists of two steps with each their own set of challenges. These steps are: the extraction of text snippets that mention an ADE and mapping of these snippets to the correct medical concept in an ontology (see Figure 1.2). The first step is most closely related to the classical NLP task called Named Entity Recognition (NER) in which named entities such as person names or locations are

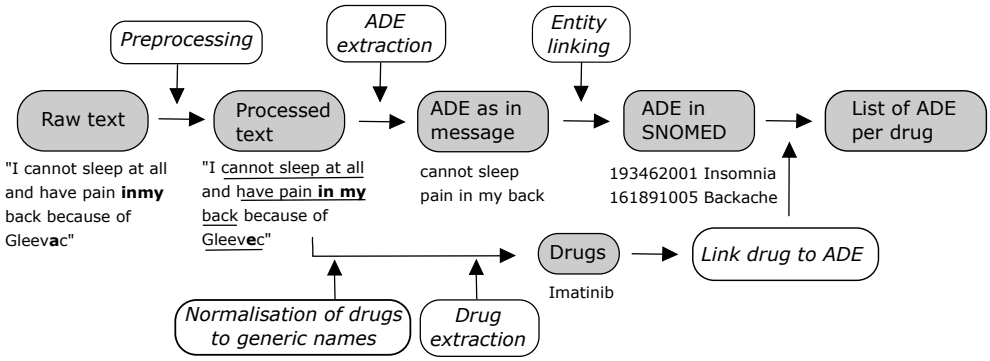


Figure 1.2: An overview of the pipeline for harvesting ADEs from social media. Italicized parts indicate technical steps. An example message is provided to clarify each step

extracted [149]. NER is a type of sequence labeling task: each word in the text is labeled for whether it is part of an entity or not. An example of a text snippet containing an adverse drug event is "cannot fall asleep" (see Figure 1.2). Extracting complex concepts like ADEs is far more challenging than extracting locations or person names as unlike named entities they are not proper nouns and can be described descriptively. Furthermore, there is large variation in the language that can be used to describe the same medical concept. For example, one could describe their headache with 'my head is bursting', 'throbbing pain in my head' or 'pounding headache' to name a few. We will discuss NER of adverse drug events in Section 1.5.1. The second mapping step is called entity linking (e.g., "cannot fall asleep" would be linked to the concept Insomnia). Two major challenges here are the large number of medical concepts that an entity could be linked to, and the large difference in language between layman terms and formal medical terms. We will discuss this second step in Section 1.5.2. Finally, in Section 1.5.3 we will consider previous work on determining which drug an ADE is linked to according to the patient. As most of the methodological work on ADE extraction has been on Twitter data and there is a strong overlap between Twitter and patient forum messages, we will cover both methodological work on Twitter and patient forums indiscriminately.

1.5.1. ADE EXTRACTION

The first study to perform NER for adverse drug events was a study by Leaman et al. [173]. They used a lexicon-based approach[6] with a list of pre-compiled ADE mentions. This is a common technique in earlier studies [30, 188, 216, 258, 344, 345] because labeled data was often absent whereas extensive medical resources such as the UMLS[7] were available for building such lists [266].

However, lexicon-based approaches are not able to deal with the creative and descriptive explanations that patients use to describe their ADEs like "messing up my sleeping patterns" [112, 224, 266]. Consequently, with the growth of annotated data sets

[6] A lexicon is a vocabulary of a words of a certain branch of knowledge, in this case ADEs

[7] The Unified Medical Language System is an integrated terminology from various biomedical vocabularies and standards and can be found at: `https://www.nlm.nih.gov/research/umls/index.html`

for ADE extraction, there was a shift towards machine learning approaches [116]. The first study to use machine learning was Nikfarjam and Gonzalez [216]. They used association rule mining to mine language patterns that are used for ADEs and then used these rules to extract them. Later studies use sequence labeling methods, in which the model attempts to determine for each word in the sentence whether it is part of an ADE. Specifically, they made use of Conditional Random Fields (CRF) models [203, 217]. The primary reason for their success was their ability to incorporate contextual information; Users might use a variety of creative terms for the same concept, but similar concepts are likely to occur in the same context [116].

While CRF models learn from features (e.g., the previous word and the part-of-speech tag[8]), deep neural network models, like RNN, do not require feature engineering. They learn directly from the raw text. They use examples to find features that will help them learn the goal of the task. Generally, deep learning models will improve as the amount of data they are trained with increases. They are able to learn from raw text because of how they are designed: Deep learning models consist of layers that are increasingly more complex. This architecture was inspired by the human brain. The lower layers of the model learn to recognize specific features, while the upper layers will use the information from the lower layers to recognize increasingly complex features. For instance, lower layers may recognize stripes whereas upper layers compile features from lower levels to recognize zebras. The ability of models to learn from raw data without manual selection of features also has a flip side as it makes it hard to understand what a model is basing its predictions on.

A deep neural network was shown to outperform CRF models for extraction of ADEs [67]. The BiLSTM RNN models made use of pre-trained skip-gram word embeddings[9] trained on Twitter data. This was shown to improve recall. The BiLSTM RNN models also had improved precision due to their contextual awareness. BiLSTM models process sequences both in a forward and a backward direction, allowing these models to learn dependencies in both directions.

In recent years, models based on the highly efficient Transformer architecture [326] have dominated the field. Transformer models rely in the principle of transfer learning: the reuse of a language model pre-trained on a large amount of unlabeled text. These pre-trained language models can be fine-tuned to perform a specific task such as ADE extraction using training data specific to the task. Yet, transformer models are not the first models to use transfer learning. For instance, the use of word embeddings such as Word2Vec as was done by Cocos et al. [67] is also transfer learning. However, the introduction of BERT models [84] greatly improved the potential of transfer learning, because BERT was able to produce context-dependent embeddings. This means that a token is represented differently depending on the context it occurs in, e.g., “bright” in “the lamp is bright” and “the child is bright” will be represented differently. In contrast, traditional word embeddings like Word2Vec only compute one static representation for each word.

A major advantage of using pre-trained language models is that they can be shared. For the medical domain, numerous domain-specific BERT models have entered the stage,

[8] A part of speech tag denotes a word’s function like noun or verb

[9] A word embedding is a vector representation for text in which words that are similar are close together

such as BERT models retrained on biomedical articles (BioBERT [174]), retrained on clinical records (ClinicalBERT [6]) or trained from scratch on scientific articles (SciBERT [28] and PubmedBERT [119]) or on patient forum messages (EndrBERT [303]).

In the most recent edition of the Social Media Mining for Health (SMM4H) shared task for ADE extraction, all entries were based on Transformer models [193]. Of those models, the one that attained the best performance (F_1 score of 0.29) for the overall ADE extraction pipeline (i.e., filtering for relevant tweets + extraction + linking) included an extraction component based on the EndrBERT model [257]. We do not know whether EndrBERT will perform equally well on forum data as the language in forum messages differs from the language used in tweets. The model that performed best on the extraction component of the pipeline alone was a BioBERT model with a multi-task learning strategy [86]. In multi-task learning, a model is trained on multiple tasks simultaneously.

Some studies add an additional step before extraction in which they detect messages that contain an ADE. This was introduced by the SMM4H shared task in 2017 [265]. The underlying idea was that pre-selecting tweets with ADE would aid performance and end-to-end extraction. However, extraction can also be affected detrimentally if messages with ADE are wrongly filtered out leading to error propagation in the pipeline. In a recent collaborative study [194], we found that adding a prior ADE classifier for relevant messages can still be beneficial for BERT models.

1.5.2. ADE NORMALIZATION

After finding phrases that describe ADEs, they need to be linked to medical concepts to aggregate them (i.e., to recognize different descriptions of the same ADE). Generally, social media data is more challenging than ADEs from other data sources like scientific abstracts due to the language gap between the lay public and medical professionals [304]. Initial methods used string matching or lexicons, but these performed very poorly [204, 300] because patients use laymen language instead of formal medical terms.

More recent work treats the problem as a classification task[10] with medical concepts as target classes. Here, textual mentions of ADEs (e.g., 'feeling dizzy') are only treated as phrases and its context (i.e., the rest of the sentence) is not taken into account. Various studies [27, 124, 184, 302] used deep neural network models to classify ADEs in this manner. A more recent study by Miftahutdinov and Tutubalina [206] found that BERT models are able to outperform deep neural network models.

In contrast, the current state-of-the-art model BioSyn [291, 304] treats ADE normalization as a ranking task in which the target concepts are candidates that are ranked according to their likelihood of being the correct concept. The context of the ADE is still ignored. To rank the target concepts, the BioSyn model uses both dense BERT embeddings and sparse embeddings based on Tf-idf term weighting[11] for representing the entities and calculating their similarity to the target concepts and their synonyms. Supervised data is used to maximize the marginal probability of positive synonyms of an ADE mention. The model outputs a ranking of the most similar synonyms that has been found for an ADE mention. For implementation and comparison to classifier approaches,

[10] Classification is a task in which a model predicts for each item which category or class it belongs to

[11] Term frequency-inverse document frequency (tf-idf) reflects how important a word is to the document in the collection

only the top ranked concepts is selected for each ADE mention. Besides showing that BioSyn outperforms BERT for ADE normalization of social media, [304] showed that this is also true if the test set only contains unseen entities albeit with a drop in performance of 23.3 percent points (from 83.8% to 60.5% accuracy). We also found reduced performance on unseen entities in our work with Magge et al. [194].

The BioSyn model is capable of predicting target concepts for which it does not have training data because of its reliance on a pretrained embedding space. The inclusion of *all* medical concepts as targets greatly increases the number of possible classes and essentially forces models to be able to deal with zero-shot scenarios.[12] However, this development is essential, as creating training data for all possible medical concepts is unrealistic. This is only a recent, albeit key, development in the field. The sixth edition of the SMM4H task [193] was the first to include zero-shot cases in the test set: 257 new MedDRA[13] classes were part of the test set, whereas 669 classes overlapped between test and training data. Nonetheless, the performance on these zero-shot cases is not evaluated separately and thus it remains unclear how well models perform in a zero-shot scenario.

Another recent development within the field is an increased focus on end-to-end ADE resolution (i.e., NER and subsequent normalization) instead of on individual components of the pipeline. An end-to-end ADE resolution task was introduced in the SMM4H shared task of 2018 [335]. Weissenbacher et al. [335] found that extraction was the main bottleneck with normalization alone attaining up to 88.7% accuracy while the performance of end-to-end resolution remained low at an F_1 score[14] of 0.432. This has been corroborated by later shared tasks [159, 193] and our own collaborative work [194].

1.5.3. DETERMINING ADE–DRUG RELATIONS

Most studies at present have used a particular list of drugs as a starting point. This introduces an additional challenge, namely forum posts or tweets pertaining to that particular drug need to be filtered prior to extraction. There is often no information available on which disorder the patients taking the drug have. Using our approach, the reverse is true: as we focus on disease-specific forums, we do know which particular patient population we are assessing, yet do not know which drug the ADE is reported for.

There has been some previous work on linking ADEs to their respective drugs as reported by the patient. Early studies simply used co-occurrence as a basis for ADR-drug relations [30, 173]. Yet, these methods had a low precision [188] and could not deal with multiple drugs mentioned in one message. Yang et al. [344] used more advanced co-occurrence methods that calculated the actual co-occurrence probability based on their independent occurrences and co-occurrences. Later studies have attempted to further specify the relations between ADEs and drugs to identify those that indicate a causal relationship using statistical learning [188, 258]. Here, causality refers to whether the patient reports a causal relationship between a drug and an ADE, as opposed to factual causality. In the work of Liu and Chen [188], causality detection was a two-stage process

[12] A zero-shot scenario is a case where a model must classify an instance of a class without having observed any instances of that class during training

[13] The Medical Dictionary for Regulatory Activities (MedDRA) is a medical ontology that contains 79.507 classes total and is maintained by the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH).

[14] F_1 score is the harmonic mean of precision and recall. It is a common metric in NLP.

consisting of relation detection and relation classification. For each step, a classifier is trained in which the first classifier distinguishes whether entities have a relation or not and a second, in turn, defines the relation type. In this manner, Liu and Chen [188] differentiated between true ADEs, negated ADEs (i.e., the drug did not cause side effect) and drug indications (i.e., the disease for which the patient takes the drug). They found that although co-occurrence attained a recall of 1, the precision was only 0.385, resulting in an F_1 score of 0.556. Their method performed better with an F_1 score of 0.669 (Precision at 0.82 and Recall at 0.565). Sampathkumar et al. [258] used a very different approach, namely they predicted the presence of a causal relationship using a list of keywords (e.g., 'effects from') that were extracted from annotated causal relations of the training set using a Hidden Markov Model (HMM). Their classifier outperformed a co-occurrence baseline with an average F_1 of 0.76 compared to a F_1 of 0.575.

There have also been various studies[62, 217, 262] that treat the extraction of ADEs as a combined entity and relation extraction task by including the necessity of having a reported causal relation to a drug as a prerequisite for being an ADE. They explored the use of linguistic features to differentiate ADEs from drug indications or disease symptoms that lack such a relation. In a similar vein, one recent data set [353] explicitly annotated various other medical categories that ADEs could be confused with namely: withdrawal symptoms (i.e., symptoms you get from reducing drug intake), symptoms of the illness, and drug indications. Although including the patient reported causal relation as a prerequisite for ADEs may be beneficial, it does not resolve *which* drug the ADE relates to and thus these methods do not suffice for disease-specific forums.

1.6. RESEARCH QUESTIONS AND THESIS OUTLINE

The work presented in this thesis lies at the intersection between computer science and medicine. As most of the work has been on technical developments, the main reading audience is computer science researchers, and researchers in natural language processing in particular. However, we have added footnotes to explain text mining concepts throughout the introduction and we provide tailored abstracts for each of the chapters in an effort to make this thesis more accessible to medical researchers.

The focus of this thesis is the discovery of experiential knowledge from patient forums through text mining methods and its complementary value to traditional sources of medical knowledge for scientific hypothesis generation. The main research question answered in this thesis is thereby:

Main RQ To what extent can automated extraction of experiential knowledge from patient forum posts aid knowledge discovery to yield hypotheses for clinical research?

Qualitative investigations of patient forums have already revealed that patients share a large variety of experiential knowledge, for example on when and how to take medication (i.e., 'chronomedication'); on to how to psychologically deal with the disease; and on which adverse drug events occur and how to best cope with them. In this work, we will focus on the latter to build upon the work that has already been done on the extraction of adverse drug events (ADEs) from social media. Nonetheless, many of the methodological

challenges we deal with for extracting these specific types of experiential knowledge are also relevant for extracting other experiences.

We address three main methodological challenges or bottlenecks to the extraction of ADEs and the coping strategies patients recommend in parts 1 through 3. In part 1, we present methods to deal with the noise present in social media data, and medical social media in particular. In part 2, we address the text mining challenges presented by the extraction of the text snippets containing adverse drug events from patient forum messages. In part 3, we will investigate how to extract coping strategies from patient forum posts and link them to the relevant ADE.

Finally, in part 4, we present the results for a case study on a specific patient forum for Gastro-Intestinal Stromal Tumor (GIST) patients and demonstrate the value of extracting ADEs from patient forum posts for post-market drug monitoring. Moreover, we compare the ADE reported on a GIST-specific patient forum to responses to standard patient-reported outcome measurement tools amongst Dutch GIST patients. We also explore how representative the patient population active on a patient forum is for GIST patients since representativeness is a commonly noted concern for social media data and online communities[13, 23, 32, 58, 276, 287].

1.6.1. PART 1: REDUCING NOISE IN MEDICAL SOCIAL MEDIA TEXT

The first step in a natural language processing pipeline is preprocessing, or the cleaning of data prior to data analysis. Examples of possible preprocessing steps are spelling correction, removing punctuation or lowercasing text. See Figure 1.2 of preprocessing a forum post about an adverse drug event.

There are two types of noise we focus on in the preprocessing stage of patient forum messages, namely noise within the text of the message itself and noise presented by the large number of irrelevant messages compared to relevant messages for the task at hand

The first type of noise is the difference between the noisy language used on patient forums and more formal and 'standard' language used in academic articles and dictionaries. A larger variation in language use is harder for models to learn for subsequent steps and moreover, most methods and models are developed for more formal language and thus are not equipped to deal with social media messages. This type of noise reduces the ability of machine learning algorithms to extract entities and map them to medical concepts [173, 216, 224, 229].

Unlike formal text, social media messages contain informal, conversational language (i.e., colloquial language) and frequent misspellings and abbreviations [116, 229]. The noise is aggravated on medical social media by laymen medical terms such as 'high blood pressure' instead of 'hypertension' and 'cannot sleep' instead of 'insomnia' [116]. These laymen terms are not present in specialized medical dictionaries. Moreover, medical terms like drug names are both essential to the knowledge extraction task but also particularly prone to spelling mistakes.

As previous work [261] dealt with the normalization (or standardization) of abbreviations and informal language, we focus on investigating how to deal with misspellings of medical terms. In Chapter 2, we investigate:

RQ1 To what extent can corpus-driven spelling correction reduce the out-of-vocabulary

rate in medical social media text and improve the accuracy of subsequent classification tasks?

The second type of noise that complicates knowledge extraction is the low signal-to-noise ratio. A review by Golder et al. [114] showed that around 8% of the posts on a patient forum were reports of adverse drug events. In Chapter 3, we look at identifying patient experiences (also coined narratives) as a way to reduce this type of noise. In Chapter 4, we look at including the conversational context in which the message was posted to better identify which messages are relevant. These chapters revolve around the following questions:

RQ2 Which features distinguish patient narratives from other social media text and how can they best be identified?

RQ3 To what extent can the addition of conversational context to state-of-the-art models improve the identification of relevant posts?

1.6.2. PART 2: EXTRACTING ADVERSE DRUG EVENTS

Since Weissenbacher et al. [335] found named entity recognition (NER) to be the main bottleneck in end-to-end ADE resolution, the second part of our work focused on further understanding the weaknesses of state-of-the-art NER methods and tackling obstacles for end-to-end ADE extraction.

In Chapter 5, we investigate the efficacy of several default transfer learning methods for extracting ADEs and subsequently normalizing them to a standard medical ontology. In Chapter 6, we explore the weaknesses of the state-of-the-art transfer learning model BERT for Named Entity Recognition (NER) through targeted attacks (i.e., adversarial attacks). In Chapter 7, we deal with the problem of discontinuous entities: entities that either overlap with other entities (e.g., *hand* and *foot pain*) or are split (e.g., *eyes* are feeling *dry*). We propose a fuzzy representation schema for these entities and explore its impact on both the extraction and normalization of ADEs.

These chapters center around the following questions:

RQ4 How effective are default transfer learning methods for extracting and normalizing adverse drug events?

RQ5 How vulnerable are BERT models for Named Entity Recognition to adversarial attack and to which variation are they most vulnerable?

RQ6 To what extent can a fuzzy continuous representation of discontinuous entities improve the extraction and normalization of adverse drug events?

1.6.3. PART 3: EXTRACTING RECOMMENDED COPING STRATEGIES

Aside from sharing which adverse drug events they are experiencing, patients also give each other advice on how to deal with them. In part 3, we focus on extracting these coping strategies. As this is a novel task, we develop an ontology for coping strategies and create three annotated data sets, namely for NER, normalization and extracting relations between adverse drug events and coping strategies (CS). In chapter 8, we introduce this

task and the resources we have created. We present baseline results for CS extraction and perform a case study on the GIST patient forum with the first end-to-end pipeline for extracting coping strategies.

This chapter answers for the following question:

RQ7 To what extent can coping strategies for adverse drug events be extracted automatically from online patient discussions?

1.6.4. PART 4: COMPLEMENTARY VALUE OF DISEASE-SPECIFIC PATIENT FORUMS AS A SOURCE OF PATIENT-REPORTED OUTCOMES

In part 4, we present three studies outlining and exploring the complementary value of automatic ADE extraction for a case study. We focus on a large forum for patients with Gastrointestinal Stromal Tumor (GIST), a rare oncological condition. In Chapter 9, we argue that ADEs from patient forums can be used to complement current pharmacovigilance (i.e., post-market drug monitoring) systems. We show with examples that a patient forum can provide real-world evidence for pertinent ADEs according to patients, long-term ADEs, and ADEs not found in registration trials. In Chapter 10, we explore the overlap and differences between ADEs reported on a GIST-specific patient forum and responses to standard patient-reported outcome measurement tools amongst Dutch GIST patients. In Chapter 11, we look at the bias in the patient population that is active on online forums through a survey amongst Dutch GIST patients. The latter two studies were done in collaboration with the Radboud University and the Netherlands Cancer Institute.

This part thus answers the following questions:

RQ8 How can the automated gathering of real-world evidence of adverse drug events from online patient forums complement pharmacovigilance for rare cancers?

RQ9 To what extent are the ADE reported on a GIST patient forum covered by existing patient-reported outcome measures, namely the EORTC QLQ-C30 and the EORTC Symptom Based Questionnaire?

RQ10 To what extent are the GIST patients active on patient forums representative for the GIST population and which sample biases does this data source suffer from?

1.7. SCIENTIFIC CONTRIBUTIONS

Despite our specific focus on extracting experiences from patient forums, our work can contribute to the larger field of natural language processing. On the one hand, it contributes to research on how to improve text mining for user-generated content. On the other hand, it provides valuable insights for research in the biomedical NLP domain and specifically on how to extract and map medical concepts. More broadly, there are numerous other niche domains that similarly deal with both noisy data and relatively small quantities of labeled data and may thus benefit from our work.

The main contributions of this thesis are:

Contribution 1 We developed three methods for reducing the noisiness of medical social media data and improving downstream tasks.

We provide an unsupervised spelling correction algorithm suitable for medical social media data to reduce noise from misspellings as part of a lexical normalization pipeline[15]. This pipeline is also suited for other domain-specific social media data. We present two collected data sets of spelling mistakes from medical social media for future research. [16]

We also provide two methods for reducing the signal-to-noise ratio between relevant and irrelevant posts, namely a classifier for identifying patient narratives[17], and a set of classifiers that can use conversational context to identify relevant posts[18].

Contribution 2 We show the strengths and weaknesses of transfer learning methods for entity extraction in the biomedical domain.

Transfer learning models show excellent performance for a range of NLP tasks. We evaluate how well they work default for ADE extraction and normalization. Scripts[19] and models[20] for this pipeline are open-source. Our pipeline attains a competitive performance with default models, highlighting the strengths of this approach.

However, BERT, a popular transfer learning model, was shown to be vulnerable to deliberate attempts to fool the model (so-called adversarial attacks) for classification tasks in prior work [139, 143, 180, 290, 347]. We expanded this work to sequence labeling tasks[21] to investigate how vulnerable BERT models are for Named Entity Recognition (NER). We compare general to domain-specific models and investigate to what extent the vulnerability may be impacted by domain-specific data. The weaknesses of BERT for NER we uncover can inform future work on mitigating these vulnerabilities. Moreover, the methods we developed are not limited to BERT models alone but can be employed to attack and compare robustness of other transfer learning models for sequence labeling tasks.

Contribution 3 We present a novel simplified representation schema for discontinuous entities in user-generated biomedical text that can benefit end-to-end performance.

We advocate for a continuous representation of discontinuous entities, comprising of both composite (e.g., *hand* and *foot pain*) and disjoint entities (e.g., *eyes* are feeling *dry*). We show how this representation can benefit end-to-end performance of ADE discovery in electronic patient records as well as medical social media compared to the current conventional representation for discontinuous entities (BIOHD). We also make the code to transform data from BIOHD to

[15] Available at: `https://github.com/AnneDirkson/LexNorm`
[16] Available at: `https://github.com/AnneDirkson/SpellingCorpus`
[17] Available at: `https://github.com/AnneDirkson/NarrativeFilter`
[18] Available at: `https://github.com/AnneDirkson/ConversationAwareFiltering`
[19] Available at: `https://github.com/AnneDirkson/SharedTaskSMM4H2019`
[20] Available at: `https://data.mendeley.com/datasets/rxfzx6nbvw/2`
[21] The methods we developed to conduct adversarial attacks on sequence labeling tasks are available at: `https://github.com/AnneDirkson/breakingBERT`

our representation and to compare performance on extraction and end-to-end normalization available to the community[22].

Contribution 4 We propose a novel task, the extraction of ADE-related coping strategies, and introduce an ontology for mapping extracted coping strategies to.

Patients share advice on how to deal with ADEs with each other on online forums. An overview of which coping strategies are recommended could directly benefit patients but can also spur academic research into the potential beneficial or harmful effects of these strategies. We developed the first supervised data set for the extraction and normalization of ADE-related coping strategies from a patient forum and used this labeled data to build the first pipeline for completing these tasks. We are unfortunately not able to share this data with the research community.

However, we introduce this task in the hope others will build on this idea. We compared different possible NLP conceptualizations (i.e., NER with subsequent normalization versus multi-label classification) which can provide a starting point for future work. In order to be able to normalize coping strategies, we built an ontology from parts of existing ontologies to promote transferability. This ontology is not disease-specific, although we recognize new categories may need to be added to accommodate different diseases. This ontology and an end-to-end pipeline for the extraction of coping strategies are publicly available[23].

Contribution 5 We outline the complementary value of disease-specific patient forums as a source of real-world knowledge for pharmacovigilance.

We shed new light on the discussion of the value of social media for pharmacovigilance with our work into ADE extraction from disease-specific patient forums to find ADEs experienced by a particular patient population as opposed to assessing ADEs for a particular drug in patients for whom we do not know their disease. Our work employs state-of-the-art methods whereas previous work relied on traditional (e.g., rule-based) machine learning. We showcase how a patient forum can provide real-world evidence for long-term and novel ADEs. Aside from discovering unknown ADEs, social media is also able to provide a patient-centric view of which ADEs are most pertinent.

Since surveys are considered another option for collecting patient-reported outcomes, we also explored to what extent the adverse drug events patients report differ between patient forums and standard surveys for collecting patient-reported outcomes. We found that outcomes reported on patient forums could disclose ADEs that were not included in the standard measurement tools although they were relevant to patients. Thus, we found that automated harvesting of patient forum data could be used to keep questionnaires up to date.

Representativeness of online patient communities is a commonly noted barrier for using social media as a source of patient-reported outcomes [13, 23, 32, 58, 276, 287].

[22] Available at: `https://github.com/AnneDirkson/FuzzyBIO`

[23] Available at https://github.com/AnneDirkson/CopingStratExtract

Our study into sample bias on patient forums for GIST patients allowed us to indicate which patients are under- and which are over-represented. This in turn can guide the interpretation of patient-reported outcomes harvested from patient forums. We hope to also stimulate research into relevant bias mitigation strategies.

1.8. GUIDE FOR THE READER

This thesis is a collection of peer-reviewed and published papers and papers that are currently under review. This means that chapters can be read independently from the other chapters. An overview of how the thesis chapters interlink and contribute to the extraction and analysis of adverse drug events and coping strategies is provided in Figure 12.1 of the discussion.

Chapter 2: **Anne Dirkson**, Suzan Verberne, Abeed Sarker & Wessel Kraaij (2019), *Data-Driven Lexical Normalization for Medical Social Media*, Multimodal Technologies and Interaction 3(3): 60.

Chapter 3: **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2019), *Narrative Detection in Online Patient Communities.* Proceedings of Text2Story — Second Workshop on Narrative Extraction From Texts co-located with 41th European Conference on Information Retrieval (ECIR 2019). 21-28.

Chapter 4: **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2020). *Conversation-aware Filtering from Online Patient Forums.* Proceedings of the Fifth Social Media Mining for Health Applications (SMM4H) Workshop @ COLING.

Chapter 5: **Anne Dirkson** & Suzan Verberne (2019), *Transfer Learning for Health-related Twitter Data. Proceedings of the Fourth Social Media Mining for Health Applications* (SMM4H) Workshop & Task. Association for Computational Linguistics. 89-92.

Chapter 6: **Anne Dirkson**, Suzan Verberne and Wessel Kraaij (2021). *Breaking BERT: Understanding its Vulnerabilities for Named Entity Recognition through Adversarial Attack.* ArXiv. `https://arxiv.org/abs/2109.11308`

Chapter 7: **Anne Dirkson**, Suzan Verberne and Wessel Kraaij (2021), *FuzzyBIO: A proposal for Fuzzy Representation of Discontinuous Entities*, Proceedings of the 12th Health Text Mining and Information Analysis at EACL 2021.

Chapter 8: **Anne Dirkson**, Suzan Verberne, Gerard van Oortmerssen, Hans Gelderblom and Wessel Kraaij (2022). *How do others cope? Extracting coping mechanisms for adverse drug events from social media.* Journal of Biomedical Informatics.

Chapter 9: **Anne Dirkson**, Suzan Verberne, Wessel Kraaij, Gerard van Oortmerssen and Hans Gelderblom (2022). *Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers.* Scientific Reports, 12 (10317).

Chapter 10: Dide den Hollander, **Anne Dirkson**, Suzan Verberne, Wessel Kraaij, Gerard van Oortmerssen, Hans Gelderblom, Astrid Oosten, Anna K.L. Reyners, Neeltje Steeghs, Winette T.A. van der Graaf, Ingrid Desar and Olga Husson (2022). *Symptoms reported by Gastrointestinal Stromal Tumour (GIST) patients on imatinib treatment: combining questionnaire and forum data.* Supportive Care in Cancer.

Chapter 11: **Anne Dirkson**, Dide den Hollander, Suzan Verberne, Ingrid Desar, Olga Husson, Winette T.A. van der Graaf, Astrid Oosten, An Reyners, Neeltje Steeghs, Wouter van Loon, Hans Gelderblom and Wessel Kraaij (2022). *Sample bias in online patient generated health data of Gastrointestinal Stromal Tumor patients: Survey study.* JMIR Formative Research.

PART I:

PREPROCESSING MEDICAL SOCIAL MEDIA TEXT

Noise! Noise! Noise!
That's one thing he hated! The NOISE!

Dr. Suess, *How the Grinch stole Christmas!*

2

DATA-DRIVEN SPELLING CORRECTION

Edited from: **Anne Dirkson**, Suzan Verberne, Abeed Sarker & Wessel Kraaij (2019), *Data-Driven Lexical Normalization for Medical Social Media*, Multimodal Technologies and Interaction 3(3): 60.

The extraction of knowledge from medical social media is complicated by colloquial language use and misspellings. This noisiness can be reduced through lexical normalization: the transformation of non-standard text to a standardized vocabulary. Yet, lexical normalization of such data has not been addressed effectively.

To this end, we present a data-driven lexical normalization pipeline with a novel spelling correction module for medical social media. We find that our method significantly outperforms state-of-the-art spelling correction methods and can detect mistakes with an F_1 of 0.63 despite an extreme imbalance in the data.

Additionally, we present the first corpus for spelling mistake detection and correction in a medical patient forum. We make this corpus publicly available for the community to facilitate further research on this topic.

2.1. INTRODUCTION

In recent years, user-generated data from social media that contains information about health, such as patient forum posts or health-related tweets, has been used extensively for medical text mining and information retrieval (IR) [116]. This user-generated data encapsulates a vast amount of knowledge, which has been used for a range of health-related applications, such as the tracking of public health trends [267] and the detection of adverse drug events [266]. However, the extraction of this knowledge is complicated by non-standard and colloquial language use, typographical errors, phonetic substitutions, and misspellings [65, 229, 261]. This general noisiness of social media text is only aggravated by the complex medical domain [116].

The noisiness of medical social media can be reduced through lexical normalization: the transformation of non-standard text to a standardized vocabulary. Nonetheless, lexical normalization for medical social media has not been explored thoroughly. Medical lexical normalization methods (i.e., abbreviation expansion [210] and spelling correction [168, 230]) have mostly been developed for clinical records or notes. Although clinical records also contain many domain-specific abbreviations and misspellings, their contents are typically focused solely on the medical domain. In contrast, social media text typically covers a wider vocabulary including colloquial language and layman's terms for medical concepts [116, 352]. For medical social media, some recent studies have explored the related task of concept normalization (i.e., the mapping of tokens to standardized concept IDs in an ontology) [116].[1] Community-driven research on the topic has been boosted by the public release of relevant annotated data sets.[2] However, these available annotated data sets for concept normalization do not annotate misspellings explicitly and are thus not suitable for evaluating lexical normalization. As of yet, there are no publicly available annotated data sets for lexical normalization in medical social media.

Currently, the most comprehensive benchmark for lexical normalization in *general-domain* social media is the ACL W-NUT 2015 shared task[3] [19]. The current state-of-the-art system for this task is MoNoise [318]. However, this system is supervised and uses a lookup list of all replacement pairs in the training data as one of its important features. The training data from the task consists of 2,950 tweets with a total of 3,928 annotated non-standard words [19]. As extensive training data is unavailable for medical social media, such supervised systems cannot be employed in this domain. The best unsupervised system available is a modular pipeline with a hybrid approach to spelling, developed by Sarker [261]. Their pipeline also includes a customisable back-end module for domain-specific normalization. However, this back-end module relies on (i) a standard dictionary supplemented manually with domain-specific terms to detect mistakes and (ii) a language model of distributed word representations (word2vec) of generic Twitter data to correct these mistakes (for more detail see Section 2.3.2). For domains that have many out-of-vocabulary (OOV) terms compared to the available dictionaries and language models,

[1] For example, lexical normalization of 'pounding hed' would output 'pounding head', whereas concept normalization would aim to map it to the concept of Headache in a medical ontology such as SNOMED CT. A major difference between lexical and concept normalization is that the latter is constrained to terms of a predefined category (e.g., symptoms), whereas lexical normalization is unconstrained and can include any term.

[2] CADEC [151], PsyTAR [353] and the shared tasks of the SMM4H task [268, 335]

[3] `https://noisy-text.github.io/norm-shared-task.html`

such as medical social media, this is problematic.

Manual creation of specialized dictionaries is an unfeasible alternative: medical social media can be devoted to a wide range of different medical conditions and developing dictionaries for each condition (including laymen terms) would be very labor-intensive. Additionally, there are many different ways of expressing the same information and the language use in the forum evolves over time. In this chapter, we present an alternative: a corpus-driven spelling correction approach. Our method is designed to be conservative and to focus on precision to mitigate one of the major challenges of correcting errors in domain-specific data: the loss of information due to the erroneous correction of already correct domain-specific terms. Although dictionary-based retrieval will capture all mistakes, because any word that is not in the dictionary is considered a mistake, thereby attaining a high recall, its precision will be low. This is a result of words that are correct but not present in the dictionary as they will be erroneously marked as mistakes. Many domain-specific terms will fall in this category. In contrast, data-driven methods can capture patterns to recognize these non-mistakes as correct words and thereby improve precision, while recall could go down as these patterns might miss mistakes, for example because they are common. A data-driven detection approach will thus be more precise than dictionary-based retrieval.

In this chapter, we address two research questions:

1. To what extent can corpus-driven spelling correction reduce the out-of-vocabulary rate in medical social media text?

2. To what extent can corpus-driven spelling correction improve the accuracy of health-related classification tasks with social media text?

Our contributions are (1) an unsupervised data-driven spelling correction method that works well on specialized domains with many OOV terms without the need for a specialized dictionary[4] and (2) the first corpus for evaluating mistake detection and correction in a medical patient forum.[5]

The rest of the paper is organized as follows: In Section 2.2, we briefly review related work. In Section 2.3, we discuss the data sets we employ (Section 2.3.1) followed by a detailed description of our methodology (Section 2.3.2). In Section 2.4, we present our evaluation results, which are discussed further in Section 2.5. Lastly, in Section 2.6 we conclude our paper with final insights and an outline of future work.

2.2. RELATED WORK

2.2.1. CHALLENGES IN CORRECTING SPELLING ERRORS IN MEDICAL SOCIAL MEDIA

A major challenge for correcting spelling errors in small and highly specialized domains is a lack of domain-specific resources. This complicates the automatic creation of relevant dictionaries and language models. Moreover, if the dictionaries or language models are not domain-specific enough, there is a high probability that specialized terms will be

[4]Our lexical normalization pipeline is available at: `https://github.com/AnneDirkson/LexNorm`
[5]The corpus is available at `https://github.com/AnneDirkson/SpellingCorpus`

incorrectly marked as mistakes. Consequently, essential information may be lost as these terms are often key to knowledge extraction tasks (e.g., a drug name) and to specialized classification tasks (e.g., does the post contain a side effect of drug X?).

This challenge is further complicated by the dynamic nature of language on medical social media: in both the medical domain and social media novel terms (e.g., novel drug names) and neologisms (e.g., group-specific slang) are constantly introduced. Unfortunately, professional clinical lexicons are also unsuited for capturing the domain-specific terminology on forums, because laypersons and health care professionals express health-related concepts differently [348]. Another complication is the frequent misspellings of key medical terms, as medical terms are typically difficult to spell [352]. This results in an abundance of common mistakes in key terms, and thus, a large amount of lost information if these terms are not handled correctly.

2.2.2. LEXICAL NORMALIZATION OF SOCIAL MEDIA

The emergence of social networks and text messaging has redefined spelling correction to the broader task of lexical normalization, which may also involve tasks like abbreviation expansion [19]. In earlier research, text normalization for social media was mostly unsupervised or semi-supervised (e.g., [121]) due to a lack of annotated data. These methods often pre-selected and ranked correction candidates based on phonetic or lexical string similarity [120, 121]. Han and Baldwin [120] additionally used a trigram language model trained on a large Twitter corpus to improve correction. Although these methods did not rely on training data to correct mistakes, they did rely on dictionaries to determine whether a word *needed* to be corrected [120, 121]. The opposite is true for modern supervised methods: they do not rely on dictionaries but do rely on training data for both misspelling detection and correction. For instance, the best performing method at the ACL W-NUT shared task of 2015 used canonical forms in the training data to develop their own normalization dictionary [144]. Other competitive systems were also supervised and often used deep learning to detect and correct mistakes [175, 208] (for more detail on W-NUT systems see Baldwin et al. [19]). More recent competitive results for this shared task include MoNoise [318]. As mentioned, this system is also supervised and uses a lookup list of all replacement pairs in the training data as an important feature in their spelling correction. Since such specialized resources (appropriate dictionaries or training data) are not available for medical forum data, a method that relies on neither is necessary. We address this gap in this chapter.

Additionally, recent approaches (e.g., [261]) often make use of language models for spelling correction. Language models, however, require a large corpus of comparable text from the same genre and domain [261], which is a major obstacle for employing such an approach in niche domains. Since forums are often highly specialized, the resources that could capture the same language use are limited. Nevertheless, if comparable corpora are available, language models can contribute to effectively reducing spelling errors in social media [261] due to their ability to capture the context of words and to handle the dynamic nature of language.

Recent developments in the NLP field towards distributional language models based on byte-pair (BPE) or character-level encoding instead of word-level encoding call into question the need for prior spelling correction. In general, character-level models

are more robust to noise than word-level models, as they can exploit the remaining character structure regardless of errors. Niu et al. [218] recently developed a character-level attentional network model for medical concept normalization in social media which can alleviate the problem of out-of-vocabulary (OOV) terms by using a character-level encoding. Their model is robust to misspellings resulting from double characters, swapping of letters, adding hashtags and deletions.

However, firstly, the robustness to noise of character-based models appears to rely on whether they have been trained on noisy data [26, 132]. Otherwise, they are prone to breaking when presented with synthetic or natural noise [26, 132]. Thus, if sufficiently large amounts of data with similar types of noise are available, character-based models may negate the need for spelling correction. However, in domains lacking such resources, spelling correction in the pre-processing stage is still needed. Secondly, character-based models have computational disadvantages: their computational complexity is higher and it becomes harder to model long-range dependencies [132]. Alternatively, word embeddings designed to be robust to noise [196] could be used. Yet, also for this method, sufficiently large amounts of comparable noisy data are necessary. To provide an indication, Malykh et al. [196] use the Reuters CV-1 corpus consisting of 800,000 news stories ranging from a few hundred to several thousand words in length [177] to generate their robust English word embeddings.

2.2.3. LEXICAL NORMALIZATION OF CLINICAL RECORDS

Like medical social media, clinical notes made by doctors are user-generated and noisy. In fact, Ruch et al. [255] reported about one spelling error per five sentences. Yet, most normalization research for clinical notes has focused on concept normalization instead of lexical normalization [116]. A prominent shared task for concept normalization of clinical notes is Task 2 of the CLEF e-Health workshop in 2014. Its aim was to expand abbreviations in clinical notes by mapping them to the UMLS database [210]. The best system by Wu et al. [343] applied four different trained tagging methods depending on the frequency and ambiguity of abbreviations. Unfortunately, the abbreviations used by doctors are not the same as the ones used by patients, and thus these methods do not transfer.

To correct misspellings in clinical notes, Lai et al. [168] developed a spell checker based on the noisy channel model by Shannon [273]. Noisy channel models interpret spelling errors as distortions of a signal by noise. The most probable message can then be calculated from the source signal and noise models. This is how spelling correction is modeled traditionally [64]. Although their correction accuracy was high, their method relied on an extensive dictionary compiled from multiple sources to detect mistakes. Similarly, the method by Patrick et al. [230] also used a compiled dictionary for detecting errors. For correction, Patrick et al. [230] used edit distance-based rules to generate suggestions which were ranked using a trigram model. Fivez et al. [110] was the first to leverage contextual information to correct errors in clinical records. They developed an unsupervised, context-sensitive method that used word and character embeddings to correct spelling errors. Their approach outperformed the method proposed by Lai et al. [168] for the benchmark MIMIC-III [146]. However, they did not perform any mistake detection, as they simply tried to correct the annotated misspellings of MIMIC-III. In conclusion, the methods developed for spelling correction in clinical records either only

focus on correction or rely solely on extensive, compiled dictionaries to find mistakes. Therefore, they are not applicable in domains lacking such resources.

2.3. MATERIALS AND METHODS

2.3.1. DATA

Data collection For evaluating spelling correction methods, we use an international patient forum for patients with Gastrointestinal Stromal Tumor (GIST). It is moderated by GIST Support International (GSI). This data set was donated to Dr. Verberne by GSI in 2015. We use a second cancer-related forum to assess the generalisability of our methods: a sub-reddit community on cancer, dating from 16/09/2009 until 02/07/2018.[6] It was scraped using the Pushshift Reddit API.[7] The data was collected by looping over the timestamps in the data. This second forum is roughly four times larger than the first in terms of the number of tokens (See Table 2.1).

Table 2.1: Raw data without punctuation. IQR: Inter-quartile range

	GIST forum	Reddit forum
# Tokens	1,255,741	4,520,074
# Posts	36,277	274,532
Median post length (IQR)	20 (35)	11 (18)

Data annotation Spelling mistakes were annotated for 1000 randomly selected posts from the GIST data. Each token was classified as a mistake (1) or not (0) by the first author. For the first 500 posts, a second annotator checked if any of the mistakes were false positives. In total, 99 of the 109 non-word spelling errors were annotated for correction experiments. The remaining 10 errors were found later during error detection experiments and were therefore only included in these experiments. The corrections for the 53 unique mistakes present in the first 500 posts were annotated individually by two annotators, of which one was a GIST patient and a forum user. Annotators were provided with the complete post to determine the correct word. The initial absolute agreement was 89.0%. If a consensus could not be reached, a third assessor was used to resolve the matter. The remaining mistakes were annotated by the first author. For the correction 'reoccurrence', the synonym 'recurrence' was also considered correct. As far as we are aware, no other spelling error corpora for this domain are publicly available.

To tune the similarity threshold for the optimal detection of spelling mistakes, we used 60% of the annotated data as a development set. The split was done per post and stratified on whether a post contained mistakes or not. Since the data is extremely unbalanced, we balanced the training data to some extent by combining the mistakes with a ten-fold of random correct words with the same word length distribution (see Table 2.2). These words were not allowed to be numbers, punctuation, or proper nouns, because these are ignored by our error detection process. The development set was split in a stratified manner into 10 folds for cross-validation.

[6] http://www.reddit.com/r/cancer

[7] https://github.com/pushshift/api

Table 2.2: Annotated data for spelling detection experiments. *excluding punctuation, numbers and proper nouns.

	Mistakes (%)	Total word count*
Training set	57 (9.1%)	627
Test set	45 (0.42%)	10760

Corpus for calculating weighted edit matrix Since by default all edits are weighted equally when calculating Levenshtein distance, we needed to compute a weighted edit matrix in order to assign lower costs and thereby higher probabilities to edits that occur more frequently in the real world. We based our weighted edit matrix on a corpus of frequencies for 1-edit spelling errors compiled by Peter Norvig.[8] This corpus is compiled from four sources: (1) a list of misspellings made by Wikipedia editors, (2) the Birkbeck spelling corpus, (3) the Holbrook corpus and (4) the Aspell error corpus.

Specialized vocabulary for OOV estimation in cancer forums To be able to calculate the number of out-of-vocabulary terms in the two cancer forums, a specialized vocabulary was created by merging the standard English lexicon CELEX [46] (73,452 tokens), the NCI Dictionary of Cancer Terms [215] (6,038 tokens), the generic and commercial drug names from the RxNorm [314] (3,837 tokens), the ADR lexicon used by Nikfarjam et al. [217] (30,846 tokens) and our in-house domain-specific abbreviation expansions (DSAE) (42 tokens) (see 2.3.2 for more detail). As many terms overlapped with those in CELEX, the total vocabulary consisted of 118,052 tokens (62.2% CELEX, 5.1% NCI, 26.1% ADR, 6.5% RxNorm and <0.01% DSAE).

2.3.2. METHODS

Preprocessing URLs and email addresses were replaced by the strings -URL- and -EMAIL- using regular expressions. Furthermore, text was lower-cased and tokenized using NLTK. The first modules of the normalization pipeline of Sarker [261] were employed: converting British to American English and normalizing generic abbreviations (see Figure 2.1). Some forum-specific additions were made: Gleevec (British variant: Glivec) was included in the British-American spelling conversion, one generic abbreviation expansion that clashed with a domain-specific one was substituted (i.e., 'temp' defined as *temperature* instead of *temporary*), and two problematic medical terms were removed from the slang dictionary (i.e., 'ill' corrected to 'i'll' and 'chronic' corrected to 'marijuana').

Moreover, the abbreviations dictionary by Sarker [261] was lower-cased. As apostrophes in contractions are frequently omitted in social media posts (e.g., im instead of i'm), we expanded contractions to their full form (e.g., i am). Firstly, contractions with apostrophes were expanded and subsequently those without apostrophes were expanded only if they were not real words according to the CELEX dictionary. Lastly, domain-specific abbreviations were expanded with a lexicon of domain-specific abbreviation expansions (DSAE). The abbreviations were manually extracted from 500 randomly selected posts of the GIST forum data. This resulted in 47 unique abbreviations. Two annotators, of which

[8] http://norvig.com/ngrams/count_1edit.txt

one was a domain expert, individually determined the correct expansion term for each abbreviation, with an absolute agreement of 85.4%. Hereafter, they agreed on the correct form together.[9]

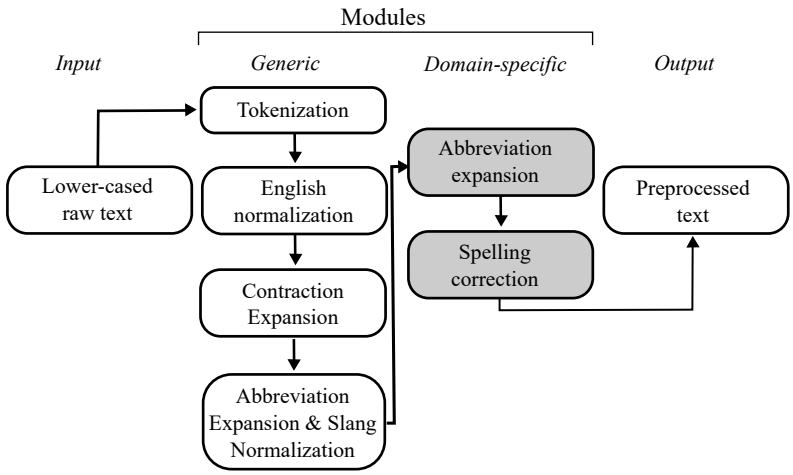


Figure 2.1: Sequential processing pipeline

Spelling correction

Baseline methods We used the method by Sarker [261] as a baseline for spelling correction. Their method combines normalized absolute Levenshtein distance with Metaphone phonetic similarity and language model similarity. For the latter, distributed word representations (skip-gram word2vec) of three large Twitter data sets were used. In this chapter, we used the largest available version of the DIEGO LAB Drug Chatter Corpus (around 1 Billion tweets) [263], as it was the only health-related corpus of the three. We also use a purely data-driven spelling correction method for comparison: Text-Induced Spelling Correction (TISC) developed by Reynaert [248]. It compares the anagrams of a token to those in a large corpus of text to correct mistakes. These two methods are compared with simple absolute and relative Levenshtein distance and weighted versions of both. To evaluate the spelling correction methods, the accuracy (i.e., the percentage of correct corrections) was used. The weights of the edits for weighted Levenshtein distance were computed using the log of the frequencies of the Norvig corpus. We used the log to ensure that a 10x more frequent error does not become 10x as cheap, as this would make infrequent errors too improbable. In order to make the weights inversely proportional to the frequencies and scale the weights between 0 and 1 with lower weights signifying lower

[9] This abbreviations lexicon is shared at `https://github.com/AnneDirkson/LexNorm`

costs for an edit, the following transformation of the log frequencies was used: Weight Edit Distance = $\frac{1}{1+log(frequency)}$.

Correction candidates Spelling correction methods were first compared using the terms from the specialized vocabulary for cancer forums (see section 2.3.1) as correction candidates. This enables us to evaluate the methods independently of the vocabulary present in the data. Hereafter, we assessed the impact of using correction candidates from the data itself, since our aim is to develop a method that is independent of manually compiled lexicons. Numbers, proper nouns, and punctuation are ignored as possible correction candidates.

We inspected whether restricting the pool of eligible correction candidates based on their corpus frequency relative to that of the token aids correction. We use relative corpus frequency thresholds ranging from at least 0 times (no restriction) to 10 times more frequent than the token. The underlying idea is that the correct word will be used more often than the incorrect word and by restricting the candidates we prevent implausible but similar words from hindering correction. This, for instance, prevents mistakes from being corrected by other similar and roughly equally frequent mistakes. A relative, instead of absolute, threshold that depends on the frequency of the mistake enables us to also correct mistakes even if they occur more commonly (e.g., misspellings of a complex medication name). Candidates are considered in order of frequency. Of the candidates with the highest similarity score, the first is selected.

We tried two different approaches to further improve correction by altering the pool of correction candidates. Firstly, we tested whether prior lemmatization of the spelling errors with or without prior lemmatization of the correction candidates could improve spelling correction. Secondly, we investigated the effect of imposing an additional syntactic restriction on the correction candidates, namely only allowing those with the same Part-of-Speech tag at least once in the data or the same surrounding POS tags to the left and right (i.e., the POS context) at least once in the data. McNemar tests were used to test whether the predictions of various correction methods are significantly different. In all follow-up experiments, correction candidates were derived from the respective data set and constrained by the optimal relative corpus frequency threshold.

Improving the baseline method For the best baseline method with data-driven candidates, we explored whether the context of the token could aid the correction further using (1) language models of the forum itself or (2) publicly available distributed and sequential language models of health-related social media data. This last category includes the distributed word2vec (dim= 400) and sequential trigram language models developed by Sarker and Gonzalez [263] and the distributed word2vec (dim = 200) HealthVec model developed by Miftahutdinov et al. [207]. The models by Sarker and Gonzalez [263] are based on around 1 billion Twitter posts derived from user timelines where at least 1 medication is mentioned. A smaller version of this language model is used in the current state-of-the-art normalization pipeline for general social media [261].[10] The HealthVec model is based on the Health Dataset consisting of around 2.5 million

[10] Language models can be obtained from: `https://data.mendeley.com/datasets/dwr4xn8kcv/3`

user comments from six web resources: WebMD, Askapatient, patient.info, Dailystrength, drugs.com, and product reviews from the Amazon Dataset.[11] Besides employing these language models, we explored whether adding double Metaphone phonetic similarity [233] improves correction. Phonetic similarity is a measure of how phonetically similar an error is to the potential correction candidate.

The best baseline method was combined with these similarity measures (i.e., phonetic similarity or the similarity based on a language model) in a weighted manner with weights ranging from 0 to 1 with steps of 0.1. The inverse weight was assigned to the baseline similarity measure. For all language models, if the word was not in the vocabulary, then the model similarity was set to 0, essentially rendering the language model irrelevant in these cases. To investigate the impact of adding these contextual measures, Pearson's r is used to calculate the correlation between the correction accuracy and the assigned weight.

Correcting Concatenation Errors If a word is not in the Aspell dictionary[12], but is also not a spelling mistake, our method checks if it needs to be split into two words. It is split only if it can be split into two words of at least 3 letters which both occur more in the corpus more frequently than the relative corpus frequency boundary. For each possible split, the frequency of the least frequent word is considered. The most plausible split is the one for which this lower frequency is the highest (i.e., the least frequent word occurs the most). Words containing numbers (e.g., 3months) are the exception: they are split so that the number forms a separate word.

Spelling mistake detection We manually constructed a decision process, inspired by the work by Beeksma et al. [25], for detecting spelling mistakes (See Figure 2.7). The optimal relative corpus frequency threshold determined for spelling correction in our earlier experiments is adopted. On top of this threshold, the decision process uses the similarity of the best candidate to the token to identify mistakes. If there is no similar enough correction candidate available, then the word is more likely to be a unique domain-specific term we do not wish to correct than a mistake. The minimum similarity threshold is optimized with a 10-fold cross validation grid search from 0.40 to 0.80 (steps of 0.02). The loss function used to tune the parameters was the $F_{0.5}$ score, which places more weight on precision than the F_1 score. We believe it is more important to not alter correct terms than to retrieve incorrect ones. Candidates are considered in order of frequency. Of the candidates with the highest similarity score, the first is selected. The error detection automatically labels numbers, punctuation, proper nouns, and words present in the Aspell dictionary as correct. We used the word list 60 version of the Aspell dictionary, as is recommended for spelling correction. To verify that medication names were not being classified as proper nouns and thereby excluded from spelling correction, we checked the part-of-speech tags of the most common medication for GIST patients (gleevec) and two of its common misspellings (gleevic and gleevac). For gleevec, 81.4% of the mentions were classified as nouns (NN). The next two largest categories were adjectives (JJ) (7.2%), plural nouns (NNS) (4.7%) and verbs (VB) (3.9%). The remaining 2.8% were divided over 10 POS-tags (ranging from 0.6% to 0.0005%). Most importantly, none were classified as

[11] Available at: `http://jmcauley.ucsd.edu/data/amazon`
[12] Available at: http://Aspell.net/

proper nouns (NNP or NNPS). Similarly, gleevic and gleevac were labeled as nouns (NN) 78.1% and 83.9% of the time and neither was ever labelled as a proper noun. For gleevic, the remaining cases were divided amongst plural nouns (11.4%), adjectives (8.3%) and verbs (2.2%). For gleevac, the remainder was divided between verbs (11.9%) and adjectives (4.2%).

We compared our optimized decision process with and without concatenation error detection (see Section 2.3.2) with error detection using two commonly used dictionaries, CELEX [46] and Aspell, with Microsoft Word and with TISC, another data-driven detection method [248]. Significance was calculated with McNemar tests. Any mistakes overlapping between the training and test set were not included in the evaluation.

Impact of the corpus size on detection To measure the influence of the size of the corpus on spelling mistake detection, we varied the size of the corpus from which correction candidates are derived. The token frequencies of errors and candidates were both calculated using this corpus. Therefore, the frequencies of mistakes and potential corrections would vary and we could estimate for each corpus size how much the error detection in 1000 posts would change. We used Jaccard similarity to measure the overlap between the error predictions of each possible combination of two different corpus sizes.

As our relative corpus frequency threshold is a minimal threshold, bigger corpora and thus larger differences between the token frequency of the error and that of the correct variant would not pose a problem. Consequently, we randomly selected posts to artificially downsize our two cancer forums exponentially. We used sizes ranging from 1000 posts to all forum posts. The 1000 posts for which errors were detected were always included in the corpus. For the GIST forum, we used the 1000 annotated posts.

Impact of the degree of noisiness of the data To investigate the impact of the level of noise in the data on spelling correction and detection, we simulated data sets with varying proportions of misspellings. As our method was designed on data with few errors (< 1% in our sample), this will help us to understand to what extent our method can generalize to more noisy user-generated data. We generated artificial data by altering the number of misspellings in two cancer-related fora.

In line with the work by Niu et al. [218], we generated artificial noise typical of social media text by (i) deleting a single letter, (ii) doubling a letter and (iii) swapping two adjacent letters. Niu et al. [218] also added hashtags to words, but as this is only relevant for Twitter we omit this transformation. Words are randomly selected based on a predetermined probability of occurrence (1,2,3,4,8 and 16%). Which letter is removed or swapped in the word is dependent on the normalized likelihood of a deletion or swap occurring in real-word data. We use the normalized log frequencies of the Norvig corpus [219]. Additionally, the log frequencies were normalized per word to sum to 1. Which letter is doubled is randomly selected, as frequencies for such operations are not available. We evaluated the spelling correction and detection for each forum with the average of three runs of 1000 randomly selected posts with 3 different seeds.

Effect on OOV rate The percentage of out-of-vocabulary (OOV) terms is used as an estimation of the quality of the data: less OOV-terms and thus more in-vocabulary (IV)

terms are a proxy for cleaner data. As the correction candidates are derived from the data itself, one must note that words that are not part of Aspell may also be transformed from IV to OOV. OOV analysis was done manually.

External validation To evaluate the impact of lexical normalization as a preprocessing step on the performance of separate downstream tasks, we perform extrinsic evaluation of our pipeline by running six text classification experiments. We obtained six publicly available health-related Twitter data sets ranging in size from 588 to 16,141 posts (see Table 2.3). As can be seen in Table 2.3, the data sets also have varying degrees of imbalance. It is not uncommon for social media data sets to be highly imbalanced and thus we investigate whether the impact of spelling correction is influenced by imbalance. The data sets were retrieved from the data repository of Dredze[13] and the shared tasks of Social Media Mining for Health Applications (SMM4H) workshop 2019.[14]

Text classification was performed before and after normalization using default sklearn classifiers: Stochastic Gradient Descent (SGD), Multinomial Naive Bayes (MNB) and Linear Support Vector Machines (SVC). Unigrams were used as features. A 10-fold cross-validation was used to determine the quality of the classifiers and a paired t-test was applied to determine significance of the absolute difference. Only the best performing classifier is reported per data set. For the shared tasks of the SMM4H workshop, only the training data was used.

Table 2.3: Six classification data sets of health-related Twitter data. *SMM4H: Social Media Mining for Health Applications workshop

Data set	Task	Size	Positive Class
Task 1 SMM4H 2019*	Presence adverse drug reaction	16,141	8.7%
Task 4 SMM4H 2019* Flu vaccine	Personal health mention of flu vaccination	6,738	28.3%
Flu Vaccination Tweets [141]	Relevance to flu vaccination	3,798	26.4%
Twitter Health [231]	Relevance to health	2,598	40.1%
Task4 SMM4H 2019* Flu infection	Personal health mention of having flu	1,034	54.4%
Zika Conspiracy Tweets [91]	Contains pseudo-scientific information	588	25.9%

To evaluate our method on generic social media text, we used the test set of the ACL W-NUT 2015 task [19]. The test set consists of 1967 tweets with 2024 one-to-one, 704 one-to-many, and 10 many-to-one mappings. We did not need to use the training data, as our method is unsupervised. We omitted the expansion of contractions from our normalization pipeline for the W-NUT task, because expanding contractions was not part of the goals of the task. Error analysis was done manually on the 100 most frequent errors.

[13] http://www.cs.jhu.edu/~mdredze/data/

[14] https://healthlanguageprocessing.org/smm4h/challenge/

2.4. RESULTS

In this section, we will report the distribution of spelling errors in our corpus (2.4.1), the evaluation of spelling correction (2.4.2) and detection methods (2.4.3) on our spelling corpus and the impact of corpus size (2.4.4) and the level of noise in the corpus (2.4.5) on the efficacy of our method. Hereafter, we assess the impact of our method on the OOV rate in two cancer-related fora (2.4.6) and on classification accuracy of six health-related Twitter benchmarks (2.4.7). We also evaluate the performance of our method on the W-NUT shared task for generic social media normalization (2.4.7).

2.4.1. ERROR DISTRIBUTION

Spelling errors can be divided into non-word errors (i.e., errors that are not valid words) and real-word errors (i.e., errors that result in another valid word) [164]. Incorrect concatenations and splits can be either. For example, 'scan' to 'scant' is a real word error whereas 'side effects' to 'sideeffects' is a non-word error. We focus on correcting non-word errors, as we are not interesting in correcting syntactic or semantic errors [164].

Nonetheless, we investigate the prevalence of these error types in the data to gain insight into which types of errors are made in medical social media text. As can be seen in Table 2.4, our corpus of 1000 medical posts from the GIST forum mainly contains non-word errors. Moreover, non-word errors contain the highest percentage of medical misspellings (47.7%). Comparatively, only 20% of real word errors are medical terms. Most posts do not contain any errors (see Figure 2.2), but for those that do, there was in most cases only one error per post.

Table 2.4: Error distribution in 1000 GIST posts

Error type	Non-word	Incorrect splits	Incorrect concatenations	Real word
Amount	109	17	24	30
Non-Medical/Medical	57/52	25/5	14/3	18/6
Percentage of tokens	0.32%	0.05%	0.07%	0.09%
Example mistake	gleevac	gall bladder	sideeffects	scant
Example correction	gleevec	gallbladder	side effects	scan

2.4.2. SPELLING CORRECTION

The normalization step prior to spelling correction (see Figure 2.1) corrected 12 of the 99 spelling errors, such as 'feelin' to 'feeling'. These errors are all on the fuzzy boundary between spelling errors and slang. Thus, spelling correction experiments were performed with the remaining annotated 87 spelling errors.

The state-of-the-art method for generic social media by Sarker [261] performs poorly for medical social media: it corrects only 19.3% of the mistakes (see Table 2.5). In fact, it performed significantly worse ($p < 0.0001$) than all edit distance based methods. Computationally, it is also much slower (see Table 2.6). A second established data-driven approach, TISC [248], performed even more poorly (14.8%). TISC was also significantly worse than all edit-based methods ($p < 0.0001$). Relative weighted edit distance performed the best with an accuracy of 68.2%. The theoretical upper bound

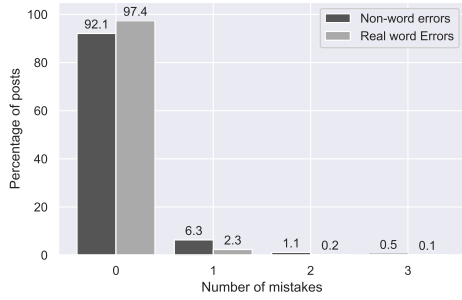


Figure 2.2: Distribution of non-word and real word errors across posts in the GIST forum.

for accuracy was 92.0%, because not all corrections occur in the specialized dictionary. Examples of corrections can be seen in Table 2.7.

Table 2.5: Correction accuracy using a specialized vocabulary. AE: absolute edit distance. RE: relative edit distance. WAE: weighted absolute edit distance. WRE: weighted relative edit distance. *Only the best corpus frequency threshold is reported

Source of candidates	Ceiling	**AE**	**RE**	**WAE**	**WRE**	**Sarker**	**TISC**
Specialized vocabulary	92.0%	58.0%	64.7%	63.3%	68.2%	19.3%	14.8%
GIST forum text*	97.6%	**73.9%**	**73.9%**	70.4%	72.7%	44.3%	-

Table 2.6: Mean computation time over 5 runs

AE	**RE**	**WAE**	**WRE**	**Sarker**
13.36 ms	14.04 ms	29.45 ms	32.00 ms	904.33 ms

However, when using candidates derived from the data itself, unweighted absolute and relative edit distance perform the best. Relative edit distance accurately corrects 73.9% of all mistakes at a relative corpus frequency threshold (θ) of 9, while absolute edit distance does so at a θ of 2 to 5 (See Table 2.5 and Figure 2.3). A θ of 9 means that candidates are only considered plausible if they occur 9 times more frequently than the spelling error. We elect to use relative edit distance, because it is more fine-grained than absolute edit distance, especially for short words. Using data-driven candidates increases the theoretical upper bound from 90.2% to 97.6%. This showcases the limitations of using dictionaries for correction.

Nonetheless, simply using all words from the data as possible candidates (i.e., a corpus frequency threshold of 0) for every spelling error results in a very low correction accuracy (see Figure 2.3). However, imposing the restriction that the corpus frequency of a viable correction candidate must be at least double (2x) that of the mistake, significantly improves correction ($p < 0.0001$) for all correction methods. In that case, for a mistake occurring 10 times, only words occurring at least 20 times are considered. Thus, the

Table 2.7: Corrections by different methods with candidates from a specialized vocabulary. *Gleevec and Sutent are important medications for GIST patients.

Mistake	Correction	AE	RE	WAE	WRE	Sarker	TISC
gleevac	gleevec*	**gleevec**	**gleevec**	**gleevec**	**gleevec**	colonic	gleevac
stomack	stomach	**stomach**	**stomach**	smack	**stomach**	smack	smack
ovari	ovary	**ovary**	**ovary**	**ovary**	**ovary**	ova	atari
sutant	sutent*	mutant	mutant	**sutent**	**sutent**	mutant	dunant
mestastis	metastasis	miscasts	**metastasis**	**metastasis**	**metastasis**	miscasts	mestastis

assumption that corrections are more common than mistakes appears to hold true. However, at any threshold all edit distance based methods still significantly ($p < 0.001$) outperform the state-of-the-art method [261], in line with previous results (Table 2.5). Examples of corrections with data-driven candidates are reported in Table 2.8.

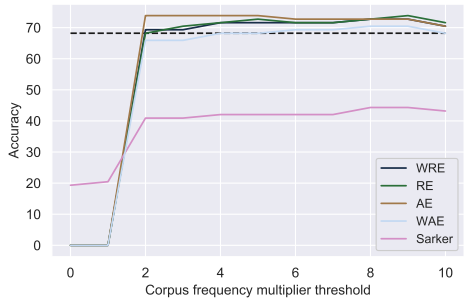


Figure 2.3: Correction accuracy of unique mistakes using correction candidates from the data at various minimum relative corpus frequency thresholds. Dotted line indicates the best correction accuracy using dictionary-derived candidates.

Table 2.8: Corrections by different methods with data-driven candidates. AE: absolute edit distance. RE: relative edit distance. WAE: weighted absolute edit distance. WRE: weighted relative edit distance.

Mistake	Correction	AE	RE	WAE	WRE	Sarker
gleevac	gleevec	**gleevec**	**gleevec**	**gleevec**	**gleevec**	**gleevec**
stomack	stomach	**stomach**	**stomach**	**stomach**	**stomach**	stuck
ovari	ovary	**ovary**	**ovary**	**ovary**	**ovary**	ovarian
sutant	sutent	**sutent**	**sutent**	**sutent**	**sutent**	mutant
mestastis	metastasis	metastis	metastis	metastis	metastis	metastis

The accuracy of the best baseline method, namely relative edit distance with a θ of 9, is unaffected by prior lemmatization of the spelling errors (see Table 2.9). It thus appears that if prior lemmatization can correct the error, our method automatically does so. In contrast, additional lemmatization of their corrections and of the correction candidates significantly reduces accuracy ($p = 0.021$ and $p = 0.011$) compared to omitting prior

lemmatization. Thus, lemmatization of the data or candidates prior to spelling correction is not recommended.

NoLemmatization	LemmatizedInput	+ LemmatizedOutput	+ LemmatizedCandidates
73.6%	73.6%	64.7%	67.0%

Table 2.9: Effect of lemmatization of the errors (LemmatizedInput), their corrections (LemmatizedOutput) and correction candidates (LemmatizedCandidates) on spelling correction accuracy using RE (θ= 9)

Adding weighted phonetic similarity Previous research has shown that when users are faced with the task of writing an unfamiliar, complex word like a drug name, they tend to revert to phonetic spelling [235]. Therefore, we investigate whether adding a weighted phonetic component may improve correction. This is not the case: The weight assigned to phonetic similarity has a strong negative correlation (-0.92) with the correction accuracy ($p < 0.0001$) (see Figure 2.4). This suggests that such phonetic errors are already captured by our frequency-based method.

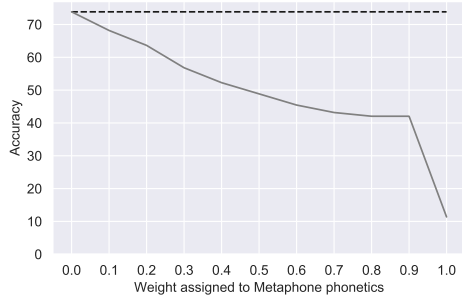
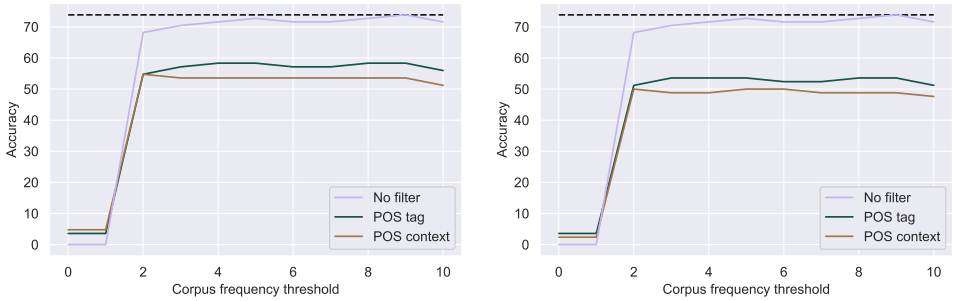


Figure 2.4: Correction accuracy with additional weighted double Metaphone phonetic similarity. Dotted line indicates the best accuracy with relative edit distance alone.

Adding weighted contextual similarity Previous work has indicated that the context of spelling mistakes might be helpful to improve spelling correction [110]. Since domain-specific resources are scarce, one potential approach is to use the contextual information present in the corpus itself. Based on work by Beeksma et al. [25], we tried to use the Part-of-Speech (POS) tag of the error or the POS tags of its neighbors to constraint correction candidates. However, as can be seen in Figure 2.5, adding these constraints reduces correction accuracy, although not significantly. Aside from some additional errors, using POS context as a constraint results in identical errors as enforcing a similar POS tag for potential correction candidates, regardless of whether NLTK or Spacy is used.

As many modern methods use language models to aid spelling correction [261], we also examine whether we can leverage contextual information by using language models of the corpus itself to improve correction accuracy. For both Word2vec and FastText



(a) NLTK POS tags

(b) Spacy POS tags

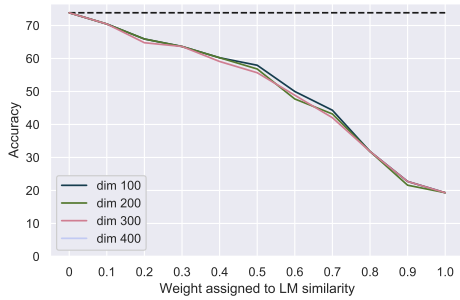
Figure 2.5: Correction accuracy of spelling mistakes with additional POS tag filters. Dotted lines indicate the best accuracy with relative edit distance alone.

distributed models of the data, we find that the higher the weight assigned to the language model similarity, the more the accuracy drops. This inverse correlation is significant and almost equal to -1 for all dimensionalities ($p < 0.000001$) (see Figure 2.6a and 2.6b). Our data is possibly too sparse to place contextual constraints on the correction candidates or to employ language model similarity in this manner. It is also too small for building a sequential trigram model [327].

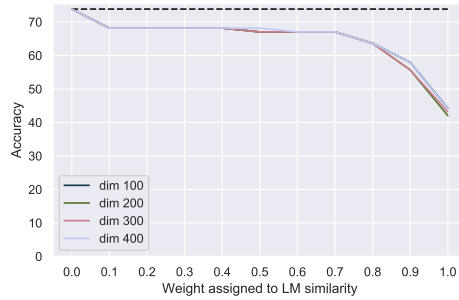
Alternatively, we can employ more generic language models based on medical social media, but not specific to a particular disease domain. We find that a distributed language model based on a collection of health-related tweets, the DIEGO Drug chatter corpus [263], does not manage to improve correction accuracy (see Figure 2.6c). Nevertheless, a sequential trigram model based on this same Twitter corpus does improve correction accuracy with 2.2% point to 76.1% at a weight of 0.6 (see Figure 2.6c). The weight assigned to the probability of a trigram with the correction in place of the error is positively correlated ($r = 0.58$) with the correction accuracy. However, the HealthVec distributed language model can improve the correction accuracy up to 79.5% at a weight of 0.6 (see Figure 2.6d). Overall, its assigned weight is also positively correlated ($r = 0.63$) with the correction accuracy. Table 2.10 shows that adding the HealthVec model mostly improves accuracy for non-medical errors (e.g., 'explane') and for medical errors for which it is difficult to determine whether they should be singular or plural (e.g., 'ovarie and surgerys'). One medical term (i.e., 'surgerys') is no longer corrected accurately. We opt to employ this weighted method due to its higher overall accuracy, but one could opt to not include the HealthVec model depending on the importance of non-medical terms for the downstream task.

2.4.3. SPELLING MISTAKE DETECTION

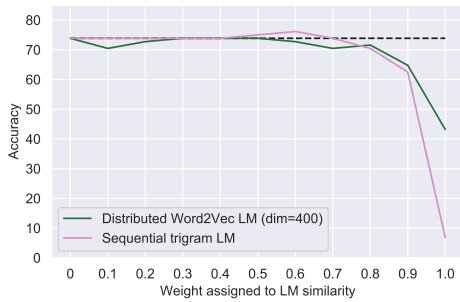
A grid search results in an optimal similarity score threshold of 0.76. As higher similarity scores indicate that tokens are more dissimilar, this means that if the best correction candidate has a higher similarity score than this threshold, the token is not corrected (see Figure 2.7). This combination attains the maximum $F_{0.5}$ score for 8 of 10 folds. For the other two folds, 0.74 was optimal. See Figure 2.7 for the tuned decision process. On the



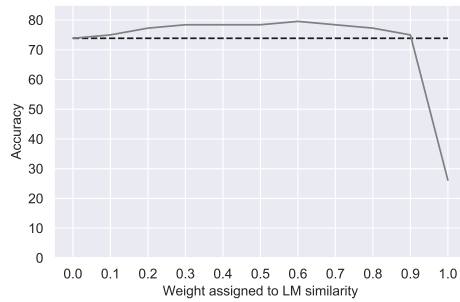
(a) Word2Vec distributed language model trained on the GIST forum data



(b) Fasttext language model trained on the GIST forum data



(c) Language models trained on the DIEGO Drug chatter corpus



(d) Distributed Word2vec (dim= 200) model (HealthVec) trained on Health Dataset

Figure 2.6: Correction accuracy of spelling mistakes with additional weighted language model (LM) similarity. Weight of the LM similarity is the inverse of the weight of the relative edit distance. Dotted line indicates the best accuracy with relative edit distance alone.

Table 2.10: Changes in corrections when HealthVec is added (weight = 0.6) to the relative edit distance (weight = 0.4) with $\theta = 9$. LM = language model.

	Error	Correct word	Correction	
			Without LM	With LM
Improved	alse	else	false	else
	lm	im	am	im
	esle	else	resolve	else
	explane	explain	explained	explain
	ovarie	ovary	ovary	ovaries
	surgerys	surgeries	surgeries	surgery
Missed	surgerys	surgery	surgery	surgury

test set, our method attains a significantly higher precision ($p < 0.0001$) and $F_{0.5}$ score ($p < 0.0001$) than all other detection methods (see Table 2.11). Our method does attain a slightly lower recall than dictionary-based methods, although its recall is very high at 0.91. Adding concatenation correction to our method improves recall and precision by 0.05 and 0.01, respectively. See Table 2.12 for some examples of errors made by our decision process and the corrections our method will output.

Although the recall of generic dictionaries is maximal at 1.0, their precision is low (0.11 and 0.26). Both are logical: The high recall is a result of dictionary-based methods classifying all terms *not* included in the dictionary as mistakes, which will include all non-word errors, whereas the low precision is a result of the misclassification of correct domain-specific terms that are not included in the dictionary. Aspell outperforms CELEX due to its higher coverage of relevant words such as 'oncologist', 'metastases' and 'facebook'. Microsoft Word and TISC perform the worst overall: their precision is low but they also have a lower recall than both dictionary-based methods and our method.

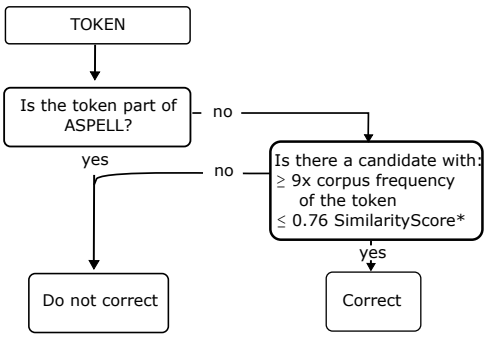


Figure 2.7: Decision process. *SimilarityScore = 0.6 * LM similarity + 0.4 * RE

Table 2.11: Results for mistake detection methods on the test set

Method	Mistakes found	Recall	Precision	$F_{0.5}$	F_1
CELEX	395	**1.0**	0.11	0.13	0.20
Aspell dictionary	163	**1.0**	0.26	0.31	0.42
TISC	270	0.74	0.12	0.14	0.21
Microsoft word	395	0.88	0.10	0.12	0.18
Our method (RE = 0.76)	90	0.91	0.46	0.51	0.61
Our method (RE= 0.76) + ConcatCorrection	92	0.96	**0.47**	**0.52**	**0.63**

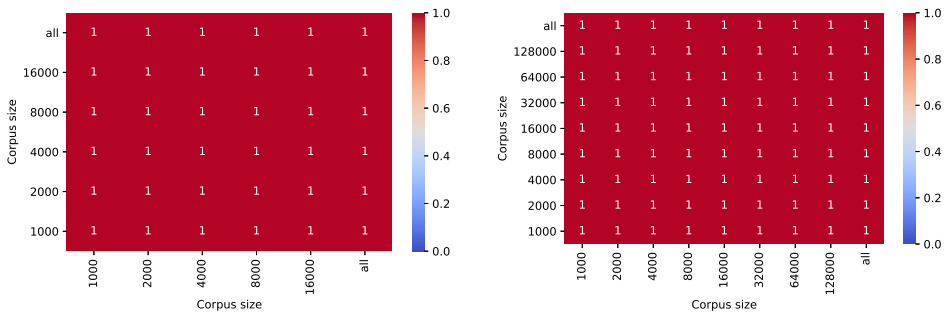
2.4.4. IMPACT OF CORPUS SIZE

Despite the fact that a relative corpus frequency threshold is more robust to different corpus sizes than an absolute one, it is likely that the ratio between tokens and their corrections will vary if the corpus size becomes smaller. Thus, we investigated to what extent the multiplication factor of 9 would be robust to such ratio changes.

Table 2.12: Examples of false positives and negatives of our error detection method.

	Mistakes (their corrections with our method)			
False positives	intolerances (intolerant)	resected (removed)	reflux (really)	condroma (syndrome)
False negatives	istological (histological)	vechile (vehicle)		

Figure 2.8 shows that our threshold is highly robust to corpus size with maximal Jaccard similarity (1.0) for all comparisons. Figure 2.9 demonstrates this with an example of one common ('gleevac') and one uncommon misspelling ('gllevec') for the medication Gleevec. The corpus frequency for each misspelling relative to the corpus size is shown with unbroken lines. The minimum corpus frequency threshold for correction candidates of each misspelling is indicated with dotted lines of the same color for the range of corpus sizes. Irrespective of the corpus size, the correct variant 'gleevec' (the purple line) remains above the minimum corpus frequency (i.e., the dotted lines) for the complete range of corpus sizes.



(a) GIST forum of 36,277 posts

(b) Reddit forum of 274,532 posts

Figure 2.8: Stability of error detection in 1000 posts with varying corpus size

2.4.5. IMPACT OF THE DEGREE OF NOISINESS OF THE DATA

As our method was designed on data with few errors (< 1% in our sample), we investigate to what extent our method can generalize to more noisy user-generated data using simulated data sets with varying proportions of misspellings. As can be seen in Figure 2.10a and 2.10b, correction accuracy is either stable or increases when the level of noise increases from 1 to 8%, whereas it appears to diminish at a noise level of 16%. As relative Levenshtein distance does not depend on the noise in the corpus, this possibly indicates that at 16% noise the corpus is affected to the degree that the frequency of correct counterparts of errors often drops below the θ of 9 times the frequency of the error. This is not surprising: due to the equal probability that each word has of being transformed into a mistake, increasingly more words necessary for correction are transformed into

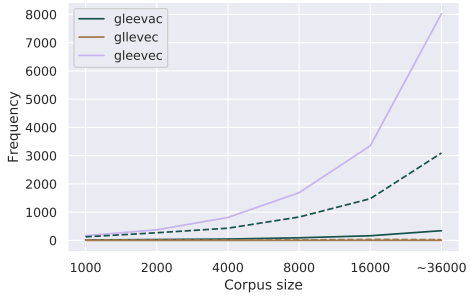
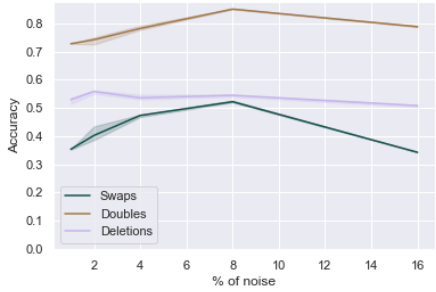


Figure 2.9: Corpus frequency of one uncommon and one common misspelling of the medication Gleevec in the GIST forum with increasing corpus size. The dotted line indicates the corpus frequency threshold for correction candidates for each misspelling.

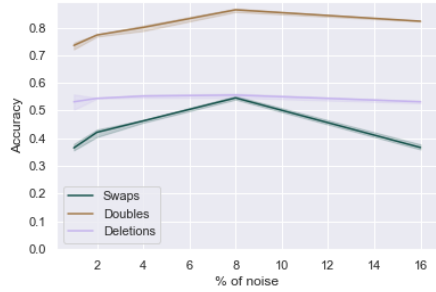
errors. However, no conclusions can be drawn about the exact turning point, as we did not measure the impact of noise levels between 8 and 16%. If necessary, re-tuning of the threshold on a more noisy corpus may resolve this issue.

Except for errors due to doubling of letters, the absolute correction accuracy is far lower than on our real-world data set (79.5%). We believe there may be two reasons for this: firstly, users are more likely to misspell medical terms than other words [352] and thus this random distribution is unrealistic. Such medical terms are likely to be longer than the average word in social media text. Indeed, we find that in our real-world sample of 1000 posts from the GIST forum the 109 non-word errors are significantly longer than average ($p < 1e-22$) according to a Mann Whitney U test: The errors have a mean character length of 6.8 compared to an overall average of 4.2 characters. Since deletions or swaps in shorter words lead to more ambiguous words (e.g., 'the' to 'te') or even other real words (e.g., 'the' to 'he'), this will lower the overall correction accuracy of methods designed to correct non-word medical errors. The second reason ties into this: these artificial data sets do not allow for differentiation between real word and non-word errors and thus are not suited to evaluating absolute non-word error correction. Nonetheless, although absolute accuracy on synthetic data may thus not be a reliable indicator, the relative accuracy at different noise levels does provide a good indication for the impact of the level of noise in the data on the efficacy of our method.

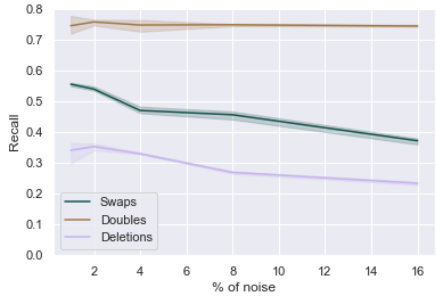
Regarding the detection of errors, recall appears to drop as the level of noise increases for swaps and deletions and remains roughly constant for errors due to doubling of characters (i.e., doubles) (see Figure 2.10c and 2.10d). In contrast, precision increases with increasing noise for swaps and doubles and remains mostly stable for deletions (see Figure 2.10e and 2.10f). These results may indicate that the relative frequency ratios of false positives to their predicted corrections are more frequently close to the detection threshold (θ) of 9 than those of true positives. As an artificial increase in noise by a certain percentage (e.g., 4%) will cause the frequency of correct words to drop by approximately that percentage due to random chance selection of words to transform into errors, increasing noise will lead to a slight drop in the ratio between a token and its predicted correction. If the ratio was far larger than 9, this does not alter the outcome.



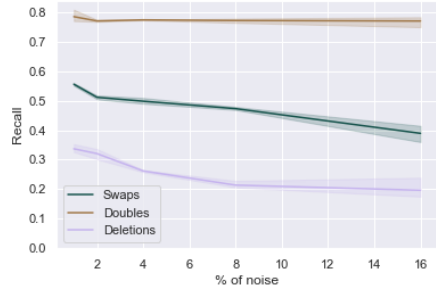
(a) GIST forum
Correction accuracy



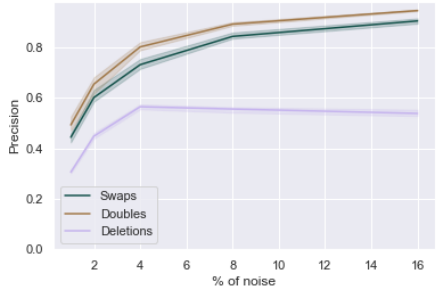
(b) Reddit forum
Correction accuracy



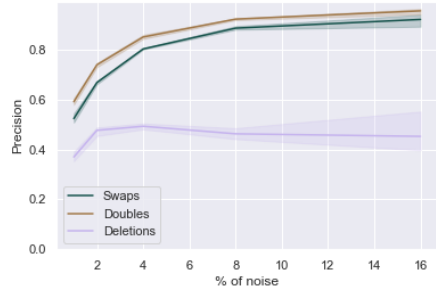
(c) GIST forum
Recall for error detection



(d) Reddit forum
Recall for error detection



(e) GIST forum
Precision for error detection



(f) Reddit forum
Precision for error detection

Figure 2.10: Impact of degree of noisiness of the data (1,2,4,8 and 16% noise) on the detection (c-f) and correction accuracy (a-b) of three types of spelling errors (deletions of a single letter, doubling of a single letter and swaps of adjacent letters) in two cancer-related forums. The lines indicate the mean result while the band indicates the variance in results over three runs.

However, if the ratio was only slightly above 9, then it is liable to dropping below the detection threshold when the noise is increased. In that case, the token will no longer be marked as an error. Thus, if false positives more frequently have ratios slightly above 9 than true positives do, this could explain the increase in precision.

To investigate this idea, we consider swaps in the GIST forum at different levels of noise. It appears that indeed false positives have a higher % of ratios between 9 and 10 than true mistakes at lower noise levels (2,4 and 8%) across all random seeds. This flips for 16%: false positives now have a lower percentage of ratios liable to dropping below the θ of 9 than true positives. Thus, possibly false positives that were 'at risk' for dropping below the required θ have done so. This increased precision does come at a cost: some errors will also have ratios close to 9 leading to a drop in recall with increasing noise levels.

Due to the presence of common errors, the impact of noise might be less pronounced for real data. Although the artificial data does contain common errors (e.g., 'wtih' (218x)), their frequency depends on the frequency of the word of origin (e.g., 'with' (9635x)) because each word has an equal, random probability of being altered. Consequently, their ratio will be much higher and they will be easier to detect than real common errors. Moreover, absolute precision and recall on synthetic data may not be transferable. Overall relative trends, however, do provide an first indication for the generalisability of our method to noisier data sets. Further experimentation with noisier, annotated real world data will be necessary to assess the true effect of noise on our error detection.

For both error correction and detection, results are consistent across the two forums and variance of the results is low except at tail end (16%). This can be explained by the random assignment of transformations for each run: depending on which words are randomly transformed in a certain run, the frequency of certain correct words may either fall below the θ of 9 or not.

2.4.6. EFFECT ON OOV RATE

The reduction in out-of-vocabulary (OOV) terms is higher for the GIST (0.64%) than for the Reddit forum (0.36%) (See Figure 2.11b). As expected, it appears that in-vocabulary terms are occasionally replaced with out-of-vocabulary terms, as the percentage of altered words is higher than the reduction in OOV (0.72% vs 0.64% for the GIST and 0.50% vs 0.36% for the Reddit forum). The vast majority of the posts do not contain any mistakes and of the posts with mistakes, the majority have only one (see Figure 2.11a). Thus, it appears that the spelling mistakes are not caused by a select group of individuals that are poor at spelling, but by various forum users making the occasional mistake.

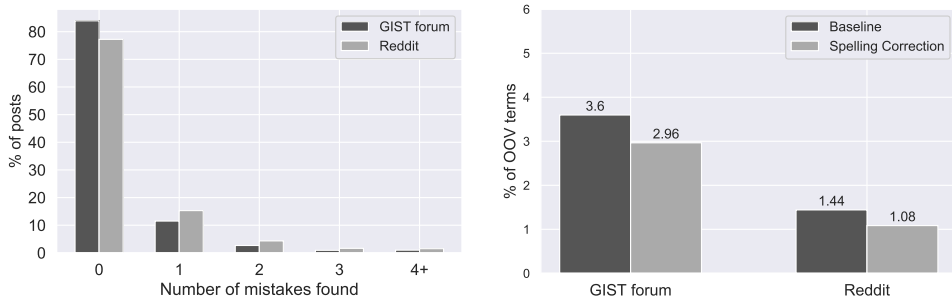
Interestingly, the prior OOV count of the GIST forum is more than double that of the sub-reddit on cancer. This could be explained by the more specific nature of the forum: it may contain more words that are excluded from the dictionary, despite the fact that the dictionary is tailored to the cancer domain. This again underscores the limitations of dictionary-based methods.

Many of the most frequent corrections made in the GIST forum are medical terms (e.g., gleevec, oncologists, tumors). Similarly, the most frequent mistakes found in this forum are common misspellings of medical terms (e.g., gleevac and gleevic) (see Figure 2.12a). It appears that for common medical corrections, there are often various less commonly occurring misspellings per term since their misspelt equivalents do not show

up amongst the most common mistakes. We also found that our method normalizes variants of medical terms to the more prevalent one (e.g., reoccurrence to recurrence). Thus, although the overall reduction in OOV-terms may seem minor, our approach appears to target medical concepts, which are highly relevant for knowledge extraction tasks. In addition, our method incorrectly alters plural to singular variants (e.g., gists to gist), probably due to their higher prevalence in the data. Additionally, due to the addition of the distributed language model, prevalent terms can be replaced by their synonyms as 'corrections' (e.g., resected to removed). Fortunately, the resulting information loss will be minimal for medical downstream tasks.

In the sub-reddit on cancer, frequent corrections include medical terms (e.g., chemotherapy, medication and hospital), normalization from plural to singular (e.g., wifes to wife) but also both incorrect alterations of slang (e.g., gon to got) and of medical terms (e.g., immunotherapy) (see Figure 2.12b). Additionally, the change from didn to did is problematic due to the loss of the negation. Our method thus appears to work less well for more general fora.

Nonetheless, when we consider the 50 most frequent remaining OOV terms, only a small proportion of them are non-word spelling errors, although slang words could arguably also be part of this category (see Table 2.13 for examples). A significant portion consists of real words not present in the specialized dictionary. Importantly, also some drug names and medical slang (e.g., 'scanxiety' or anxiety about being scanned) are considered OOV. Since they can be essential for downstream tasks, it is promising that they have not been altered by our method.



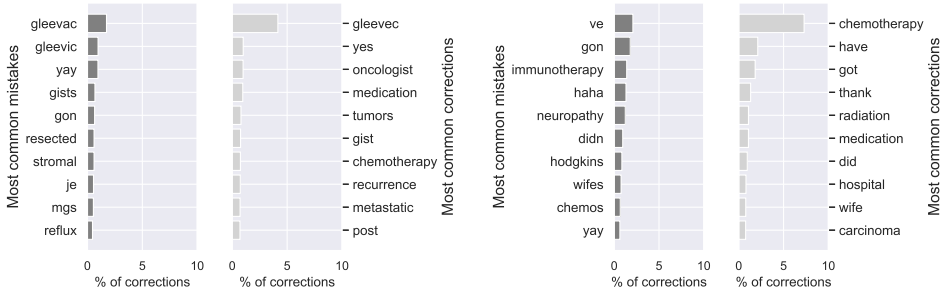
(a) Distribution of found mistakes across posts

(b) Change in out-of-vocabulary terms

Figure 2.11: Internal validation on two cancer forums

2.4.7. EXTERNAL VALIDATION

As can be seen in Table 2.14, normalization leads to a significant change in the F_1 score for two of the six classification tasks ($p = 0.0096$ and $p = 0.0044$). For the Twitter Health corpus, this change is mostly likely driven by a significant increase in recall ($p = 0.0040$), whereas for the detection of flu infection tweets (Task4 SMM4H2019) it is the precision that is increased significantly ($p = 0.0064$). In general, these changes are of the same order of magnitude as those made by the normalization pipeline of Sarker [261]. Although the



(a) GIST forum

(b) Reddit forum on cancer

Figure 2.12: Most frequent mistakes and corrections in two cancer forums

Table 2.13: Manual error analysis of 50 most frequent OOV terms after spelling detection

	GIST	Example	Reddit	Example
Real word	33	unpredictable, internet	42	misdiagnosed, website
Spelling mistake	5	side-effects, wildtype, copay	2	side-effects, inpatient
Abbreviation	2	mos, wk	3	aka
Slang	6	scanxiety, gister	1	rad
Drug name	2	stivarga, mastinib	1	ativan
Not English	2	que, moi	-	
TOTAL	50		50	

overall classification accuracy on Task 1 of the SMM4H workshop is low, this is in line with the low F_1 score (0.522) of the best performing system on a comparable task in 2018 [335].

Especially the expansion of contractions and the splitting of hash tags (e.g., '#flushot' to '#flu shot') appear to impact the classification outcome. In contrast, neither the goal of the task, the relative amount of corrections nor the initial result seem to correlate with the change in F_1 score. The lack of a correlation between the amount of alterations and the change in F_1 score may be explained by the weak reliance of classification tasks on individual terms. Unlike in Sarker [261], the improvements also do not seem to increase with the size of the data. This is logical, as we do not rely on training data. The imbalance of the data may be associated with the change in accuracy to some extent: the two most balanced data sets show the largest increase (see Table 2.3). Further experiments would be necessary to elucidate if this is truly the case.

On generic social media text, our method performs only slightly worse than the state-of-the-art methods (see Table 2.15). We did not need to use the training data, as our method is unsupervised. For comparison, our method attains a F_1 of 0.726, a precision of 0.728, and a recall of 0.726 on the W-NUT training data.

Error analysis reveals that 46 of the 100 most frequent remaining errors are words that should not have been altered according to the W-NUT annotation (see Table 2.16). Yet, in fact, these words are often slang that our method expanded correctly (e.g., info to information). It is thus debatable whether these are errors. Of the remainder, 33 are either

Table 2.14: Mean classification accuracy before and after normalization for six health-related classification tasks. Only the results for the best performing classifier per data set are reported. ∗∗ indicates p<0.005; ∗ indicates p<0.01; † indicates absolute change

		F1		Recall		Precision	
Data set	Words altered	Pre	Δ †	Pre	Δ †	Pre	Δ †
Task1 SMM4H 2019	1.53%	0.410	-0.0007	0.373	+0.014	0.470	-0.025
Task4 SMM4H 2019 Flu Vaccination	0.50%	0.780	+0.006	0.834	+0.008	0.733	+0.005
Flu Vaccination Tweets	0.50%	0.939	+0.002	0.935	+0.004	0.943	+0.0004
Twitter Health	0.71%	0.702	+0.016*	0.657	+0.028*	0.756	-0.0009
Task4 SMM4H 2019 Flu Infection	0.57%	0.784	+0.012**	0.842	+0.013	0.735	+0.019**
Zika Conspiracy	0.36%	0.822	-0.005	0.817	+0.012	0.835	-0.021

uncorrected abbreviations or slang terms. This may partially be explained by the fact that the slang usage of medical forum users differs from the general Twitter population. Lastly, 16 of these 100 can be considered non-word errors that were missed by our method and another 4 are errors that were correctly detected but corrected inaccurately.

	F_1	Precision	Recall
MoNoise [318]	**0.864**	**0.934**	0.803
Sarker's method [261]	0.836	0.880	0.796
IHS_RD [292]	0.827	0.847	**0.808**
USZEGED [31]	0.805	0.861	0.756
BEKLI [24]	0.757	0.774	0.742
LYSGROUP [89]	0.531	0.459	0.630
Our method	0.743	0.734	0.753

Table 2.15: Results for unconstrained systems of ACL W-NUT 2015

2.5. DISCUSSION

The state-of-the-art normalization method for generic social media [261] performs poorly for medical social media with a spelling correction accuracy of only 19.3%. It is significantly outperformed by all edit-based methods regardless of whether the correction candidates are derived from a specialized vocabulary or the data itself. The highest correction accuracy (73.9%) is attained by unweighted relative edit distance combined with the constraint that correction candidates need to be at least 9 times more frequent than the error. This accuracy is further increased by 5.6% point to 79.5% when it is combined with model similarity based on the HealthVec language model. Our preceding decision process is capable of identifying mistakes for subsequent correction with a $F_{0.5}$ of 0.52 and a significantly higher precision than all other methods while retaining a very high recall of 0.96. Additionally, it is almost completely independent of the size of the corpus for the two cancer-related forums, which is promising for its usage in other even smaller,

Table 2.16: Manual analysis of 100 most frequent errors in W-NUT. *also considered non-word mistakes

Type of error	Freq.	Example	Our correction	W-NUT annotation
Should not have been altered	46	info, kinda	information, kind of	info, kinda
Abbreviation not or incorrectly expanded	19	smh	smh	shaking my head
Uncorrected slang	14	esp	esp	especially
Missed concatenation error*	6	incase	incase	in case
Missed apostrophe*	5	youre	youre	you're
Wrong correction	4	u	your	your
Missed split mistake*	3	i g g y	i g g y	iggy
Missed non-word spelling mistake	2	limites	limites	limits
American English	1	realise	realize	realise
TOTAL	100			

domain-specific data sets. Our method can also function well for more noisy corpora up to a noise level of 8% (i.e., 1 error in every 12.5 words).

In the two cancer forums that we used for evaluation, the spelling correction reduces OOV-terms by 0.64% point and 0.36% point. Although the reduction may seem minor, relevant medical terms appear to be targeted and, additionally, many of the remaining OOV-terms are not spelling errors but rather real words, slang, names, and abbreviations. Furthermore, our method was designed to be conservative and to focus on precision to mitigate one of the major challenges of correcting errors in domain-specific data: the loss of information due to the 'correction' of correct domain-specific terms.

Our method also significantly improves the classification accuracy on two tasks, although the absolute change is marginal. On the one hand, this could be because classification tasks do not rely strongly on individual terms. On the other hand, it may be explained by our use of only unigrams as features. Feature extraction would likely also benefit from normalization and could possibly increase performance differences. Further experimentation is required to further assess the full effect of lexical normalization in downstream tasks.

As named entity recognition (NER) tasks rely more strongly on individual terms, we speculate that our method will have a larger impact on such tasks. Unfortunately, NER benchmarks for health-related social media are limited. We have investigated three relevant NER tasks that were publicly available: CADEC [151], ADRMiner [217], and the ADR extraction task of the SMM4H 2019. For all three tasks, extracted concepts could be matched exactly to the forum posts, thus negating the potential benefit of normalization. The exact matching can perhaps be explained by the fact that data collection and extraction from noisy text sources such as social media typically rely on keyword-based searching [264].

Our study has a number of limitations. Firstly, the use of OOV-terms as a proxy for the quality of the data relies heavily on the vocabulary that is chosen and, moreover, does not allow for differentiation between correct and incorrect substitutions. Secondly, our method is currently targeted specifically at correcting non-word errors and is therefore

unable to correct real word errors. Thirdly, the evaluation data set for developing our method is small: a larger evaluation data set would allow for more rigorous testing. Nonetheless, as far as we are aware, our corpora are the first for evaluating mistake detection and correction in a medical patient forum. We welcome comparable data sets sourced from various patient communities for further refinement and testing of our method.

2.6. CONCLUSIONS AND FUTURE WORK

To what extent can corpus-driven spelling correction reduce the out-of-vocabulary rate in medical social media text? Our corpus-driven spelling correction reduces the OOV rate by 0.64% point and 0.36% point in the two cancer-related medical forums we used for evaluation. More importantly, relevant medical terms appear to be targeted.

To what extent can corpus-driven spelling correction improve accuracy of health-related classification tasks with social media text? Our corpus-driven method could significantly improve the classification accuracy on two of the six tasks. This is driven by a significant increase in precision for one and by a significant increase in recall for the second.

In conclusion, our data-driven, unsupervised spelling correction method can improve the quality of text data from medical forum posts. We have demonstrated the success of our method on data from two cancer-related forums. The automatic spelling corrections significantly improve the F_1 score for two of the six external classification tasks that involve medical social media data. Our method can also be useful for user-generated content in other highly specific and noisy domains, which contain many OOV terms compared to available dictionaries. Future work will include extending the pipeline with modules for named entity recognition, automated relation annotation and concept normalization. Another possible avenue for future work could be to determine whether a word is or is not from the domain at hand (the medical domain in our case) prior to normalization and apply different normalization techniques in either case. Furthermore, despite a lack of domain-specific, noisy corpora for training character-level language models, it would be interesting to investigate to what extent our spelling correction can improve classification accuracy using character-level language models pretrained on other source domains.

3

DETECTING PERSONAL EXPERIENCES

Edited from: **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2019), *Narrative Detection in Online Patient Communities.* Proceedings of Text2Story — Second Workshop on Narrative Extraction From Texts co-located with 41th European Conference on Information Retrieval (ECIR 2019). 21-28.

In this chapter, we discuss the extraction of messages containing the experiences of patients (hereafter called narratives) from patient fora. This subset will also include messages in which patients share their experiences with adverse drug events and may thereby aid in their extraction.

Prior to this study, the systematic detection and analysis of patient narratives was limited to a single study in which lower-cased words were used to identify narratives. In contrast, here we examine whether psycho-linguistic features or document embeddings could aid their identification. We also investigate which features distinguish narratives from other social media posts. Moreover, this study is the first to automatically identify the topics discussed in narratives on a patient forum.

We find that for the identification of patient narratives, character 3-grams outperform psycho-linguistic features and document embeddings. Additionally, we find that narratives are characterized by the use of past tense, health-related words and first-person pronouns, whereas non-narrative text is associated with the future tense, emotional support words and second-person pronouns. Topic analysis of the patient narratives uncovered fourteen different medical topics, ranging from tumor surgery to side effects. Future work will use these methods to extract experiential patient knowledge from social media.

3.1. INTRODUCTION

Nowadays, online patient forums are the main medium by which patients exchange their narratives. These narratives mainly recount their own experiences with their condition. As such, they contain experiential knowledge [38], defined as the knowledge that patients gain from their own experiences. In recent years, such experiential knowledge has increasingly been recognized as valuable and complementary to empirical knowledge [50]. Consequently, more health-related applications are making use of patient forum data, for instance to track public health trends [267] and to detect adverse drug events [266]. Experiential knowledge is also valuable for patients themselves: patients indicate that they strongly rely on experiences and information provided on patient forums [277]. This is especially true for patients with a rare disease, for which medical professionals often lack expertise and the number of studies is limited [15].

To understand the experiential knowledge on patient forums, forum posts that contain narratives must first be identified. As of yet, research into systematically distinguishing patient narratives on patient forums is limited to a single study on Dutch forum data [328], which uses words as only features. We expand upon this work using a different data set by examining whether document embeddings and psycho-linguistic features can improve the identification of patient narratives. We expect so, because these aggregated features are less dependent on individual terms, which may overlap significantly between narratives and factual statements about the same topic. Secondly, we explore how narratives differ from other types of posts by studying which features are influential in identifying narratives and which posts are classified incorrectly. Thirdly, we analyze how prevalent narratives are on a cancer patient forum and which topics these narratives discuss.

3.2. RELATED WORK

Narratives on patient forums have mainly been studied qualitatively (e.g., [325]). The automatic identification of narratives on a patient forum is limited to the study by Verberne et al. [328] on a Dutch cancer forum. They identified narratives with a F_1 of 0.911 using only the lower-cased words of the posts as features. They also found that various linguistic factors (1st person singular, 3rd person and negations) and psychological processes (social processes and religion) were correlated with the presence of narratives. These psycho-linguistic features were measured using the Linguistic Inquiry and Word Count (LIWC) method [297].

Additionally, research into self-reported adverse drug events (ADE) has led to the development of classifiers for differentiating between factual statements of ADE and personal experiences of ADE on social media [33, 217, 262]. However, these classifiers are highly specific and thus not suitable for identifying patient narratives in general.

Another closely related field is the classification of personal health mentions on social media, i.e., posts that mention a person who is affected as well as their specific condition, such as: 'my granddad has Alzheimer's'. Presently, only two studies have investigated this task. The first by Lamb et al. [169] focused on separating flu awareness from actual flu reports on social media. More recently, Karisani and Agichtein [152] introduced WESPAD, a classifier for personal health mentions, which attains state-of-the-

art performance for seven different health domains including stroke, depression, and flu infection. Nonetheless, a personal health mention alone is not sufficient to consider the post a narrative, and thus these classifiers are also inadequate for our purpose.

3.3. METHODS

3.3.1. DATA

Our data consists of an open, international Facebook forum for patients with Gastrointestinal Stromal Tumor (GIST)[1]. It is moderated by GIST Support International and consists of 36,722 posts with a median length of 20 tokens.

3.3.2. PREPROCESSING

The data was lower-cased and tokenized with NLTK. Due to the noisy nature of user-generated content, especially in the spelling of medical terms, we applied a tailored preprocessing pipeline[2] to our data. Firstly, an existing normalization pipeline for social media[3] [261] was used to normalize tokens to American English and to expand generic abbreviations used on social media. Hereafter, domain-specific abbreviations were expanded with a lexicon of 42 non-ambiguous abbreviations, generated based on 1000 posts and annotated by a domain expert and the first author. Spelling mistakes were detected using a combination of relative frequency and edit distance to possible candidates and corrected using weighted Levenshtein distance. Correction candidates were derived from the corpus itself. Drug names were normalized using the RxNorm database [314]. Non-English posts were removed using langid [190]. Punctuation was removed, but stop words were not, as we expect function words to play a role in the expression of narratives.

3.3.3. SUPERVISED CLASSIFICATION

Manual annotation of example data We randomly selected 1050 posts for annotation. The annotators were asked to indicate per message whether it contains a personal experience. They were not provided with its context. Personal experiences did not need to be about the author but could be about someone else. This definition was based on earlier work by Verberne et al. [328] and van Uden-Kraan et al. [324]. The first 50 posts were annotated individually by the first author and another PhD student to improve the annotation guidelines.[4] The remaining 1000 posts were divided equally into six sets of 200 posts, with 40 posts (20%) overlapping between all sets. The overlap was used to calculate the pairwise Cohen's kappa. There were seven annotators in total: six PhD students and one GIST patient. Each sample was assigned to an annotator, apart from one sample which was divided between two PhD students. To be able to include the overlapping sample in the classification, we opted to use the annotations of the GIST patient for these 40 posts.[5]

[1] `https://www.facebook.com/groups/gistsupport/`
[2] The preprocessing scripts can be found at: `https://github.com/AnneDirkson/LexNorm`
[3] `https://bitbucket.org/asarker/simplenormalizerscripts`
[4] The annotation guidelines can be found at: `https://github.com/AnneDirkson/NarrativeFilter`
[5] The annotated data is available upon request in order to protect the privacy of the patients

Feature sets Four feature sets were derived from the text data: word unigrams, character n-grams (using the CountVectorizer function in sklearn), psycho-linguistic features, and document embeddings. For both word unigrams and character n-grams, we investigated whether TF-IDF weighting would improve performance compared to raw counts. Additionally, we explored whether stemming or lemmatizing the data prior to extracting the unigrams could improve performance. Psycho-linguistic features were based on the LIWC 2015 [297]. Punctuation categories were discarded, resulting in 82 LIWC features in total. LIWC is a well-known method for investigating psychological processes in text and includes both linguistic (e.g., first-person pronouns) and psychological categories (e.g., positive emotions). The last feature set consisted of document embeddings: a doc2vec model [172] was trained on the labeled training data for each fold in the cross-validation. We combine a distributed memory model with a distributed bag of words model, as recommended by Le and Mikolov [172]. We also attempted to train document embeddings first on the unsupervised data and then retrain on the supervised data, but this led to nonsensical classification features.

Supervised classification algorithms Classifiers were evaluated separately for each feature set. We ignored all posts that had been left empty by the annotator (the annotator chose neither yes nor no): three posts were ignored for this reason. For word unigrams, character n-grams, and psycho-linguistic features, we compared four sklearn classification algorithms: Multinomial Naive Bayes (MNB), linear Support Vector Classification (LinearSVC), Stochastic Gradient Descent (SGD) with log loss, and K Nearest Neighbors (KNN). These were chosen according to the following criteria: (1) known to perform well on text data, (2) recommended for small data sets, and (3) able to calculate probabilistic outcomes. The latter enabled us to use probabilistic ensembles. The doc2vec representations combined with Logistic Regression were used as classifier in itself: the document representations were tagged with the labels of the training data. This model was then used to derive vector representations for new documents. To test if a combination of feature types could improve performance, we evaluated soft voting (argmax of the sums of the predicted probabilities) of the best individual classifiers for the best performing variants of each feature set. Significance testing was done with pair-wise t-tests.

To evaluate the performance, the average F_1 score of a 10-fold cross validation was used. For each run, hyper-parameters were tuned for that specific training set using a 10-fold grid search on the training data. The tuning grids were based on sklearn documentation: C from 10^{-3} to 10^{3} (steps of x10) for LinearSVC and Logistic Regression; number of neighbors from 3 to 11 (steps of 2) for KNN; and max iterations from 2 to 2048 (steps of x2) and alpha from 10^{-8} to 10^{-2} (steps of x10) for SGD. The dimensionality of the document vectors was tuned on a grid of 100 to 400 (steps of 100).

3.3.4. TOPIC MODELING OF THE WHOLE DATA SET

To label the remaining data, the best performing classifier was used with the hyper-parameter settings that were optimal in the majority of the training sets. To investigate which topics are discussed in the patient narratives, we used topic modeling with non-Negative Matrix Factorization of the TF-IDF weighted tokens without stopwords. Topic coherence, measured using TC-W2V [223], was used to select the number of topics. Topic

labels were assigned manually by exploring the words with the highest weights and the top-ranked (i.e., most relevant) messages per topic.

3.4. RESULTS

3.4.1. ANNOTATED DATA

The data was slightly imbalanced, with 37.7% of the posts containing a narrative, resulting in a majority baseline of roughly 0.62. The inter-annotator agreement was substantial ($\kappa = 0.69$).

3.4.2. CLASSIFIER EVALUATION

A Linear SVC on character 3-grams achieves the highest F_1 score (Table 3.1), although character 4-grams ($p = 0.526$), stemmed unigrams ($p = 0.930$) and lemmatized unigrams ($p = 0.587$) do not perform significantly worse. Character 5- and 6- grams also do not perform worse overall ($p = 0.122$ and $p = 0.169$), but their recall is significantly lower ($p = 0.023$ and $p = 0.029$). The classifiers for the best performing document embeddings (DBOW+DM) and psycho-linguistic features, however, are significantly worse overall than character 3-grams ($p = 0.0055$ and $p = 0.026$ respectively). Employing TF-IDF weighting does not aid any of the unigram or character n-gram features. Additionally, neither feature selection (F_1=0.761) nor word boundaries (F_1=0.796) improve the performance of character 3-grams. Using a range of character n-grams, namely 3-to-4 (F_1=0.814), 3-to-5 (F_1=0.814), or 3-to-6 (F_1=0.812), also does not boost performance.

Ensemble classification did not perform better than character 3-grams alone (see Table 3.2). Nevertheless, an ensemble of all four feature types is significantly more precise than all other classifiers ($p = 0.0048$ compared to the second best). To further explore why ensemble classification does not manage to improve overall performance, we investigated the predictions of individual classifiers. As can be seen in Table 3.3, there is a high degree of overlap between the predictions based on character 3-grams and the other feature sets (88.3%, 83.8% and 84.4% respectively). Consequently, the vast majority of the predictions cannot be improved by complementing character 3-grams with these feature sets. Interestingly, 4.7% of the posts are misclassified by all feature sets. Considering the non-overlapping predictions, the percentage of correct predictions was higher for character 3-grams than for either document embeddings or psycho-linguistic features in a pairwise comparison. Thus, it appears that adding these features would be more detrimental than beneficial to narrative classification.

3.4.3. INFLUENTIAL FEATURES

Narratives are typically distinguished by terms relating to the past tense (*was, had, years*), health (*imatinib, tumor, surgeri*) and first-person narrative (*my, i*) (see Figure 3.1). This is corroborated by the character 3-grams, psycho-linguistic features and document embeddings. Some of the important terms for non-narrative texts are also health-related (*patients, gist*) and first-person narrative (*we, us*), which showcases the difficulty of the task at hand. In general, non-narrative texts seem to focus more on emotional support (*prayer, share, may*), second-person narrative (*you, your*) and the future (*may, will*). The psycho-linguistic features additionally reveal that narratives contain more mentions of

Table 3.1: Mean test score (10-fold CV) for best classifiers per feature set

Feature set		Size	Classifier	F_1	R	P
Unigrams	**Original**	4,078	SGD	0.795 ± 0.025	0.788 ± 0.074	0.811 ± 0.055
	Stemmed	3,205	SGD	0.814 ± 0.031	0.793 ± 0.047	0.840 ± 0.049
	Lemmatised	3,777	SGD	0.808 ± 0.039	0.810 ± 0.059	0.813 ± 0.070
Character n-grams	**3-grams**	5,086	SVC	**0.815 ± 0.035**	**0.844 ± 0.047**	0.793 ± 0.058
	4-grams	16,496	SVC	0.811 ± 0.027	0.827 ± 0.068	**0.844 ± 0.029**
	5-grams	36,349	SGD/SVC	0.796 ± 0.023	0.784 ± 0.059	0.817 ± 0.069
	6-grams	60,443	SGD	0.793 ± 0.040	0.797 ± 0.042	0.795 ± 0.079
LIWC		82	SVC	0.773 ± 0.031	0.805 ± 0.044	0.752 ± 0.077
Doc2vec	**DBOW**	400	LogReg	0.737 ± 0.029	0.751 ± 0.056	0.735 ± 0.066
	DM	400	LogReg	0.762 ± 0.039	0.749 ± 0.062	0.785 ± 0.070
	DM+DBOW	800	LogReg	0.77 ± 0.037	0.803 ± 0.064	0.749 ± 0.055

Table 3.2: Mean test score (10-fold CV) for ensemble classification. * DM+DBOW variant.

Feature sets	F_1	R	P
3-grams + LIWC + Doc2vec* + Stemmed Unigrams	0.770 ± 0.029	0.703 ± 0.065	**0.859 ± 0.053**
3-grams + LIWC + Doc2vec*	**0.795 ± 0.037**	**0.772 ± 0.072**	0.829 ± 0.065
3-grams + LIWC	0.706 ± 0.032	0.624 ± 0.059	0.828 ± 0.073
3-grams + Doc2vec*	0.755 ± 0.048	0.735 ± 0.089	0.786 ± 0.040

causality and negative emotions. In contrast, non-narrative texts seem to contain more positive emotions. Lastly, as predicted, function words appear important for classifying narratives in social media, and it is thus advisable to not remove stopwords.

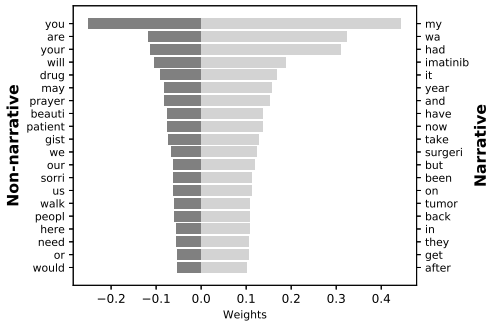
3.4.4. ERROR ANALYSIS FOR THE BEST PERFORMING CLASSIFIER

Error analysis reveals that a significant proportion of the errors is due to incorrect annotation: 36.9% of the false positives and 36.2% of the false negatives were labeled incorrectly (see Table 3.4). Specifically, annotators have difficulty correctly labeling discussions about personal medical facts or side effects as narratives (e.g *'i have been on imatinib 5 months and lost 1/3 of my hair'*). Conversely, annotators may incorrectly judge posts that give emotional support, external information or advice to be narratives while they are not (e.g., *'i may be wrong but total gastrectomy sounds very extreme for two small gist'*).

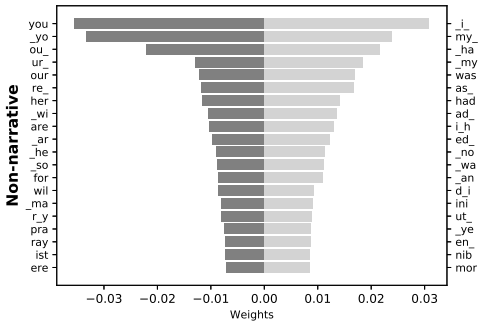
The incorrect labeling may have impacted the automated classification such that these categories are also more difficult for the computer to distinguish. The classifier does, however, appear to outperform human judgment and to some extent 'correct' their mistakes. In fact, its performance may be underestimated by the metrics based on these incorrect labels. Other types of posts that appears challenging for the computer are posts that lack context or contain questions. The former are often answers to unknown questions posed earlier in the thread.

Table 3.3: Comparison of predictions of classifiers for different feature sets. * DM+DBOW variant.

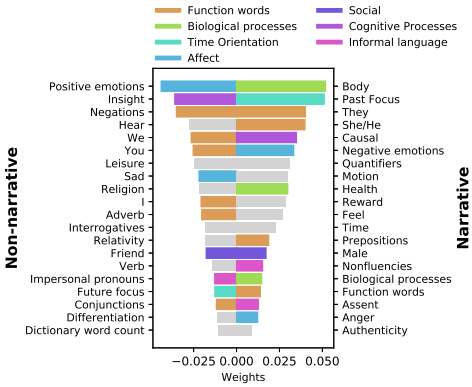
	Compared to	Both		Difference	
		Correct(%)	Incorrect(%)	In Favor of 3-grams(%)	In Favor of Other Method(%)
Character 3-grams	**LIWC**	75.0	8.8	8.4	7.7
	Doc2Vec*	74.8	9.6	8.6	6.9



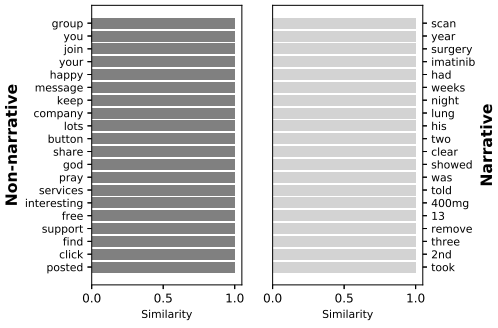
(a) Stemmed Unigrams



(b) Character 3-grams



(c) LIWC



(d) Doc2vec DM model

Figure 3.1: The 20 most influential features in individual classifiers. In (b) underscores represent spaces.

3.4.5. FREQUENCY AND CONTENT OF PATIENT NARRATIVES

Automated narrative detection in unsupervised data The percentage of narratives in the unlabeled data is 37.0 %, which is comparable to the annotated sample. This results in a total of 13.436 posts for topic modeling.[6]

[6]The code for unsupervised narrative filtering is shared at: `https://github.com/AnneDirkson/NarrativeFilter`

Table 3.4: Error analysis for best classifier (character 3-gram classification of narratives)

False positives		**False negatives**	
Reasons for misclassification	Frequency	Reasons for misclassification	Frequency
Mislabeling	24	Mislabeling	17
Emotional support/thanks	15	Unknown	12
Information/advice	13	Lack of context	7
Lack of context	7	Question	5
Question	4	Non-medical narratives	3
Unknown	1	Hypothetical	1
Empty post	1	Empty post	2
TOTAL	65	TOTAL	47

Topic modeling The TC-W2V metric [223] identifies the optimal number of topics to be fourteen. The resulting topics relate to different aspects of the medical process for GIST patients (see Table 3.5). Note that imatinib is the most commonly used medication.

3.5. DISCUSSION

The detection of narratives was most optimal when using character 3-grams. Their strength is in their ability to cluster relevant word types based on suffixes and prefixes. This is especially relevant in the medical domain, e.g., all cancer medication for GIST ends in *'nib'*. In contrast, psycho-linguistic features appear to suffer from oversimplification, because they aggregate words that define *different* classes into one category e.g., *we* and *my* into the umbrella category of first person pronouns (see Figure 3.1). The use of document embeddings may have been hampered by the small size of the data. An alternative explanation could be that incorrect labeling impacts these features more strongly than word-based features.

Narratives could be differentiated most strongly by their use of past tense, first-person narrative and health-related words. The first two are in line with linguistic definition of a narrative. The stronger focus on health, however, may indicate that patients prefer to share their own health experiences than health information from external sources.

Annotating narratives appears a challenging task, despite providing annotators with a guideline based on previous work [328] and validated through initial annotation by two annotators. This is underscored by our inter-annotator agreement ($\kappa = 0.69$) which was comparable to that of Verberne et al. [328] ($\kappa = 0.71$). Our classifier performed less well that their system ($F_1 = 0.91$), which may be explained by their larger sample of annotated data (2.051 posts).

Inevitably, our results depend on the choice of what constitutes a narrative and how the annotators interpret this definition. It appears that especially the line between a medical fact about oneself and a medical experience is fuzzy for annotators. Future studies could perhaps use this knowledge to develop clearer guidelines.

Table 3.5: Most important topics discussed in patient forum narratives. Topic labels were assigned manually. *A type of cancer medication

Topic labels	Top 10 words	Top-ranked post for the topic
Tumor location	tumor stomach removed liver small cm mitotic metastases rate intestine	"i only had one tumor on my stomach"
(Emotional) Coping	take get time doctor like also know imatinib* day would	"i completely understand i started 400 imatinib after surgery in and have lots of bad days [...]"
Duration of Treatment	years imatinib* almost ago 10 taking two still 11 12	"about 1 and 1/2 years"
Types of Scans	scan ct pet results next today last showed week cat	"oops one is a ct scan and one is a pet scan"
Diagnosis of GIST	gist diagnosed cancer specialist oncologist husband anyone ago surgeon found	"that was my gist"
Other Medication	sunitinib* regorafenib* sorafenib* imatinib* working 37 exon nilotinib* trial stopped drug	"i have this on sunitinib"
Side Effects	side effects imatinib* effect different fatigue eyes bad 400mg time	"and no side-effects"
Tumor Surgery	surgery remove since weeks first post surgeon second shrink done	"just had surgery"
Absence of Tumor Recurrence	disease evidence still years today post since resection year far	"no evidence of disease no evidence of disease"
Recurrence of Work, Medication or Tumor	back came come hair go went weeks took coming lost	"i started imatinib after i went back to work"
Emotional support	good luck news best far hope bad goes well keep pretty	"all my best and good luck"
Dosage of Medication	mg 400 800 imatinib* 600 take day taking since started	"11 years of imatinib since 2003 at 600 mg and since november 2009 at 800 mg [...]"
Timing of Scans	months every scans three ct six year two first month	"my doctor said 3 years"
Ingesting imatinib	one year last took imatinib* day another old got time	"take imatinib"

3.6. CONCLUSION

For the detection of patient narratives on social media, psycho-linguistic features and document embeddings are outperformed by character 3-grams. These narratives are associated with the past tense, health and first-person pronouns, whereas non-narrative text is associated with the future tense, emotional support and second-person pronouns. The patient narratives could be subdivided into discussions of fourteen different medical topics, ranging from surgery to side effects. Future work will develop automated methods for the extraction of patient knowledge from the narratives.

4

CONVERSATION-AWARE FILTERING OF RELEVANT MESSAGES

Edited from: **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2020). *Conversation-aware Filtering from Online Patient Forums.* Proceedings of the Fifth Social Media Mining for Health Applications (SMM4H) Workshop @ COLING.

In this chapter, we explore the benefit of exploiting conversational context for filtering posts relevant to a specific medical topic, such as adverse drug events. The filtering of relevant posts from a larger corpus is a commonly used first step towards knowledge extraction from social media.

Previous approaches to NLP tasks on online patient forums have been limited to single posts as units, thereby neglecting the overarching conversational structure. Here, we experiment with two approaches to add conversational context to the state-of-the-art BERT model: a sequential CRF layer and manually engineered features.

Although neither approach can outperform the F_1 score of the baseline, we find that adding a sequential layer improves precision for all target classes, whereas adding a non-sequential layer with manually engineered features leads to a higher recall for two out of three target classes. Thus, depending on the end goal, conversation-aware modeling may be beneficial for identifying relevant messages. We hope our findings encourage other researchers in this domain to move beyond studying messages in isolation towards more discourse-based data collection and classification.

4.1. INTRODUCTION

In the past decade, social media has emerged as a source of valuable knowledge in the health domain [116], for instance during the COVID-19 pandemic [159, 269]. In order to use social media to answer a medical question, it is necessary to identify posts on the forum that are relevant to the question at hand e.g., posts mentioning adverse drug events (ADEs) [183], personal experiences [87], medication abuse [267] or medical misinformation [158]. This filtering step is often the first step of the analysis pipeline. In this chapter, we will refer to this specific type of filtering as relevance classification.

Previous automatic methods for medical relevance classification generally consider posts as units without context, thereby ignoring any information that can be gained from the conversational context. One example of such an approach is the recent shared task on ADE relevance classification [337]. Yet, including the conversational context may prove beneficial to relevance classification, as responses in a thread often relate to previous responses. For example, responses to a question or comment about a specific side effect are likely to also concern this side effect. To test this hypothesis, we investigate how positive labels are distributed across and within conversational threads.

At present, only one study into medical relevance classification has included some engineered features to capture aspects of the conversational structure [158]. However, as this study includes only two discourse-based features, the effect of including manually engineered features that capture conversational structure is still largely unknown for relevance classification tasks.

Furthermore, including the relation between posts on a discourse level may also be able to improve classifier performance. Each post serves a conversational function in a dialogue, e.g., a question, explanation or statement [14]. These functions are called *dialogue acts* [288]. We have not found any study that included dialogue acts as features for medical relevance classification.

As an alternative to using manually engineered features, conversational threads can also be modeled with a sequential model. This has proven beneficial in other fields such as rumor classification in social media discussions [354]. As of yet, the use of sequential models for medical relevance classification has also not been explored.

We address the following research questions in this chapter:

RQ1 To what extent can the addition of a sequential model on top of state-of-the-art non-sequential models improve medical relevance classification of social media data?

RQ2 To what extent can the addition of manually engineered features for conversational structure and discourse improve medical relevance classification?

We use two different data sets for answering our questions. In our current research, we are particularly interested in discovering ADEs in online discussions. We have collected and annotated a data set about this topic. Since this data set is new, no other results have been published for it. We therefore use one other data set for evaluating our methods: the medical misinformation data set by Kinsora et al. [158]. We use a BERT-based model as baseline. BERT models constitute the current state of the art for most NLP tasks [84] including ADE relevance classification [337].

In the following section, we will elaborate on related work. Hereafter, we describe our methodology and data in Section 4.3 and 4.4 respectively. Finally, we present and discuss our results in Section 4.5 and 4.6.

4.2. RELATED WORK

The use of conversational structure for improving the performance of classifiers of social media posts is prevalent in the field of rumor classification [354] and related fields like disagreement detection [254]. Conversational structure has previously been exploited through (a) manually engineered features or (b) sequential classifiers.

The most commonly employed engineered features to model the conversational structure are the similarity to the previous message and to the thread in general [354]. In addition to these features, the current state-of-the-art model on a leading shared task for rumor stance classification (RumourEval-2019) uses the label of the previous message and the distance to the start of the thread [181]. In the health domain, the only study that employs manually engineered features for conversational structure is Kinsora et al. [158]. Specifically, they use the running count of positive labels and the distance to the previous positive label. In this study, we will employ the above features as well as expand upon them with additional discourse-related features.

Other studies have used sequential classifiers to model the discursive nature of social media, although according to Zubiaga et al. [354] this is "still in its infancy" (p. 276). Their comparison of various classifiers for rumor stance classification revealed that sequential classifiers outperform non-sequential classifiers overall. This is probably due to their ability to leverage information about sequential structure and preceding labels. Furthermore, Zubiaga et al. [354] found that sequential classifiers did not benefit from contextual features representing thread context (e.g., similarity to the source tweet) whereas non-sequential classifiers did. They speculate that sequential classifiers take the surrounding context into account implicitly. To see if this also holds true for relevance classification in medical social media, we will compare the addition of conversation-aware features to both sequential and non-sequential models.

4.3. METHODS

4.3.1. MODELS

CRF As a sequential model we use Conditional Random Fields (CRF). We train the models using the implementation in sklearn-crfsuite. L1 and L2 regularization parameters were tuned for each fold.

Linear SVM As a non-sequential counterpart, we use the sklearn implementation of Linear Support Vector Machines. The hyper-parameter C is tuned per fold with a grid of 10^{-3} to 10^{3} in steps of $\times 10$.

DistilBERT As BERT model, we opt for DistilBERT (distilbert-base-uncased), which is a lighter, more computationally efficient variant of BERT [260]. We use the Huggingface implementation [339] with the wrapper ktrain [195] to train our models. The initialization

seed is set to 1. We use the default learning rate of 5×10^{-5} and tune the number of epochs (3 or 4) per fold.

Ensemble models To investigate the benefit of adding a sequential model on top of the DistilBERT model, we experiment with a blending-based ensemble method: we input the raw confidence scores from DistilBERT for each label as features in a CRF model (i.e., CRF + BERTpred). We create an equivalent non-sequential baseline by using the same approach with an SVM (i.e., SVM + BERTpred).

4.3.2. FEATURE ANALYSIS

To explore the benefit of manually engineered features that capture thread context, we use step-wise greedy forward feature selection using the features in Table 4.1. For each step-wise iteration, we select the best feature to add to the model until the F_1 score no longer improves. We use 10-fold cross-validation in which for each fold features are selected on the development data (10%) and tested on a held-out test set (10%). For a fair comparison, we keep the folds and hyper-parameters the same as for the respective base model. Since the label distribution features could leak information, we omit these gold annotated features for evaluation. Instead, we perform an initial run without these features and use the resulting predictions to calculate them for the final evaluation.

4.3.3. MODEL COMPARISON

We used 10-fold cross-validation in all experiments. Instead of splitting per message, we split on whole discussion threads to ensure possible dependencies between posts do not bias the outcome. Statistical comparisons of model performance are done using Wilcoxon signed rank tests across the 10 folds. To avoid the multiple testing problem, we only compare the three best models, namely those with the highest F_1 score, precision, and recall, with the BERT baseline.

4.4. DATA

Data collection At present, there is only one publicly available medical relevance classification data set that includes the conversational structure: the Medical Misinformation Data set [158]. It is based on MedHelp data and annotated for the presence of misinformation. We collected a second data set from a Facebook group of Gastrointestinal Stromal Tumor (GIST) patients. We selected 527 discussions based on their likelihood to contain an ADE: We selected the threads that contained (1) at least one drug name according to a match with RxNorm [314] and (2) a high percentage of posts in which authors shared experiences. The latter criterion was included since sharing that you had an ADE is an example of experience sharing. To estimate this, we used the classifier described in Chapter 3. According to our classifier, at least 80% of the posts within each selected thread is a personal experience. Due to privacy issues and ownership of the data by the GIST International patient organization, we are not able to share this data set at present. See Table 4.2 for more details on the data sets.

[1] We opt for USE instead of BERT embeddings, as cosine similarity cannot be applied directly to BERT embeddings

Feature type	Name	Description	Explanation (if applicable)
Local	+Emb	Sentence Vectors	We use Universal Sentence Encoder (USE) [57] to encode sentences into 512 dimensional vectors based on pre-trained embeddings so their cosine similarity (normalized between 0 and 1) approximates their semantic similarity.[1]
	+BERTpred	distilBERT predictions	The raw confidence scores for each label
Relational	+PrevSim	Similarity to previous message	Similarity is calculated using the USE sentence vectors
	+ThreadSim	Thread similarity	Similarity to USE sentence vector of all other posts in the thread combined into one vector
Positional	+Dist	Absolute distance from start of thread	
Label distribution	+PrevLbl	Label of previous post	We use the true labels for training and the predicted labels for testing for all label distribution features.
	+CountPos	Absolute running count of preceding positive labels in thread	
	+CountNeg	Absolute running count of preceding negative labels in thread	
	+RelPos	Percentage of preceding positive labels	
	+DistPos	Distance from previous positive label	
	+DistNeg	Distance from previous negative label	
Discourse	+DA	Dialogue act of post	Dialogue acts are calculated using the Dialogue Act tagger as trained by Tortoreto et al. [299]
	+PrevDA	Dialogue act of previous post	

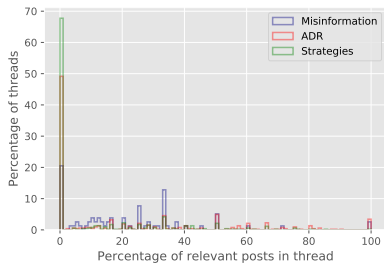
Table 4.1: Manually engineered features to model conversational structure

Data annotation Following a pilot annotation round, the data was annotated by the first author and three patients for the presence of ADEs and coping strategies for dealing with ADEs (hereafter also called: Strategies) using an annotation guideline.[2] The pairwise inter-annotator agreement was substantial for ADEs (mean κ =0.71) and moderate for Coping Strategies (mean κ =0.54).

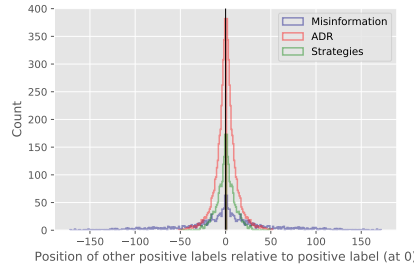
[2]Available at: `https://github.com/AnneDirkson/ConversationAwareFiltering`

Data set	Target	#Posts	#Discussions	Median length	% Positive
Medical Misinformation Dataset [158]	Misinformation	1,566	78	8.0	15.0 %
ADE Discussions (In-house)	Adverse Drug Event (ADE) & Coping Strategies	4,195	527	6	22.9 % & 12.3%

Table 4.2: Statistics on the data sets. The ADE Discussions data set has two target classes.



(a) Distribution of target posts *across* threads



(b) The relative position of target posts to each other *within* threads

Figure 4.1: Distribution of the target class (i.e., positively labeled posts)

4.5. RESULTS

4.5.1. DISTRIBUTION OF THE TARGET CLASS IN THE DISCUSSION THREADS

As visualized in Figure 4.1a, the target class is not distributed equally across the discussion threads for any of the data sets; There appear to be many threads with few or no target posts. According to z-tests, the distribution is significantly different from normal. An inspection of the relative position of target posts *within* discussion threads reveals that the target posts also cluster together (see Figure 4.1b). The probability that the post after a target post is also a target post is 27% for Misinformation and 40% and 34% for ADEs and Coping Strategies respectively. These probabilities are higher than is to be expected based on the percentage of positively labeled posts (see Table 4.2). Thus, it appears that the conversational structure is indeed related to the probability of a post being relevant and consequently incorporating conversational structure or discourse may be able to improve the performance of relevance classifiers.

4.5.2. MODEL COMPARISON

The results of model evaluation are presented in Table 4.3. It appears that neither the addition of a sequential layer nor manual features can improve upon the F_1 score of the BERT model. Misinformation detection appears to be the exception, as any additional layer, sequential or not, outperforms the BERT baseline model. The highest overall

	Misinformation		
	F_1	P	R
BERT	0.366 ± 0.155	0.386 ± 0.154	0.396 ± 0.235
SVM+Emb	**0.478 ± 0.083**	0.492 ± 0.109	0.482 ± 0.111
+ Features	0.392 ± 0.089	0.457 ± 0.169	0.405 ± 0.156
CRF+Emb	0.424 ± 0.155	**0.565 ± 0.148**	0.352 ± 0.162
+Features	0.457 ± 0.137	0.557 ± 0.155	0.420 ± 0.167
SVM + BERTpred	0.443 ± 0.078	0.449 ± 0.082	0.479 ± 0.151
+Features	0.454 ± 0.070	0.449 ± 0.081	**0.492 ± 0.140**
CRF + BERTpred	0.434 ± 0.079	0.453 ± 0.100	0.447 ± 0.138
+Features	0.428 ± 0.078	0.435 ± 0.092	0.446 ± 0.126

	ADEs		
	F_1	P	R
BERT	**0.714 ± 0.034**	0.715 ± 0.038	0.718 ± 0.062
SVM+Emb	0.640 ± 0.054	0.673 ± 0.055	0.613 ± 0.069
+Features	0.610 ± 0.068	0.621 ± 0.087	0.624 ± 0.128
CRF+Emb	0.654 ± 0.059	0.710 ± 0.036	0.611 ± 0.086
+Features	0.638 ± 0.067	0.695 ± 0.037	0.601 ± 0.110
SVM + BERTpred	**0.714 ± 0.035**	0.724 ± 0.043	0.707 ± 0.056
+Features	0.677 ± 0.121	0.673 ± 0.164	**0.738 ± 0.103**
CRF+ BERTpred	**0.714 ± 0.038**	**0.728* ± 0.040**	0.704 ± 0.062
+Features	0.713 ± 0.039	0.726 ± 0.040	0.705 ± 0.060

	Strategies		
	F_1	P	R
BERT	**0.581 ± 0.060**	0.622 ± 0.087	**0.563± 0.111**
SVM+Emb	0.517 ± 0.101	**0.660 ± 0.111**	0.434 ± 0.111
+Features	0.502 ± 0.108	0.603 ± 0.137	0.453 ± 0.128
CRF+Emb	0.441 ± 0.134	0.597 ± 0.120	0.373 ± 0.151
+Features	0.512 ± 0.106	0.609 ± 0.110	0.462 ± 0.143
SVM+Bertpred	0.578 ± 0.059	0.632 ± 0.091	0.545 ± 0.089
+Features	0.561 ± 0.095	0.601 ± 0.146	0.552 ± 0.087
CRF + BERTpred	**0.581 ± 0.065**	0.629 ± 0.087	0.558 ± 0.115
+Features	0.573 ± 0.058	0.635 ± 0.090	0.539 ± 0.100

Table 4.3: Evaluation results of mean model performance over 10 folds. Features are selected through step-wise greedy feature selection. **<0.01 *<0.05

performance is attained by an SVM model based on USE sentence vectors (+Emb), which were specifically designed for representing whole sentences. Perhaps sentence vectors perform better than BERT embeddings when the BERT model performs poorly ($F_1 = 0.366$). Additional research will be necessary to substantiate this.

Despite a lack of improvement in the F_1 score for the detection of ADEs and Strategies, an additional layer does seem to offer flexibility in tailoring the model towards a higher recall or precision. On the one hand, recall can be improved for two target classes by adding a non-sequential SVM layer with manual features to the BERT model. On the other hand, precision can be improved through the addition of a sequential CRF layer on top of BERT predictions for all target classes. Adding manually engineered features in addition to the sequential layer only improves the precision further for the detection of coping strategies. Our findings are thereby in line with Zubiaga et al. [354]. They speculated that sequential classifiers may take the surrounding context into account implicitly and therefore do not benefit from features representing thread context.

The only significant increase according to Wilcoxon signed rank tests is the precision for ADE detection. This may be related to the high variance between folds. Further research is necessary to validate these results and advance our understanding of how conversation-aware modeling can be best be used for relevance classification. We believe that this first study shows that this is a promising direction.

4.5.3. ANALYSIS OF SELECTED FEATURES

The variation in which features are selected per fold is large. Manual inspection of the selected features shows that features relating to the distribution of labels in the thread are chosen most often, especially the running count of negative and positive labels in the thread (CountNeg, CountPos), and the label of the previous post (PrevLbl) (see Table 4.1). Features of this type may therefore be the most promising for future work. The number of features that is chosen is more consistent; On average, 1 or 2 of the 11 features are chosen.

To further explore why certain features are chosen, we compute the correlations between the target label and the manually engineered features and between the BERT predictions and the manually engineered features (see Figure 4.2). We find, firstly, that features relating to the label distribution indeed appear to correlate most strongly with the ground truth labels. Secondly, the correlation between these features and the BERT predictions is often equal to or stronger than the respective correlation to the ground truth. This might indicate that this variance is already captured by the BERT model and therefore manually engineered features have little to add to the baseline model.

4.6. DISCUSSION

We find that the distribution of target posts across discussion threads is skewed and that within a conversational thread posts cluster together. Thus, our hypothesis that the probability of a target post occurring is related to the conversational structure appears valid.

In answer to **RQ1**, we find that adding a sequential CRF layer on top of a BERT model improves precision slightly, although only significantly so for ADE detection. In answer to **RQ2**, we find that the addition of manually engineered features representing

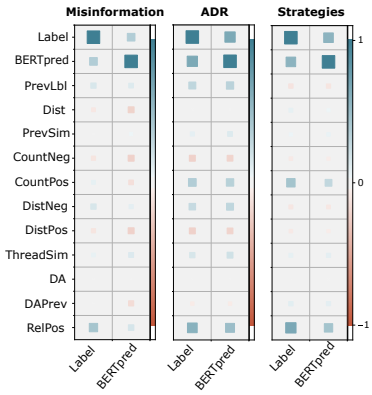


Figure 4.2: Correlation matrix of ground truth labels and BERT predictions with the manually engineered features. The size and colour of the squares corresponds to the strength of the correlation

thread context often does not aid performance. One consistent exception is when manually engineered features are combined with a non-sequential SVM layer on top of a BERT model. This combination can improve recall for all target classes, although not significantly so. An additional layer on top of a BERT model that is able to capture the thread context appears to offer flexibility in tailoring the model towards a higher recall or precision. In future work, we plan to investigate the benefit of including conversational context for other tasks such as concept normalization of ADEs.

For all data sets included in this study, a preselection of discussion threads was made prior to annotation to ensure a higher proportion of target posts. We expect that both sequential models and manually engineered features of thread context may prove more beneficial when such a preselection does not occur and the target class is even more imbalanced. Thus, our results may be an underestimation of the benefit of conversational context for finding 'needles in the haystack'.

Finally, our findings call into question the practice of splitting data into folds without taking the discussion context into account. In this study, we split the folds per discussion thread and we recommend others to consider doing so when dealing with multiple posts from the same thread, as neglecting to do so when there are dependencies between posts may bias model performance. This is especially important when threads contain duplicate posts.

PART II:

EXTRACTING ADVERSE DRUG EFFECTS (ADES)

For some minutes it puffed away without speaking, but at last it unfolded its arms, took the hookah out of its mouth again, and said, “So you think you’re changed, do you?”
...
“I’m afraid I am, sir,” said Alice; “I can’t remember things as I used—and I don’t keep the same size for ten minutes together!"

Lewis Caroll, *Alice in Wonderland*

5

TRANSFER LEARNING FOR ADE EXTRACTION FROM TWITTER

Edited from: **Anne Dirkson** & Suzan Verberne (2019), Transfer Learning for Health-related Twitter Data. Proceedings of the Fourth Social Media Mining for Health Applications (SMM4H) Workshop & Task. Association for Computational Linguistics. 89-92.

In this chapter, we introduce the use of transfer learning methods for extracting and normalizing adverse drug events from Twitter data. We also apply transfer learning to the task of identifying personal health mentions in health-related tweets.

Transfer learning is an umbrella term for methods that re-use a model trained on one (usually larger) set of data as a starting point for training a model for another task. These methods are especially promising for domains that suffer from a shortage of annotated data or resources, such as health-related social media.

This work was done as part of the 2019 Social Media Mining for Health Applications (SMM4H) Shared Task.

5.1. INTRODUCTION

Transfer learning is promising for NLP applications, as it enables the use of universal pre-trained language models (LMs) for domains that suffer from a shortage of annotated data or resources, such as health-related social media. Universal LMs have recently achieved state-of-the-art results on a range of NLP tasks, such as classification [137] and named entity recognition (NER) [3]. For the Shared Task of the 2019 Social Media Mining for Health Applications (SMM4H) workshop we focused on employing state-of-the-art transfer learning with universal LMs to investigate its potential in this domain.

5.2. TASK DESCRIPTIONS

The SMM4H shared task consisted of four subtasks:

ADE extraction The purpose of **Subtask 1** (S1) is to classify tweets as containing an adverse drug event (ADE) or not. Subsequently, these ADE mentions are extracted in **Subtask 2** (S2) and normalized to MedDRA concept IDs in **Subtask 3** (S3). MedDRA (Medical Dictionary for Regulatory Activities) is an international, standardized medical terminology.[1]

Personal Health Mention Extraction The goal of **Subtask 4** (S4) is to identify tweets that are personal health mentions, i.e., posts that mention a person who is affected as well as their specific condition [152], as opposed to posts discussing health issues in general. Generalisability to both future data and different health domains is evaluated by including data from the same domain collected years after the training data, as well as data from an entirely different disease domain.

5.3. OUR APPROACH

5.3.1. PREPROCESSING

We preprocessed all Twitter data using the lexical normalization pipeline by Sarker [261]. We also employed an in-house spelling correction method (see Chapter 2). Additionally, punctuation and non-UTF-8 characters were removed using regular expressions.

5.3.2. ADDITIONAL DATA

Personal Health Mentions For S4, the training data consists of data from one disease domain, namely influenza, in two contexts: having a flu infection and getting a flu vaccination. To improve generalisability, we supplemented this data with six labeled data sets from different disease domains [152]. We refer to this combined data set as S4+. For each subset, 10% was used for a combined validation set. For fine-tuning the ULMfit universal language model based on 28,595 Wikipedia articles (Wikitext-103) [200], the DIEGO Drug Chatter corpus [263] was combined with the data from S1 and S4+ to form a larger unsupervised corpus of health-related Twitter data ('TwitterHealth'). For S4, fine-tuning was also attempted with only the S4+ data.

[1] `https://www.meddra.org/`

	S1	S2*	S3	S4	S4+
Dev	-	130	76	-	-
Train	14,634	910	1,756	6,996	11,832
Validation	1,626	130	76	777	1,314
Test	5000	1000	1000	ND	ND

Table 5.1: Data sets. *Only tweets containing an ADE were used for developing the system. ND: Not disclosed

Concept Normalization The MedDRA concept names and their aliases in both MedDRA and the Consumer Health Vocabulary[2] were used to supplement the data from S3. This data set is hereafter called S3+.

5.3.3. TEXT CLASSIFICATION (S1 AND S4)

Text classification was performed with fast.ai ULMfit [137]. As recommended, the initial learning rate (LR) of 0.01 was determined manually by inspecting the log LR compared to the loss. Default language models were fine-tuned using AWD_LSTM [201] with (1) 1 cycle (LR = 0.01) for the last layer and then (2) 10 cycles (LR = 0.001) for all layers.

Subsequently, this model is used to train a classifier with F_1 as the metric, a dropout of 0.5 and a momentum of (0.8,0.7), in line with the recommendations. Training is done with (1) 1 cycle (LR = 0.02) on the last layer; (2) unfreezing of the second-to-last layer; (3) another cycle running from a 10-fold decrease of the previous LR to this LR divided by 2.6[4] (as recommended in the fast.ai MOOC).[3] This is repeated for the next layer and then for all layers. The last step consists of multiple cycles until F_1 starts to drop.

As an alternative classifier for S1, we used the absence of ADEs (noADE) according to the Bert embeddings NER method (see below) which was developed for the subsequent sub-task (S2) and aims to extract these ADE mentions. As a baseline for text classification, we used a Linear SVC with unigrams as features. The C parameter was tuned with a grid of 0.0001 to 1000 (steps of x10).

5.3.4. NAMED ENTITY RECOGNITION (S2)

We experimented with different combinations of state-of-the-art Flair embeddings [3], classical Glove embeddings and Bert embeddings [84] using the Flair package. We used pretrained Flair embeddings based on a mix of Web data, Wikipedia and subtitles; and the 'bert-base-uncased' variant of Bert embeddings. We also experimented with Flair embeddings combined with Glove embeddings (dimensionality of 100) based on FastText embeddings trained on Wikipedia (GloveWiki) or on Twitter data (GloveTwitter). Training for all embeddings was done with an initial LR of 0.1, batch size of 32, and max epochs set to 150.

As a baseline for NER, we used a CRF with the default L-BFGS training algorithm with Elastic Net regularization. As features for the CRF, we used the lower-cased word, its suffix, the word shape and its POS tag.[4]

[2] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/
[3] https://course.fast.ai/
[4] https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html

5.3.5. CONCEPT NORMALIZATION (S3)

Pretrained Glove embeddings were used to train document embeddings on the extracted ADE entities in the S3 data including or excluding the aliases from CHV (S3+) with concept IDs as labels. We used the default RNN in Flair with a hidden size of 512. Glove embeddings (dim = 100) were based on FastText embeddings trained on Wikipedia. Token embeddings were re-projected (dim = 256) before inputting to the RNN.

5.4. RESULTS

For all four subtasks, our best transfer learning system consistently performs better than the average over all runs submitted to SMM4H. For classifying ADE mentions, our overall best performing system is a ULMfit model trained on the TwitterHealth corpus (see Table 5.2). Yet, the highest recall is attained by using the absence of named entities (noADE) as a classifier. This is in line with our validation results (see Table 5.3). For extracting ADEs, our best system combines Bert with Flair embeddings without a separate classifier for sentences containing ADE mentions (see Table 5.4). However, using Bert embeddings alone *with* the ULMfit classifier from S1 appears to be more precise. During validation, we found that a combination of Glove embeddings (based on Twitter or Wikipedia) and Flair embeddings performed poorly compared to the submitted systems (see Table 5.5). For mapping the ADEs to MedDRA concepts, we only submitted one system with different preceding NER models (see Table 5.6), since adding the alias information (S3+) decreased both precision and recall (see Table 5.7). Our RNN document embeddings with only the S3 data, however, performed better than average. Lastly, for the classification of personal health mentions, our best classifier was a ULMfit model fine-tuned on the S4+ data (see Table 5.8), which outperformed the average result and the ULMfit model trained on the larger TwitterHealth corpus on all metrics. This system similarly outperformed the other ULMfit model on the validation data (see Table 5.9).

	Method	F_1 (range)	P	R
Average*		0.502 (0.331)	0.535	0.505
Run1	ULMfit[1]	**0.533**	**0.642**	0.455
Run2	noADE	0.418	0.284	**0.792**

Table 5.2: Results for ADE Classification (S1). *over all runs submitted [1]TwitterHealth data

Method	F_1	P	R
Baseline: Linear SVC (C=1.0)	0.475	0.526	0.433
ULMfit[1]	**0.574**	**0.574**	0.574
noADE	0.330	0.207	**0.823**

Table 5.3: Validation results for ADE classification (S1) [1]TwitterHealth data

		Relaxed			Strict		
	Method	F_1 (range)	P	R	F_1 (range)	P	R
Average*		0.538 (0.486)	0.513	0.615	0.317 (0.422)	0.303	0.358
Run1	Bert+Flair$^+$	**0.625**	0.555	**0.715**	**0.431**	0.381	**0.495**
Run2	Bert$^+$	0.622	0.560	0.701	0.427	0.382	0.484
Run3	Bert+ADEClassifier	0.604	**0.718**	0.521	0.417	**0.494**	0.360

Table 5.4: Results for ADE Extraction(S2). *over all runs submitted $^+$No separate classifier for sentences containing ADE

Method	Micro-F_1	P	R
Baseline: CRF	0.235	0.560	0.149
Flair+ GloveWiki	0.596	0.666	0.540
Flair+ GloveTwitter	0.577	0.655	0.515
Bert	0.640	**0.699**	0.590
Bert+Flair	**0.649**	**0.699**	**0.606**

Table 5.5: Validation results for ADE extraction (S2)

		Relaxed			Strict		
	Method	F_1 (range)	P	R	F_1 (range)	P	R
Average*		0.297 (0.242)	0.291	0.312	0.212 (0.247)	0.205	0.224
Run1$^+$	RNN Docemb.	**0.312**	**0.370**	0.270	**0.250**	**0.296**	0.216
Run2$^+$	RNN Docemb.	0.303	0.272	0.343	0.244	0.218	0.277
Run3$^+$	RNN Docemb.	0.302	0.267	**0.347**	0.246	0.218	**0.283**

Table 5.6: Results for concept normalization (S3). *over all runs submitted $^+$Runs same as S2 prior to concept normalization

Method	F_1	P	R
RNNDocembeddings with S3	**0.623**	**0.566**	**0.694**
RNNDocembeddings with S3+	0.253	0.171	0.482

Table 5.7: Validation results for concept normalization (S3)

Method			Acc. (range)	F_1 (range)	P	R
Average*			0.781 (0.263)	0.701 (0.464)	0.902	0.585
Run1	ULMfit S4+ data	*Domain1*	0.869	0.859	0.952	0.781
		Domain2	0.638	0.419	0.750	0.290
		Domain3	0.786	0.539	1.000	0.368
		Mean	**0.793**	**0.726**	**0.940**	**0.591**
Run2	ULMfit TwitterHealth data	*Domain1*	0.863	0.849	0.969	0.756
		Domain2	0.609	0.342	0.700	0.226
		Domain3	0.768	0.480	1.000	0.316
		Mean	0.786	0.716	0.928	0.583

Table 5.8: Results for personal health mention classification (S4). *over all runs submitted

Method	F_1	P	R
Baseline: Linear SVC (C=0.1)	0.615	0.678	0.572
ULMfit with S4+ data	**0.712**	**0.743**	**0.701**
ULMfit with TwitterHealth data	0.692	0.738	0.676

Table 5.9: Mean validation results for personal health mention classification (S4) averaged over eight data sets of S4+

5.5. CONCLUSIONS

Transfer learning using default settings offers above average results for various NLP tasks using health-related Twitter data. More research is necessary to investigate whether state-of-the-art performance may be possible with further domain-specific adaptation, for instance by tuning hyper-parameters, training embeddings on medical data or by dealing with domain-specific vocabulary absent in the language model.

6

VULNERABILITIES OF BERT FOR NAMED ENTITY RECOGNITION

Edited from: **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2021). Breaking BERT: Understanding its Vulnerabilities for Named Entity Recognition through Adversarial Attack. ArXiv. `https://arxiv.org/abs/2109.11308`

Both generic and domain-specific BERT models, like BioBERT and SciBERT, are widely used for natural language processing tasks including Named Entity Recognition (NER). In this chapter we investigate the vulnerability of BERT models to variation in the input data for NER using adversarial attack. Adversarial attack is the crafting of changes to the input data to deliberately try to fool the model.

We found that under these conditions BERT models are vulnerable to words in the local context of the entity being replaced with synonyms rarely seen during training. This type of variation resulted in 20.2 to 45.0% of entities being predicted completely wrong and another 29.3 to 53.3% of entities being predicted wrong partially. Often a single synonym replacement was sufficient to fool the model. The domain-specific BERT model trained from scratch (SciBERT) was more vulnerable than the original BERT model or the domain-specific model that retains the BERT vocabulary (BioBERT). We also found that BERT models are particularly vulnerable to entities that occur infrequently; BERT models could be fooled to predict 89.5% to 99.4% of entities wrongly when entities were replaced with more rare entities of the same type.

Our results chart the vulnerabilities of BERT models for NER and emphasize the importance of further research into uncovering and reducing these weaknesses.

6.1. INTRODUCTION

Self-attentive neural models, such as BERT [84], attain a high performance on a wide range of natural language processing (NLP) tasks. Despite their excellent performance, the robustness of BERT-based models is contested: Various studies [139, 143, 180, 290, 347] recently showed that BERT is vulnerable to adversarial attacks. Adversarial attacks are deliberate attempts to fool the model into giving the incorrect output by providing it with carefully crafted input samples, also called adversarial examples.

At present, the work on adversarial attack of Named Entity Recognition (NER) models is limited to a single study: Araujo et al. [11] attack biomedical BERT models by simulating spelling errors and replacing entities with their synonyms. They find that both attacks drastically reduce performance of these domain-specific BERT models on medical NER tasks.

Here, we aim to systematically test the robustness of BERT models for NER under severe stress conditions in order to investigate *which* variation in entities and entity contexts BERT models are most vulnerable to. This will, in turn, further our understanding of what these models do and do not learn. To do so, we propose two adversarial attack methods: replacing words in the context of entities with synonyms, and replacing the entities themselves with others of the same type. In contrast to previous work, the methods we propose are adaptive and specifically target BERT's weaknesses: We create adversarial examples by making the changes to the input that either manage to fool the model or bring it closest to making a mistake (i.e., lower the prediction score for the correct output) instead of randomly introducing noise or variation.

We address the following research questions:

1. How vulnerable are BERT models to adversarial attack on general and domain-specific NER?

2. To what extent is the vulnerability impacted by domain-specific training?

3. To which types of variation are BERT models for NER the most vulnerable?

Designing methods for direct adversarial attack of NER models poses additional challenges compared to the attack of text classification models as labels are predicted per word, sentences can contain multiple entities, and entities can contain multiple words. To ensure that labels remain accurate in our adversarial sentences, we constrain synonym replacements to non-entity words when altering the context of the entity (i.e., an *entity context attack*) and substitute entities only by entities of the same type when attacking the entity itself (i.e., an *entity attack*). In line with previous work [143, 180], we include a minimal semantic similarity threshold based on the Universal Sentence Encoder [57] to safeguard semantic consistency. Nonetheless, we acknowledge that for entity replacement adversarial examples may not be semantically consistent (e.g., if "Japan" is replaced with "China" in the sentence "Tokyo is the capital of Japan"). Although factually incorrect, the resulting sentences can be considered utility-preserving i.e., they retain their usefulness as valid input to the model [293], because BERT models should be able to identify that the final word in the sentence is a country even if it is not the correct country. In real-world data, sentences are not necessarily factually correct.

We assume a black-box setting, which means that the adversarial method has no knowledge of the data, parameters or model architecture [8]. This allows our methods to also be used for other neural architectures. Although we use English data, our methods are largely language-independent. Only an appropriate language model for synonym selection would be required.

The contributions of this chapter are twofold: We adapt existing adversarial attack methods to sequence labeling tasks and evaluate the vulnerability of general and domain-specific BERT models for NER. We make our code available for follow up research.[1]

6.2. RELATED WORK

In prior work, token-level black-box methods for adversarial attack have mainly been developed for classification and textual entailment [7]. Substituting tokens with their synonyms is the most popular choice for perturbing at the token level. Synonyms are often found using nearest neighbors in a word embeddings model. One major challenge when selecting synonyms based on word embeddings is that antonyms will also be close in the embedding space. To solve this issue, recent studies [139, 143, 181] require a minimal semantic similarity between the generated and original sentence. Additionally, some methods [8, 143] use word embeddings with additional synonymy constraints [211]. We will employ both techniques.

Approaches also differ in how they select the word that is perturbed: while some select words randomly, it is more common to use the importance of the word for the output [7]. The importance is often operationalized as the difference in output before and after removing the word. We follow this approach in our method.

Most adversarial attack methods were developed for attacking recurrent neural models. However, there has been a growing interest in attacking self-attentive models in the last year [11, 18, 139, 143, 180, 209, 290, 347]. Nonetheless, the only study that has attacked BERT models for NER is the study by Araujo et al. [11]. They perform two types of character-level (i.e., swapping letters and replacing letters with adjacent keys on the keyboard) and one type of token-level perturbation (i.e., replacing entities with their synonyms). The authors find that biomedical BERT models perform far worse on NER tasks when spelling mistakes are included or synonyms of entities are used.

Our work differs from Araujo et al. [11] in three ways. First, our adversarial examples are generated based on the importance of words for the correct output instead of through random changes. Thereby, we are able to test the robustness of BERT under the most severe stress conditions, while Araujo et al. [11] evaluate the scenario where the input data is noisy due to spelling mistakes and the use of synonyms. Second, we analyze the impact of replacing entities with others of the same type (e.g., 'France' with 'Britain') and replacing words in the context of entities (see Table 6.1 for an example) instead of replacing entities with their synonyms. Third, we will test our method on the original BERT model as well as biomedical BERT models and on both generic and biomedical NER.

[1] Our code (BSD-3 Clause license), URLs to the benchmark data and the annotation guideline are available at: `https://github.com/AnneDirkson/breakingBERT`

The	Republic	of	China	**bought**	flowers
<O>	<B-LOC>	<I-LOC>	<I-LOC>	<O>	<O>
The	Republic	of	China	**purchased**	flowers
<O>	<O>	<O>	<B-LOC>	<O>	<O>

Table 6.1: Example of a partial success. The **bold** word has been changed to attack the entity 'Republic of China'.

6.3. METHODS

In this section, we describe two methods for generating adversarial examples designed to fool NER models, namely through (1) synonym replacements in the entity context (*entity context attack*) and (2) entity replacement (*entity attack*). These are described in Sections 6.3.1 and 6.3.1, respectively.

6.3.1. AIM OF THE ATTACKS

We aim to generate adversarial examples in which a target entity is no longer recognized correctly. This can be either because it has become a false negative or it has been assigned a different entity type. The attack is considered a success when the correct label has been changed, unless it has changed from the I-tag to the B-tag of the same entity type under the IOB schema. An example of this can be seen in Table 6.1: Here, the start of the entity is mislabeled, but the last part of the entity is still recognized. We consider this a partial success.

METRICS FOR EVALUATION

The success of the attack and thus the vulnerability of the model is evaluated by the percentage of entities that were originally correctly labeled but are mislabeled after attack. For *entity context attacks* entities can also be partially mislabeled i.e., only some words in the entity are mislabeled. This is captured by the partial success rate: the percentage of entities for which not the whole entity but at least half of the entity is mislabeled. For context attacks we also include a metric ('Words perturbed') to measure how much the sentence needed to be changed before the attack was successful: the average percentage of words that were perturbed out of the total amount of out-of-mention words in the sentences. This metric functions as a proxy for how difficult it is to fool the model [143].

ENTITY CONTEXT ATTACK

To investigate the impact of the context on the correct labeling of the entity, we adapt the method of Jin et al. [143], which was designed for text classification, to sequence labeling tasks. For each entity in the sentence, a separate adversarial example is created, as models may rely on different contextual words for different entities.

Step 1: Choosing the word to perturb We use the importance ranking function shown in Equation 6.1 to rank words based on their importance for assigning the correct label to the entity. The importance (I_w) of a word w for a token in the entity is calculated as the change in the predictions (logits[2]) of the *correct* label before and after deleting the word

[2]Here logits refers to the vector of raw (non-normalized) predictions that the BERT model generates

Domain	Dataset	Entity types	Dev.	Train	Test	Eval. subset* (# Entities)
General	CoNLL-2003	Person, Location, Organization, Miscellaneous	3,466	14,987	3,684	500 (1,343)
General	W-NUT 2017	Person, Location, Corporation, Product, Creative-work, Group	1,008	1,000	1,287	500 (787)
Biomedical	BC5CDR	Disease, Chemical	4,580	4,559	4,796	500 (1,221)
Biomedical	NCBI disease	Disease	922	5,432	939	487 (897)

Table 6.2: Size of the data sets (number of sentences). *This subset is used for automatic evaluation

from the sentence [143]. If the deletion of the word leads to an *incorrect* label for the entity token, the importance of the word is increased by adding the raw prediction score (logits) attributed to the incorrect label.

If the entity consists of multiple words, we rank words based on their summed importance for correctly labeling each of the individual words in the entity. Besides stop words, we also exclude other entities from being perturbed. We adapt the function so that for any word with an I-tag, both the I and B label of the entity type (e.g., B-PER and I-PER) are considered correct.

Given a sentence of n words $X = w_1, w_2, ..., w_n$, the importance (I_w) of a word w for a token in the entity is formally defined as:

$$\begin{aligned} I_w = & F_Y(X) - F_Y(X_{-w}) \\ & \text{if} F(X_{-w}) = Y \vee (F(X) = Y_I \wedge F(X_{-w}) = Y_B) \\ & F_Y(X) - F_Y(X_{-w}) + F_{\_\tilde{Y}}(X_{-w}) - F_{\overline{\tilde{Y}}}(X) \\ & \text{if} F(X_{-w}) \neq Y \end{aligned} \tag{6.1}$$

where F_Y is the prediction score for the correct label, $F_{\tilde{Y}}$ is the prediction score of the predicted label, F is the predicted label, Y is the correct label, Y_I is the I-tag version of the correct label, Y_B is the B-tag version of the correct label and X_{-w} is the sentence X after deleting the word w.

Step 2: Gathering synonyms For each word, we select synonyms from the Paragram-SL999 word vectors [211] with a similarity to the original word above the threshold δ. Mrkšić et al. [211] injected antonymy and synonymy constraints into the vector space representation to specifically gear the embeddings space towards synonymy. These embeddings achieved state-of-the-art performance on SimLex-999 [134] and were also used by Jin et al. [143] and Alzantot et al. [8]. We chose 0.5 as the minimal similarity threshold δ for synonym selection in contrast to the threshold of 0 used by previous work to better guarantee semantic similarity. Regardless of δ, a maximum of 50 synonyms are selected. Examples of word pairs with a δ above 0.5 are 'bought' and 'obtained'; and

'cat' and 'puss'. Below this threshold but within the first 50 synonyms fall 'bought' and 'forfeited'; and 'cat' and 'dustpan'.

Step 3: Filtering synonyms To preserve syntax, synonyms must have the same POS tag as the original word. If the data did not include POS tags, we added POS tags using NLTK. We filter the generated sentences for a sufficiently high semantic similarity to the original sentence. Semantic similarity is calculated with the Universal Sentence Encoder (USE) [57]. We exclude synonyms that result in sentences falling below the similarity threshold ϵ.

Step 4: Selecting the final synonym After filtering, we check whether any of the synonyms can change the entity label(s) fully. If there are multiple options, we select the one that leads to the highest sentence similarity (ϵ) to the original sentence. If there are none, we select the synonym which can reduce the (summed) prediction scores of the correct label(s) the most. If no synonyms are left after filtering or none manage to reduce the prediction scores, we do not replace the original word.

For multi-word entities, it is possible that a synonym changes some, but not all, labels. From the synonyms that change the most labels, we select the one that leads to the largest reduction in the (summed) prediction scores for the unchanged labels (i.e., the labels that are still predicted correctly by the model) (see Equation 6.1). Which labels are still correct can differ per synonym.

Finalizing the adversarial examples For each word in this ranking, we go through step 2-4 until either the label(s) of the entity have been changed fully or there are no words left to perturb. Once the attack is partially successful, only the predictions of the not yet incorrectly labeled words in the entity are considered for subsequent iterations.

ENTITY ATTACK

To explore to what extent the models rely on the words of the entity itself, we replace the entity with one of the same type, e.g., we change 'Japan' to another location. If a sentence contains multiple entities, an adversarial sentence is generated for each entity. The replacement entity is selected from a list of all entities in the data that are of the same type. We randomly select 50 candidate replacements from the entity list. We exclude candidates that result in a sentence that is too semantically dissimilar from the original (i.e., falling below the semantic similarity threshold ϵ). For the remaining candidate entities, we check if the predicted label is incorrect. If so, we select the successful attack replacement with the highest semantic similarity at the sentence level. If not, the attack was unsuccessful.

6.4. EXPERIMENTS

6.4.1. DATA

We use two general-domain English NER data sets for evaluating our method: the CoNLL-2003 data [298] and the W-NUT 2017 data [83]. The goal of the latter was to investigate recognition of unusual, previously-unseen entities in the context of online discussions.

	CoNLL-2003	W-NUT 2017	NCBI-disease	BC5CDR
	BERT	BERT	BERT	BERT
Success rate (%)	36.3 ± 0.612	42.2 ± 0.677	20.2 ± 0.443	38.8 ± 0.862
Of which:				
– Missed entity (%)	47.4 ± 2.9	61.3 ± 5.1	100	90.4 ± 4.2
– Entity type error (%)	52.6 ± 2.9	38.7 ± 5.1	0	9.6 ± 4.2
Partial success rate (%)	51.0 ± 0.465	51.6 ± 1.6	29.3 ± 0.841	45.9 ± 1.1
Median semantic similarity	0.928 ± 0.009	0.926 ± 0.017	0.920 ± 0.040	0.946 ± 0.002
Words perturbed (%)	15.6 ± 0.306	13.2 ± 1.2	12.4 ± 1.0	12.3 ± 0.04

Table 6.3: Automatic evaluation results for context attacks on BERT models. Results are the mean of the three models

	NCBI-disease		BC5CDR	
	BioBERT	SciBERT	BioBERT	SciBERT
Success rate (%)	20.9 ± 0.762	26.4 ± 0.875	37.9 ± 0.388	45.0 ± 0.665
Of which:				
– Missed entity (%)	100	100	86.5 ± 2.3	87.1 ± 2.3
– Entity type error (%)	0	0	13.5 ± 2.3	12.9 ± 2.3
Partial success rate (%)	30.1 ± 1.032	39.0 ± 0.954	44.8 ± 0.331	53.3 ± 0.821
Median semantic similarity	0.921 ± 0.031	0.936 ± 0.030	0.921 ± 0.003	0.936 ± 0.008
Words perturbed (%)	9.1 ± 2.5	8.7 ± 2.9	9.8 ± 0.3	8.5 ± 0.8

Table 6.4: Automatic evaluation results for context attacks on biomedical BERT models. Results are the mean of the three models

Additionally, we use two English data sets from the biomedical domain: BC5CDR [179] and the NCBI disease corpus [90]. Both data sets have been used to evaluate domain-specific BERT models for NER in the biomedical domain [28, 174, 232]. See Table 6.2 for more details on the data sets.

6.4.2. TARGET MODELS

We fine-tune three BERT models (base-cased) for each data set with different initialization seeds (1, 2 & 4) using the Huggingface implementation [339]. We set the learning rate at 5×10^{-5} and optimized the number of epochs (3 or 4) as recommended in Devlin et al. [84] for NER. We select the number of epochs based on the first BERT model (seed=1). We find that for all data sets except W-NUT 2017, 4 epochs is optimal.

For the biomedical data sets, we additionally fine-tune two domain-specific BERT models, BioBERT (base-cased) [174] and SciBERT (scivocab-cased) [28]. Each model is trained in three-fold (seeds are 1, 2 & 4).

6.4.3. EVALUATION OF ADVERSARIAL ATTACKS

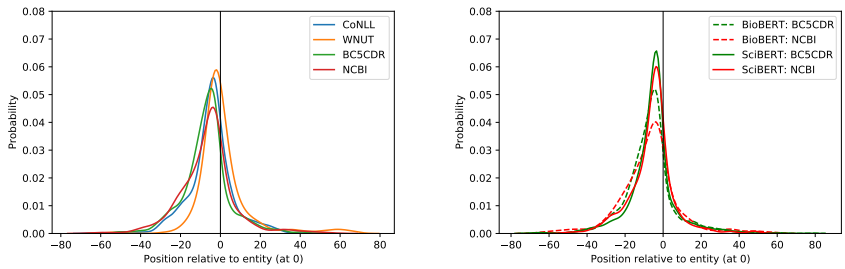
Automatic evaluation We randomly select 500 eligible sentences from each test set. Table 6.2 shows the number of entities in each subset. We considered sentences to be eligible if they contain at least one entity and one verb. For the NCBI-disease data, only 487 sentences fulfill these criteria.

We use models trained on the original training and development data to perform NER on the selected subset of the test data. We then generate one adversarial example for each entity in the sentence that was initially predicted correctly. We evaluate to what degree models are fooled only for entities that were predicted correctly in the original sentences. We set the semantic similarity threshold at $\epsilon = 0.8$ following Li et al. [181]. Experiments are run on a GPU machine (NVIDIA Tesla K80). An experiment of three runs (one model on one data set) on one GPU will take roughly 20-24hrs. The models have 110 M parameters.

Human evaluation To evaluate the quality of our adversarial examples from the CoNLL-2003 and BC5CDR data, 100 original sentences and 100 adversarial sentences from each type of attack are scored for grammaticality by human judges. Grammaticality is evaluated on a five-point scale following the reading comprehension benchmark DUC2006 [74]. Our annotators are four volunteering PhD students from our lab who have a background in linguistics[3]: two for each data set with 20% overlap. We choose to present annotators with different original sentences than the ones on which the adversarial sentences they evaluate are based to prevent bias.

6.5. RESULTS

6.5.1. ENTITY CONTEXT ATTACK



(a) BERT

(b) BioBERT and SciBERT

Figure 6.1: Distance of successful synonym replacements relative to the entity (at 0)

With an adversarial context attack, BERT models can be fooled into predicting entities partially or fully wrong (Partial + full success rate) for 87.3% and 93.8% of entities for CoNLL and W-NUT respectively. Moreover, for over 75% of the cases the BERT models were fooled by a single change.

[3]We opted for linguists as they are more acquainted with assessing grammaticality than biomedical domain experts

SciBERT appears more vulnerable than BERT, both to completely being fooled (+6.2 and +6.2% point) and being fooled partially (+9.7 and +7.4 % point) by context attacks. The domain-specific models were also often fooled by only one word being replaced with its synonym; BioBERT was fooled by a single change 65 and 75% of the time whereas SciBERT was fooled by a single change 68 and 76% of the time.

We analyzed the sentence statistics for successful and failed attacks. Specifically for BERT models, we see that the following cases are more vulnerable to attacks: longer sentences; sentences with more words that could be replaced by synonyms; and shorter entities. Manual analysis of successful attacks reveals that BERT models are vulnerable when common words are replaced by rare synonyms (e.g., replacing 'healthy' by 'salubrious').

Figure 6.1a shows where in the sentence changes have occurred in order to fool BERT. BERT models seem most vulnerable to changes in the local context of entities: only 1-2 words left or right of the entity. Manual analysis revealed that these words are often verbs. Although less influential, long distance context does appear to be used for predicting entities in some cases. We manually inspected sentences with long distance changes (>20 words). Lists stood out as a prime example of a sentence type for which long distance context is important (e.g., "The ministry said the group consisted of 13 nuns, seven Italians, and six Zaireans, and four priests, two from Belgium, one from Spain and one from Zambia.").

For the BioBERT model, the distribution is strikingly similar to that of the original BERT model (see Figure 6.1b). This is likely due to either the vocabulary or the training data[4] that these models share. SciBERT models which share neither training data nor vocabulary with the original BERT model are even more vulnerable to changes in the local context of the entity (see Figure 6.1b).

6.5.2. EVALUATING THE NECESSITY OF IMPORTANCE RANKING

To investigate the effect of adding the word importance ranking to the entity context attack, we perform an ablation study on the CoNLL-2003 test set. As can be seen in Table 6.5, removing the word importance ranking leads to a stark drop in both the average full success of adversarial attacks (from 37.3% to 9.5%) and the average partial success rate (from 52.8% to 20.1%). The number of words that need to be perturbed also drops, by 6.9% point, meaning that attacks require fewer synonym replacements on average to be successful. Thus, it appears that the word importance ranking is crucial to the success of the adversarial attack algorithm.

6.5.3. ENTITY ATTACK

The main results of adversarial *entity* attack on BERT models are presented in Table 6.6. BERT models appear highly vulnerable to adversarial attacks on the entities themselves despite the high similarity between adversarial and original sentences. On average, BERT models are fooled for 97.5% of entities that were initially predicted correctly on the CoNLL data and 89.2% on W-NUT data. BERT models appear even more vulnerable to entity attacks on domain-specific data with success rates above 99%.

[4] BioBERT includes all the original BERT training data as well as additional domain-specific data

	Importance ranking	
	Yes	No
Success rate (%)	37.3 ± 0.515	9.5 ± 4.3
Partial success rate (%)	52.8 ± 0.356	20.1 ± 9.2
Semantic similarity	0.922 ± 0.006	0.983 ± 0.006
Words perturbed (%)	13.8 ± 0.238	6.9 ± 2.2

Table 6.5: Comparison of context attacks with and without importance ranking on CoNLL-2003 data

	CoNLL-2003	W-NUT 2017	NCBI-disease	BC5CDR
	BERT	BERT	BERT	BERT
Success rate (%)	97.5 ± 0.037	89.5 ± 0.886	99.2 ± 0.114	99.4 ± 0.073
Of which:				
– Missed entity(%)	21.3 ± 12.3	71.4 ± 1.9	100	86.1 ± 0.5
– Entity type error(%)	78.8 ± 12.3	28.6 ± 1.9	0	13.9 ± 0.5
Median semantic similarity	0.959 ± 0.001	0.928 ± 0.003	0.952 ± 0.001	0.962 ± 0.000

Table 6.6: Automatic evaluation results for entity attacks on BERT models. Results are the mean of three models.

Table 6.7 shows that domain-specific BERT models do not resolve this issue. They are also highly vulnerable with over 99% of all initially correctly predicted entities now predicted incorrectly. The high success rates of entity attacks both on general domain and domain-specific data suggest that BERT models, similar to traditional models, are unable to predict entities correctly based solely on the context of the entity. Replacing the entity word itself with another of the same entity type, with the context unchanged, can easily fool the model. This suggests a strong dependency on the entities that the model has seen previously, making these models vulnerable to new or emergent entities.

This is corroborated by an analysis of which entities were chosen in successful attacks. For all BERT models and all data sets, except for the CoNLL data, these entities are significantly less frequent in training and development data than the original entities according to Wilcoxon signed rank tests ($p < 0.001$).

A possible explanation for why BERT models for CoNLL are the exception is that there is a stronger match between the pretraining data and the data at hand than for the other data sets. This may make the model less vulnerable to infrequent entities, despite not being less vulnerable to entity replacement overall. Manual inspection further revealed that BERT models appear to be sensitive to the capitalization of entities (e.g., BERT models trained on CoNLL were fooled by transforming 'New York' to 'NEW YORK').

6.5.4. RESULTS OF HUMAN EVALUATION

On CoNLL-2003, the annotators have a fair inter-annotator agreement (weighted κ = 0.353). On BC5CDR, the inter-annotator agreement is slight (weighted κ = 0.177). Investigation of the annotations reveals that this is most likely because biomedical sentences are more difficult to assess for laymen. Because of the limited agreement, we report grammaticality assessments per annotator.

	NCBI-disease		BC5CDR	
	BioBERT	SciBERT	BioBERT	SciBERT
Success rate (%)	99.2 ± 0.259	99.4 ± 0.054	99.4 ± 0.070	99.3 ± 0.089
Of which:				
– Missed entity(%)	100	100	94.0 ± 0.7	91.7 ± 1.5
– Entity type error(%)	0	0	6.0 ± 0.7	8.3 ± 1.5
Median semantic similarity	0.955 ± 0.001	0.953 ± 0.002	0.961 ± 0.001	0.962 ± 0.001

Table 6.7: Automatic evaluation results for entity attacks on biomedical BERT models. Results are the mean of three models.

	CoNLL		BC5CDR	
Annotator	1	2	3	4
Original	3.51	4.34	4.43	4.78
After context attack	3.05*	3.68**	3.86**	4.35**
After entity attack	3.30	4.37	3.85	4.67

Table 6.8: Mean grammaticality of the original and adversarial sentences. $^{*}p < 0.05$ $^{**}p < 0.01$ compared to the original sentences according to Mann-Whitney U tests.

Table 6.8 shows that although entity attacks do not significantly alter the grammaticality of the sentences, attacks on the context of the entity do. Although this reduction is consistent across data sets, the mean grammaticality of the adversarial sentences remains above 3 (acceptable) and the mean absolute reduction is less than a full point.

6.6. DISCUSSION AND LIMITATIONS

We manually analyzed the generated adversarial examples and found that our adversarial examples are susceptible to word sense ambiguity. For example, the top 50 synonyms for 'surfed' in 'surfed the Internet' includes both correct synonyms like 'googled' and incorrect ones like 'paddled'. There are also some cases where adversarial examples suffer from foreign words in the Paragram-SL999 word vectors [211]. Occasionally synonyms are not English words (e.g., 'number' to 'nombre'), or synonym choice is influenced by words that occur in multiple languages e.g., 'vie' in 'to vie for top UN post' is replaced with 'existence' which is a synonym of the French 'vie' (i.e., life).

Furthermore, our adversarial examples are susceptible to grammatical errors. Grammatically poor adversarial sentences often suffer from changes from verbs to nouns or vice versa that are not caught by the POS-filter (e.g., 'open' to 'openness' and 'influence' to 'implication'). These cases may be particularly difficult as 'open' and 'influence' can be both a verb and an adjective or noun. Another common error is singular-plural inconsistencies (e.g., 'one dossiers'). To mitigate these issues, future work could focus on removing non-English words from the embedding space, and altering how the POS-tag of the synonym is determined.

We find that semantic consistency can be an issue with broad entity types like location when attacking the entity itself. For example, in one case the country "U.S." is replaced by the village "Tavildara" (in Tajikistan). For more specific entity types like Disease, Chemical

or Person we do not encounter inconsistencies with subtypes of an entity category. On the contrary, often replacements are semantically close to the original. For instance, the anti-epileptic drug "clonazepam" was replaced by the anti-epileptic drug "lorazepam" and "Washington" in "Washington administration" was replaced by "Clinton".

Moreover, there are some caveats to keep in mind when interpreting weaknesses based on successful attacks. The architecture of self-attentive models means that the attention weight of a word is context-dependent. Thus, if changing that word fools the model, this might only be true in that context. Additionally, if multiple words were changed for a successful attack, their interaction may contribute to the success and it cannot simply be interpreted as caused by this combination of words.

6.7. CONCLUSIONS

We studied the vulnerability of BERT models in NER tasks under a black-box setting. Our experiments show that BERT models can be fooled by changes in single context words being replaced by their synonyms. They are even more vulnerable to entities being replaced by less frequent entities of the same type.

Our analysis of BERT's vulnerabilities can inform fruitful directions for future research. Firstly, our results reveal that rare or emergent entities remain a problem for both generic and domain-specific NER models. Consequently, we recommend further research into zero or few-shot learning. Moreover, the masking of entities during fine-tuning may be an interesting avenue for research. Secondly, BERT models also appear vulnerable to words it has not seen or rarely seen during training in the entity context. To combat this vulnerability, the use of adversarial examples designed specifically to include more infrequently used words could be explored. Another possible avenue for research could be alternative pre-training schemes for BERT such as curriculum learning [99]. Thirdly, we find that SciBERT is more vulnerable to changes in the entity context than BioBERT or BERT. This may be due to the domain-specific biomedical vocabulary that SciBERT employs, which could make it more vulnerable to out-of-entity words being replaced by more common English terms. This trade-off between robustness and domain-specificity of BERT models may be another worthwhile research direction.

We consider our work to be a step towards understanding to what extent BERT models for NER are vulnerable to token-level changes and to which changes they are most vulnerable. We hope others will build on our work to further our insight into self-attentive models and to mitigate these vulnerabilities.

7

FUZZY REPRESENTATION OF DISCONTINUOUS ENTITIES

Edited from: **Anne Dirkson**, Suzan Verberne and Wessel Kraaij (2021), FuzzyBIO: A proposal for Fuzzy Representation of Discontinuous Entities, Proceedings of the 12th Health Text Mining and Information Analysis at EACL 2021.

For the task of extracting entities (Named Entity Recognition) such as side effects, words in the data are generally represented using the BIO scheme (B-beginning, I-inside, and O-outside) to indicate whether a word is the beginning of an entity, inside an entity or not in an entity. However, this representation does not provide a way of dealing with discontinuous entities.

As discontinuous entities occur commonly in the biomedical domain, expansions of the BIO scheme that can handle these entity types are often used (i.e., BIOHD). However, the extra tag types make the NER task more difficult to learn. Therefore, in this chapter, we present FuzzyBIO as an alternative simpler representation scheme in which discontinuous entities are transformed into continuous sequences by annotating the non-entity words in between.

We focus on the task of Adverse Drug Event extraction and normalization to compare FuzzyBIO with BIOHD. We find that FuzzyBIO improves the recall of NER for two of the three data sets and results in a higher percentage of correctly identified disjoint and composite entities for all data sets. Using FuzzyBIO also improves end-to-end performance for continuous and composite entities in two of the three data sets. Since FuzzyBIO improves performance for some data sets and the conversion from BIOHD to FuzzyBIO is straightforward, we recommend investigating which is more effective for any data set containing discontinuous entities.

7.1. INTRODUCTION

Adverse Drug Events (ADEs), harmful reactions that result from the intake of medication, pose a major health concern [340]. Due to the limitations of clinical trials on the one hand [274] and reporting systems after release on the market on the other hand [130], many ADEs remain undiscovered. Therefore, both social media and clinical reports are being explored by the research community as alternative information sources for the semi-automatic discovery of ADEs [171, 262].

One particular challenge for the extraction of ADEs from text is the presence of discontinuous entities. These can be either composite entities (i.e., some words belong to multiple entities), such as '*lack of sleep* and *appetite*', or disjoint entities (i.e., split entities), such as '*eyes* are feeling *dry*'. These phenomena occur more commonly in the clinical than general domain. In fact, Tang et al. [295] reported that discontinuous mentions in clinical text account for about 10% of all ADE mentions. None of the traditional versions of the BIO representation scheme (B: beginning of entity, I: inside entity and O: outside entity) or common extensions such as IOBES (E: end of entity, S: singleton entity)[1] were designed to handle such mentions [241]. Therefore, Tang et al. [295] proposed extending the BIO scheme with two additional tags: the 'H' for words shared by multiple mentions and 'D' for parts of discontinuous mentions not shared by other mentions. This resulted in four new tag types (HB-, HI-, DB- and DI-). Their BIOHD representation was broadly adopted by the community [151, 275, 353]. Table 7.1 shows examples of concepts represented with the BIOHD scheme.

Sentence 1	**Muscles**	are	constantly	**quivering**	!				
BIOHD	DB	O	O	DI	O				
FuzzyBIO	B	I	I	I	O				
Sentence 2	I	have	**pain**	**in**	**my**	**hands**	and	**upper**	**arms**
BIOHD	O	O	HB	HI	HI	DB	O	DB	DI
FuzzyBIO	O	O	B	I	I	I	I	I	I

Table 7.1: Examples of discontinuous disjoint (sentence 1) and composite (sentence 2) ADE mentions represented by the BIOHD and FuzzyBIO schemes.

Although the BIOHD scheme allows for precise representation of entities, the extra tag types make the task more difficult for models to learn. Straightforward BIO rules such as 'an entity always starts with a B' are no longer valid under the BIOHD scheme. In this chapter we argue that a more simple BIO representation in which discontinuous entities are transformed into continuous sequences by including all non-entity tokens in between would improve ADE extraction by being easier to learn and reintroducing these straightforward rules. We coin this representation FuzzyBIO. Some examples of entities represented with BIOHD and FuzzyBIO can be seen in Table 7.1.

Aside from improving extraction, using FuzzyBIO instead of BIOHD may also improve subsequent concept normalization, in which ADEs are linked to standardized medical concepts (e.g., 'can't fall asleep' to the concept 'insomnia' with concept identifier 193462001 in the medical ontology SNOMED-CT). This step is essential for aggregating and thus quantifying the prevalence of ADEs. As current normalization methods are

[1] This scheme is also called BIOES, BILOU or BMEWO

mostly hampered by errors made during extraction [335], this step may also benefit from simplification of the representation scheme.

Thus, we address two research questions:

RQ1 To what extent can a fuzzy continuous representation of discontinuous entities improve NER of ADEs? (intrinsic evaluation)

RQ2 To what extent can this fuzzy representation benefit end-to-end ADE extraction? (extrinsic evaluation)

In this chapter, we present FuzzyBIO, a fuzzy continuous BIO representation of discontinuous entities. Moreover, we show it is beneficial for end-to-end ADE discovery. Our representation is applicable to other domains as well. We release our code for the purpose of follow-up research.[2]

7.2. RELATED WORK

The first shared task to deal with discontinuous medical entities was SemEval 2014 Task 7 [241]. Prior to this task, discontinuous entities were often excluded (e.g., Uzuner et al. [315]) or each part was represented as a separate continuous entity and later reassembled [203]. Various representations were proposed but the only one able to distinguish between those that share a head word (i.e composite entities) and those that do not (e.g., disjoint entities) was the BIOHD scheme [351].

This scheme was later analyzed in more detail and compared to two baseline approaches: (1) ignoring all discontinuous entities and (2) representing separate parts of discontinuous entities as individual entities [295]. In comparison to the baseline approaches, the BIOHD scheme could improve recognition of both discontinuous but also continuous entities, likely due to its ability to distinguish between the two.

Tang et al. [295] also proposed a further extension (BIOHD1234) in which numbers were added to refer to which entity a non-head ('D') entity should be combined with[3], effectively expanding the scheme from 7 to 13 tags. This representation was able to outperform BIOHD due to its ability to correctly represent multiple discontinuous entities and discontinuous entities with more than one non-head part. However, as neither BIOHD nor BIOHD1234 could handle multiple head entities in one sentence, Tang et al. [296] proposed a multi-label BIO representation in which tokens can be labeled with more than one tag, and each tag corresponds to one entity. For NER of adverse drug events, this novel representation managed to outperform BIOHD. Similarly, Shang et al. [272] allowed for multiple labels per token for extracting disorders from scientific articles.

Despite its limitations [296], the BIOHD scheme is commonly adopted [151, 275, 353]. We propose an alternative, simpler representation scheme that could improve extraction by being easier to learn.

[2] Code is available at: `https://github.com/AnneDirkson/FuzzyBIO`

[3] 1 and 2 denote nearest head and non-head entity on the left, and 3 and 4 denote nearest head and non-head entity on the right

7.3. METHODS

7.3.1. THE FUZZYBIO REPRESENTATION SCHEME

As displayed in Table 7.1, FuzzyBIO transforms discontinuous into continuous entities by annotating all tokens in between.[4] Composite entities are combined if they share an entity head. We realize this compresses two separate entities into one (e.g., the entities 'pain in my hands' and 'pain in my upper arms' in Table 7.1). However, this does not pose a problem to normalization, as the state-of-the-art normalization method [291] includes heuristic rules to split composite entities prior to normalization.

7.3.2. NAMED ENTITY RECOGNITION OF ADES

For the NER task itself, we opt for distilBERT (base-cased), a lighter, more computationally efficient version of BERT [260]. We use a one-cycle learning rate (LR) policy [281] with a maximum LR of 0.01. For each fold in the 10-fold cross-validation (CV), we select either 3 or 4 epochs based on the validation data. We use the Huggingface implementation [339] with the wrapper ktrain [195] to train our models with the initialization seed set to 1.

7.3.3. CONCEPT NORMALIZATION OF ADES

For normalization, we use the state-of-the-art BioSyn method with default parameters [291, 304]. It is possible to provide composite entities as input, as this method splits composite entities prior to normalization using the heuristics by D' Souza and Ng [70].

Our target ontology is SNOMED-CT[5]. As SNOMED-CT is too extensive for our purpose, we aim to map SNOMED concepts in our training data to a curated subset of SNOMED, the CORE Problem List Subset[6], before training the normalization model. If there is a direct mapping in the community based mappings in BioPortal [220] between the original concept and a CORE concept or the parent of the concept is in the CORE (e.g., 'moderate anxiety' to 'anxiety'), we map the mention to the respective CORE concept. We include all concepts of the CORE subset and all concepts that could not be mapped to a CORE concept in the data as candidates. Synonyms for each concept are retrieved from the community based mappings in BioPortal [220] using the REST API and from the UMLS using pymedtermino [170].

7.3.4. EVALUATION

For evaluating the NER models on a token level, our metrics are lenient and ignore the prefixes (B- I- H- D-). Additionally, we evaluate performance on an entity level. Following Magge et al. [194], an entity is considered a true positive if any part of the annotated adverse drug event is correctly identified (i.e., overlaps with the predicted ADE text). We evaluate the end-to-end performance by calculating how many entities were both extracted during NER and normalized to the correct SNOMED-CT concept.

[4] In our data, we did not find any cases where an entity was lost because it was in between two parts of another discontinuous entity. Nonetheless, this is theoretically possible and poses a potential limitation.

[5] `https://www.nlm.nih.gov/healthit/snomedct/index.html`

[6] `https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html`

Entities	CADEC	PsyTAR	CLEF
Continuous	5.360	4.508	16.261
Discontinuous			
– Disjoint	100	225	909
– Composite	828	70	286

Table 7.2: Size of data sets.

7.4. DATA

We use two data sets of social media posts annotated for ADEs: CADEC [151] and PsyTAR [353]. The former was also used by Tang et al. [296]. Both contain posts from medical fora on AskaPatient.com. Additionally, we used a data set of clinical records annotated for disorder mentions, namely the SemEval 2014 Task 7 data [241] that builds on the CLEF eHealth 2013 corpus used by Tang et al. [295]. See Table 7.2 for more details. Data sets were split into 10 folds stratified on the presence of an ADE. For PsyTAR, we chose sentences as units as they were annotated separately. For the CLEF data set, each document was split into sequences of 5 sentences for the NER task, because of memory restrictions on the input length for BERT models.

7.5. RESULTS

		Token level			Entity level recall		
Data	Scheme	Micro F_1	P	R	Continuous	Disjoint	Composite
CADEC	BIOHD	0.586	**0.636**	0.555	42.0%	55.4%	64.6%
	FuzzyBIO	**0.596**	0.612	**0.584**	**43.1**%	**58.9**%	**65.2**%
PsyTAR	BIOHD	**0.771**	**0.751**	**0.797**	**87.4**%	83.8%	79.5%
	FuzzyBIO	0.762	0.747	0.780	84.9%	**84.4**%	**87.5**%
CLEF	BIOHD	**0.312**	**0.286**	0.345	35.6%	52.8%	69.1%
	FuzzyBIO	0.309	0.276	**0.352**	**36.2**%	**55.2**%	**76.6**%

Table 7.3: Intrinsic evaluation of NER. Results are the average of a 10 fold CV.

7.5.1. INTRINSIC EVALUATION

As can be seen in Table 7.3, the FuzzyBIO scheme improves recall for two of three data sets, namely for CADEC (+0.29) and CLEF (+0.07), at a cost to precision (-0.24 and -0.1). For these data sets, using FuzzyBIO also leads to a higher percentage of correctly identified entities for both continuous (+1.1 and +0.6) and discontinuous entities (+3.5 and +3.6 for disjoint and +0.6 and +0.7 for composite entities). For the remaining data set (PsyTAR), the overall NER performance is negatively affected by using FuzzyBIO (-0.09) and continuous entities are missed more often (-2.5). Nonetheless, also for this data set discontinuous entities are extracted correctly more often (+0.8 and +8.0) when using FuzzyBIO instead of the BIOHD scheme.

7.5.2. EXTRINSIC EVALUATION

As can be seen in Table 7.4, using the FuzzyBIO scheme improves end-to-end performance for continuous and composite entities in two of three data sets, namely the CADEC (+0.4 and +1.3) and CLEF data (+0.4 and +5.7). In contrast, the end-to-end performance for disjoint entities is decreased (-15.5 and -21.0) for these data sets despite initial gains during NER. In the remaining data set (PsyTAR), the percentage of correctly identified entities after normalization is lower for all entity types when using the FuzzyBIO instead of the BIOHD scheme.

		Entity level recall		
Data	Scheme	Continuous	Disjoint	Composite
CADEC	BIOHD	23.9%	**35.9**%	21.2%
	FuzzyBIO	**24.3**%	20.4%	**22.5**%
PsyTAR	BIOHD	**43.6**%	**26.1**%	**10.6**%
	FuzzyBIO	42.8%	25.0%	7.5%
CLEF	BIOHD	21.7%	**25.8**%	26.5%
	FuzzyBIO	**22.1**%	4.8%	**32.2**%

Table 7.4: Extrinsic evaluation of ADE extraction. Results are the average of a 10 fold CV.

7.6. DISCUSSION

In answer to RQ1, we find that the FuzzyBIO scheme benefits overall recall during NER for two of the three data sets. For these data sets, it also leads to a higher percentage of correctly identified entities, both continuous and discontinuous. For the third data set (PsyTAR), more discontinuous entities are extracted correctly when using FuzzyBIO instead of the BIOHD scheme. However, more continuous entities are missed. In answer to RQ2, we find that for the same two data sets (CADEC and CLEF) the end-to-end ADE extraction is improved for continuous and composite entities. However, for the remaining data set (PsyTAR) the end-to-end performance is lower for all entity types.

We believe that the difference between PsyTAR and the other data sets may be related to either the low number of discontinuous entities or the low number of composite entities in the PsyTAR data, which may have hindered the training of an NER model for these entity types. An alternative explanation is that FuzzyBIO is less beneficial for easier NER tasks: The initial NER performance with BIOHD is far higher for PsyTAR (F_1 of 0.771) than for the other data sets (F_1 of 0.586 and 0.312). The difference between PsyTAR and the other data sets is unlikely to be related to the relative percentage of discontinuous entities, as this is similar to that of the CLEF data (5.8 vs 6.2%), or the nature of the data, as CADEC contains forum posts from the same website.

Another result that stands out is the lower end-to-end performance for disjoint entities when using the FuzzyBIO scheme despite initial gains in the extraction of disjoint entities for all data sets. We suspect that the normalization of these entities is made more challenging by the words in between the disjoint parts of the entity that are now included in the extracted entity. Therefore, in future work, we plan to investigate post-processing

steps such as the removal of stop words, which may improve the normalization of the more noisy disjoint entities represented with FuzzyBIO. As our representation is applicable for representing any type of discontinuous entity, future work may also include testing FuzzyBIO in other domains.

Although the improvement in NER is comparable for disjoint entities in medical records and user-generated content, the negative impact on normalization of disjoint entities is far stronger for the medical records. One might expect the normalization to decrease more strongly if more non-entity words (i.e., more noise) were included, but the median amount of non-entity words included in the disjoint entities is equal for all data sets (on average 1 word is added). Manual analysis also does not reveal a difference between the *type* of non-entity word included; They appear to mostly be stopwords. Thus, the most likely explanation for this difference is that the training examples for normalization from the user-generated data are already more noisy than their counterparts from the medical records. Consequently, the normalization algorithm for user-generated data might be better at dealing with noise. Future work could investigate whether training with the noisy examples instead of the original entities would be beneficial.

FuzzyBIO appears to be more beneficial for end-to-end extraction of composite entities in the medical records (+5.7) than in the user-generated data (+1.3 and -3.1). However, the number of non-entity words that is included is not lower for the medical records (median of 2 words added) compared to the user-generated data (median of 1 for CADEC and 3 for PsyTAR). Thus, this difference does not appear to be due to an increase in the fuzziness of the entities.

We also find some support for our hypothesis that the BIOHD representation makes the NER task more difficult for BERT models to learn than the FuzzyBIO representation. Overall the BERT models have difficulty learning the additional tag types; The precision for H- and D-tags is consistently lower than the precision for B-tags. In fact, on the PsyTAR data which contains few overlapping entities (1.7%), the H-tag was never predicted. It seems that FuzzyBIO makes the task easier in two ways, namely by standardizing entities into continuous sequences that always start with a B-tag and by excluding rare tags such as the H-tag. Standardizing entities makes it easier for the model to learn the underlying rules and excluding rare tags removes a goal for which there are only few examples available.

7.7. CONCLUSION

We expect FuzzyBIO to be most beneficial for NER for difficult tasks with a fair amount of discontinuous entities. However, since the conversion from BIOHD to FuzzyBIO is straightforward and deterministic, we recommend experimentally comparing which of the two is more effective for any data set that includes discontinuous entities.

PART III:

EXTRACTING COPING STRATEGIES FOR ADVERSE DRUG EFFECTS

When the dog bites, when the bee stings
When I'm feeling sad
I simply remember my favorite things
And then I don't feel so bad

Rodgers, Hammerstein & Nevin (1981), *The Sound of Music*

8

THE DISCOVERY OF RECOMMENDED COPING MECHANISMS

Edited from: **Anne Dirkson**, Suzan Verberne, Gerard van Oortmerssen, Hans Gelderblom and Wessel Kraaij. How do others cope? Extracting coping strategies for adverse drug events from social media. Submitted.

Patients advise their peers on how to cope with their illness in daily life on online support groups. To date, no efforts have been made to automatically extract recommended coping strategies from online patient discussion groups. We introduce this new task, which poses a number of challenges including complex, long entities, a large long-tailed label space, and cross-document relations. We present the first initial ontology for coping strategies as a starting point for future research on coping strategies, and the first pipeline for extracting coping strategies for side effects. We also compared two possible computational solutions for this novel and highly challenging task; multi-label classification and named entity recognition (NER) with entity linking (EL). We found that coping strategy extraction is difficult and both methods reach limited quality on held out test sets; multi-label classification outperforms NER+EL ($F_1 = 0.220$ vs $F_1 = 0.155$). An inspection of the multi-label classification output revealed that for some of the incorrect predictions, the reference label is close to the predicted label in the ontology (e.g. the predicted label 'juice' instead of the more specific reference label 'grapefruit juice'). Performance increased to $F_1 = 0.498$ when we evaluated at a coarser level of the ontology. We conclude that our pipeline can be used in a semi-automatic setting, in interaction with domain experts to discover coping strategies for side effects from a patient forum. For example, we found that patients recommend ginger tea for nausea and magnesium and potassium supplements for cramps. This can be used as input for patient surveys or clinical studies.

Patients rely heavily on the experiences of other patients for advice on how to cope with their illness in daily life [277]. Specifically, it has been found that patients use online disease-specific forums to gain information from peers [45, 129, 157]. While professionals often approach patients from a primarily medical point of view, patients need to weigh different life values of which 'taking good care of one's body' is but one [49, 56, 238]. Fellow patients are therefore often able to provide more pragmatic and holistic advice to their peers [38].

Adverse Drug Events (ADEs), harmful reactions that result from the intake of medication, are one aspect of their illness that patients need to cope with. ADEs can severely impact the quality of life of patients as well as form a barrier to medication adherence [167]. Although pharmacological management of side effects is sometimes possible, qualitative work indicates that lifestyle and diet can also impact the extent of ADEs, especially for chronic disorders [5].

Previously, qualitative studies have investigated how patients cope with side effects using questionnaires or structured interviews. The most used measurement instrument is the Side Effects Coping Questionnaire (SECope) [148] and the revised version developed by Smedt et al. [279]. It has been employed for the general population [225], patients with HIV [148], and patients with chronic heart failure [279]. The SECope measures general strategies for managing ADE, namely non-adherence, information seeking, social support seeking, and taking medication. The revised version contains two additional strategies: accepting the ADE and requesting other medication from the treating physician.

To date, the only large-scale study into which specific coping strategies patients employ for side effects is an internet survey [156] amongst patients receiving antidepressants. They found that patients employ a variety of methods including changes in lifestyle, diet, and social situations, next to pharmacological management.

Automatic extraction of coping strategies from peer-to-peer resources where patients themselves obtain advice has not been explored. Harvesting coping strategies recommended by patients could provide researchers with new hypotheses and facilitate medical research into which strategies work and why. Some strategies may work to the detriment of medication efficacy. A classic example is the consumption of grapefruit juice which can influence drug metabolism [312]. Our goal is not a fully automated method but a method that produces output that can be assessed and later used by a domain expert.

We focus on coping strategies for adverse drug events specifically. For example[1], in the sentence 'Pickle juice reduces my cramps within just a few minutes' the ADE is cramps and the coping strategy is drinking pickle juice, and in the sentence 'If you feel nauseous, eat ginger' the ADE is nausea and the coping strategy is eating ginger.

The automatic extraction of coping strategies from online patient forums poses four major challenges:

Complex entities The narrative description of coping strategies (e.g. 'take 400mg with breakfast and 400mg with dinner and a big glass of water') results in complex and long entities, which are often not proper nouns. Classic methods for entity extraction are generally not equipped to deal with.

[1] These examples are artificial variants of real sentences in the data to protect patient privacy

No ontology There is at present no ontology to normalize or link the coping strategies to, while aggregation and normalization of coping strategies is vital to be able to provide insight into overall prevalence.

Large and long-tailed label space The large variety of possible coping strategies means that extraction or classification methods will need to be able to deal with a large number of zero-shot cases (i.e. target classes for which there are no examples in the training data) as it is not feasible to collect sufficient data for all target classes.

Cross-document relations Coping strategies are only relevant in relation to a specific ADE and in online discussions these relations may span multiple messages.

An additional complicating factor is that ADE extraction is not trivial. For instance, it is challenging for models to distinguish ADEs from symptoms of the disorder or symptoms resulting from withdrawal (of a medication). The ADE extraction that we employ[2] attains an end-to-end token-level performance of F_1 0.626 and an entity-level performance of 0.716 (Chapter 9).

We address the following research questions:

RQ1 To what extent can coping strategies for ADE be extracted automatically from online patent experiences?

RQ2 How do two approaches to information extraction, namely named entity recognition (NER) with subsequent entity linking and multi-label classification compare on this task end-to-end?

We evaluate our methods on data related to Gastrointestinal Stromal Tumors (GIST), a rare cancer in the digestive system. The Facebook page of the worldwide patient organization GIST Support International (GSI)[3] is the largest online patient community for GIST patients. On the Facebook page, patients share their experiences in discussion threads. The data we work with consists of 124,103 posts in 14,631 threads.

Our main contributions to the medical informatics field are thereby: (1) the novel task of coping strategy extraction, (2) an exploration of extraction and classification methods for its end-to-end resolution and (3) the first ontology for coping strategies. Our code and ontology are publicly available.[4] Unfortunately, our annotated data cannot be shared due to privacy restrictions.

The remainder of the paper is organized as follows: In Section 8.1, we discuss related methodological work. In Section 8.2 and 8.3, we discuss the data sets we use, followed by a detailed description of our methodology. In Section 4.5, we present our results, which are discussed further in Section 8.5.

8.1. RELATED WORK

For the extraction of medical concepts, two broad approaches can be identified. The first approach is Named Entity Recognition (NER) to extract the relevant phrases or

[2] ADE extraction consists of an endr-BERT model and subsequent BioSyn entity linking to SNOMED-CT
[3] `https://www.facebook.com/groups/gistsupport/`
[4] `https://github.com/AnneDirkson/CopingStratExtract`

entities with subsequent entity linking to determine which concept from an ontology is mentioned in the phrase. This approach is widely used for the related task of extraction of ADE from social media messages [193, 266, 335]. State-of-the-art methods for ADE extraction generally rely on domain-specific BERT models [88, 193, 194]. Entity linking of ADE entities is cast as a classification task with all concepts in a medical ontology (e.g., MedDRA or SNOMED-CT) as possible target labels. Because of the large label space, which leads to sparseness in the training data for smaller categories, these methods are designed to be able to deal with zero-shot cases. Similar to coping strategies, the label space for these tasks is both long-tailed and large with over 20.000 labels in MedDRA [194]. Present competitive methods such as BioSyn [291] are often ranking-based and use dense BERT embeddings. The biggest bottleneck at present for end-to-end systems is the extraction step which leads to severe error propagation [194, 335]. Mentions of coping strategies are even longer and more diverse than ADE entities, which makes the problem challenging to be approached as an NER task. The challenge of NER for longer and fuzzy entities has been acknowledged in some recent work, for biomedical concepts [72], human senses [214], motives [332], and emotion causes [182]. We will investigate how well NER with entity linking works for coping strategies using BERT models for NER and BioSyn for entity linking.

The second approach is multi-label classification, which is employed more commonly for tasks such as automatic ICD[5] code assignment [153]. This task is comparable to coping strategy extraction; The label space is also very large and long-tailed (the ICD-9 contains over 15.000 codes and its successor the ICD-10 over 140.000 codes) and multiple labels can be assigned to a single document, i.e., the labels are not mutually exclusive. Although automatic ICD code classification has been explored since the 90s [286], methods have evaluated on the full ICD as opposed to a strict subset of ICD codes only in recent years [212]. While these methods can potentially predict zero-shot cases, they still perform very poorly.

Only a few methods have actually been designed to deal with zero-shot cases to some extent [250, 284]. Rios and Kavuluru [250] extended the CNN-based CAML-DR method of Mullenbach et al. [212] with a graph CNN that makes use of the structure of the label space. Chalkidis et al. [59] find that their model ZAGCNN outperformed transfer learning methods (i.e. BERT and RoBERTA) on few-shot cases and performed comparably on frequent labels. Their results also indicate that exploiting information from label descriptors appears more important than exploiting the label hierarchy for few-shot and zero-shot learning. Song et al. [284] further improve upon the work by Rios and Kavuluru [250] by replacing the CNN with an RNN component. They also propose a latent feature generation framework based on generative adversarial networks [117] to improve the prediction of unseen codes without compromising the prediction of seen codes. Features are generated by exploiting the label structure and label descriptions. As our data does not include label descriptions, these methods are not transferable to the task at hand.

Instead, we opted for a multi-label classification method that does not require label descriptions. We employed a ranking-based (or information retrieval) approach in which labeled data is only used to determine the optimal similarity threshold (i.e., the sentence is

[5] ICD or International Classification of Disease is a terminology for classifying diseases developed by the WHO

Named Entity Recognition (NER)	
CS	781 (2,729 tokens)
– median length CS	3 tokens (mean = 3.55, max = 29)
CS-NEG*	43 (197 tokens)
ADE	2,001 (5,983 tokens)
Negative (O-tag) tokens (included**)	187,355 (95,830)
Posts (included**)	3,715 (1,995)
– median # CS per post	0 (mean = 0.42)
Posts with CS	481
– that also contain an ADE	284 (59%)
Discussions (with CS)	527 (170)
Entity linking (EL)	
CS	824
– with >1 label	59
– with higher order label†	42
# unique concepts	284
% of CSAO in labeled data	0.6%
Posts	481
Multi-label	
CS	824
Posts with CS	481
– median # of labels	1 (max=9)
Negative cases	1514

Table 8.1: Descriptive statistics for Coping Strategy extraction data sets. The multi-label data is converted from the NER and EL data. *Converted to CS for NER **Only a subset of negative examples was included during training †If the concept does not exist in the ontology but the higher order category does

labeled with all labels scoring above this similarity score). Specifically, we used sentence-BERT models to measure the similarity between sentences and target labels. Sentence-BERT models are a class of models introduced by Reimers and Gurevych [247] that are better equipped to handle sentence-level tasks such as multi-label classification. These models employ a pretraining scheme based on Siamese networks.

8.2. DATA

We first detail the data collection and annotation for this novel task in Section 8.2.1 and 8.2.2. The ontology creation is then described in Section 8.2.3. In Section 8.2.4, we describe how we add negative examples to the annotated data.

8.2.1. DATA COLLECTION

In agreement with the GIST Support International Organization, we collected data from their Facebook group. More specifically, we accessed the Facebook official API[6] through

[6] `https://developers.facebook.com/docs/graph-api/`

Relation extraction (RE)	
# of ADE-CS relations	580
– within the same post	397 (68.4%)
– median # of possible ADE per CS	2
– median # co-referents of ADE for which CS is advised	7
# negative cases	1350
median # of annotated posts per CS	6

Table 8.2: Descriptive statistics for the Relation Extraction data between coping strategies (CS) and Adverse Drug Events (ADEs).

a Python script. We got access to the contents of the Facebook group through the account of the group admin. We then collected all posts and comments from the start of the forum. The data ranges from 24 Oct 2009 until 1 Nov 2020 and includes 124,103 posts in 14,631 threads. Our study design was in line with the privacy guidelines of Leiden University and approved by the University privacy officer. The Facebook API did not provide (pseudonymized) usernames in order to protect user privacy. Thus, we were unable to link different posts from the same user within the forum. The collected messages were stored securely, and access was restricted to the involved researchers and annotators. For the labeling of data, we did not use commercial tools but set up private servers that were only accessible to the annotators. In accordance with the GDPR (Article 9.2), we did not obtain consent from each user as the GDPR allows for the use of data from publicly accessible forums with justified cause without individual consent. We are unable to share the data according to the GDPR, because access to the forum has become restricted to members since our data collection (i.e., it is no longer publicly accessible).

8.2.2. DATA ANNOTATION

Named Entity Recognition For annotation, we selected 527 discussions (4,195 posts) based on their likelihood to contain an ADE. We automatically selected the threads that contained at least one drug name according to a match with RxNorm [314]. From these, we selected the threads with the highest percentage of posts in which experiences are shared until our data set included over 4,000 posts. Sharing that someone experienced an ADE falls under this category. In order to estimate which percentage of posts in a thread included patient experiences, we used a previously developed model (Chapter 3).

The data was first annotated by three GIST patients and the first author for the presence of ADEs and coping strategies (CS) for ADE using an annotation guideline.[7] Annotators could indicate with the CS-NEG tag (as opposed to the CS tag) that a coping strategy for an ADE was negative i.e. it entails *not* doing something (e.g. 'avoid salt'). The pair-wise inter-annotator agreement was substantial for ADE (mean κ =0.71) and moderate for CS (mean κ =0.54). The somewhat lower agreement for CS compared to ADE indicates that the CS annotation task is more difficult than the ADE annotation task, but with moderate agreement we still consider the data of sufficient quality to train and

[7] All annotation guidelines are provided at: `https://github.com/AnneDirkson/CopingStratExtract`

Tokens	Pickle	juice	reduces	my	muscle	cramps
NER tags	B-CS	I-CS	O	O	B-ADE	I-ADE
Entity linking	CS04916	CS04916	-	-	-	-

Table 8.3: Example annotation for NER and entity linking

Text	ENTITY_2 (CS)	ENTITY_1 (ADE)	Label*
ENTITY_2 reduces my ENTITY_1 but not my nausea	Pickle juice	muscle cramps	1
ENTITY_2 reduces my muscle cramps but not my ENTITY_1.	Pickle juice	nausea	0

Table 8.4: Example annotation for CS-ADE relation extraction. *1 indicates an CS-ADE relation

evaluate our models on. Data labels were converted to the FuzzyBIO annotation scheme proposed in Chapter 7. We used an online tool Doccano[8] implemented on our own private server for annotation. See Named Entity Recognition in Table 2.3 for details on the annotated data and Table 8.3 for an artificial example of what the annotated data looks like. A more extensive real annotated data fragment is provided in Appendix B (Table B.1).

Normalization The coping strategies were then annotated with concepts from our developed ontology (see Section 8.2.3) by three master students. We switched from Doccano to the annotation tool Inception[9], because Doccano is unable to annotate extracted text spans with concepts from a custom ontology. To switch from Doccano to Inception, we uploaded the earlier NER annotations (in CoNLL-2003 format) from Doccano into Inception. A pilot annotation was used to improve the annotation guideline. All three annotators annotated every post. The inter-annotator agreement was substantial (mean $\kappa = 0.706$) on a token level and moderate (mean $\kappa = 0.475$) on a document (i.e. post) level. Their annotations were curated by the first author. Agreement between at least two of the three annotators was sufficient. The remaining conflicting cases were discussed and resolved. New concepts were added to the ontology where necessary. In 42 cases, the concept was labeled with a higher order concept when the exact concept was not available, e.g., badminton would be labeled with Sport instead of Badminton. If the annotated coping strategy consisted of two strategies (e.g. 'Eat melon and kiwi' or 'Take painkillers and eat well'), the annotators needed to split the strategy to permit labeling. If it was unclear to the annotators what the patient meant, the coping strategy remained unlabeled. This only occurred in 4 cases. See Entity linking in Table 2.3 for details on the annotated data and Table 8.3 for an artificial example. A more extensive real annotated data fragment is provided in Appendix B (Table B.1).

[8] https://github.com/doccano/doccano
[9] https://inception-project.github.io/

ADE-CS relations The annotated coping strategies were coping strategies for a certain ADE. For each CS, three annotators (three different master students) annotated for which ADE the patient recommends the CS. They used the annotation tool Doccano. Annotators were provided with the six messages in the discussion before the post containing the CS. All co-referents of the ADE for which the CS is recommended were annotated. A pilot annotation was used to improve the annotation guideline. Based on an overlapping set of 100 posts, the inter-annotator agreement was measured as the average pair-wise mutual F_1 score of the annotators was 0.757.[10] For every pair-wise calculation, only instances in which at least one of the two annotators found a relation were included. See Table 8.2 for details on the data set and Table 8.4 for an artificial example of what the annotated data looks like. A more extensive real annotated data fragment is provided in Appendix B (Table B.2).

8.2.3. COPING STRATEGY ONTOLOGY

The starting point for our ontology was the experiences of GIST patients we collaborated with and our own experiences with the GIST patient forum. We used these to devise categories of coping strategies patients employ, e.g., edible substances and physical exercise. For each category, we manually selected an appropriate category in one of our source ontologies (e.g., Edible substance (SNOMED-CT 762766007)). We sourced from existing ontologies to allow for interoperability with other ontologies. We chose SNOMED-CT, NCIT and RxNORM as our source ontologies in line with the OHDSI project [222]. We added the PACO Activity Ontology [142] to better represent daily activities and exercise. From the RxNORM ontology we included all Ingredients that are also included in the OMOP vocabulary of the OHDSI project [222]. We used the five hierarchical levels of the ATC (Anatomical Therapeutic Chemical) Classification of the WHO[11] to categorise the RxNORM concepts. The ATC divides medication based on the organ or system on which they act. For normalization, we merged relevant subcategories from different ATC categories into general antibiotics, antiseptics, and antivirals labels, i.e., antiseptics acting on different organs are now grouped.

During annotation, we identified gaps in our ontology. We expanded the ontology with additional categories (e.g., the category 'position of body') and concepts (e.g., 'shampoo' in the existing category 'personal care product' under 'physical object'). These concepts were sourced from the source ontologies if possible. If no appropriate concept was available, we added a concept of our own (e.g. 'split dosage' in the category 'methods of consumption drug' in Table 8.5).

The final ontology contains 48.764 concepts, of which 70.2% from RxNORM, 13.4% from ATC, 9.7% from SNOMED-CT, 6.3% from NCIT, 0.3% from PACO and only 0.1% (64 concepts) were our own additions. The ontology was created using the Python package owlready2. See Table 8.5 for examples and descriptions of the most prominent categories of the Coping Strategy for ADE Ontology (CSAO). We also provide snapshots of the ontology and its hierarchical levels in Table 8.6 and 8.7. The ontology is publicly

[10] The pair-wise F_1 score is preferable to Cohen's kappa for calculating IAA in Named Entity Recognition, as Cohen's kappa needs the number of negative cases which is unknown for NER [41, 138]

[11] `https://www.whocc.no/atc/structure_and_principles/`

Category	Description	Example	# concepts
Adaptation	Includes mental constructs, e.g., attitude and adapting to the circumstances	Positive attitude (SNOMED 225463003)	6
Eating and drinking	Food & drinks, but also frequency and size of meals	Blueberries (SNOMED 227416001)	3,145
Intervention or Procedure	Therapeutic and surgical procedures, alternative therapies and counseling	Thoracentesis (NCIT C15392) Acupuncture Therapy (NCIT C15176)	3,052
Lifestyle	Includes activity, resting, social activities, general dietary recommendations, and clothing strategies	Swimming (PACO 10081)	202
Medication and Supplements	RxNorm medication ingredients categorized by ATC categories	Ondansetron (RxNORM 26225)	40,770
Methods of consumption drug	How and when the medication is consumed	Split dosage (new) After breakfast (SNOMED 7221000175107)	61
Physical object	Various aids, clothing items, and personal care products	Toothpaste (SNOMED 48741003) Single vision glasses (SNOMED 397287009)	1,513
Position of body	Different positions of the body	Sitting (new)	7

Table 8.5: Overview of the major categories in the Coping Strategy for ADE Ontology

available.[12] We consider our ontology – that was initially tailored to GIST – a starting point for more general research into strategies that patients use to cope with side effects.

8.2.4. ADDING NEGATIVE EXAMPLES

Previous work has shown that it is beneficial to include negative examples (i.e., sentences that do not include the item of interest) in the training set for information extraction from medical social media [194]. We found that 481 of the 4,195 posts that were subjected to NER annotation contained coping strategies, thus leaving 3,741 possible negative examples (i.e., sentences that do not contain coping strategies). To reduce the data imbalance, we selected a subset of these negative examples. Specifically, we opted to present the model with difficult negative examples by using forum messages where coping strategies are likely to occur but do not. We accomplished this by selecting the posts that contain an ADE (according to the NER annotation) and the four subsequent messages in

[12] https://github.com/AnneDirkson/CopingStratExtract/blob/main/CSAO.rdf

Eating and drinking	Edible substance	Meat			
		Seafood			
		Dairy food			
		Starchy food	Rice	Brown rice	
				White rice	
				...	
			Bread	Rye bread	
				Tortilla	
				Pita bread	
					White pita bread
					Wholemeal pita bread
				...	
			...		
		...			

Table 8.6: A snapshot of the Edible substance category under Eating and drinking. ... indicate that there are more sub-categories than listed here.

Physical object	Personal care product	Aftershave		
		Baby powder		
		Hair dye		
		Lotion		
		Lip balm		
		Deodorant		
		Mouthwash	Giving analgesic mouthwash	
			Giving antiseptic mouthwash	
			Giving warm saline mouthwash	
		...		

Table 8.7: A snapshot of the Personal care product category under the Physical object section. ... indicate that there are more sub-categories than listed.

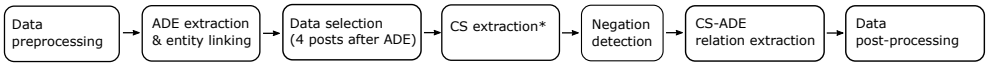


Figure 8.1: Pipeline. ADE: Adverse Drug Effect, CS: Coping Strategy. *Multi-label classification or NER with subsequent entity linking

the discussion. This provided us with 1514 posts (76%) that do not contain CS (see Table 2.3). We included these negative examples in the training set for both NER and multi-label classification.

8.3. METHODS

In Sections 8.3.1 to 8.3.6, we describe the modules of our extraction pipeline for coping strategies shown in Figure 8.1. Although additional components (such as relation extraction, and negation detection) are part of the complete pipeline of extracting coping strategies from online discussions, we define the 'end-to-end' resolution or extraction of coping strategies in this chapter as determining which coping strategies are mentioned in the text.

8.3.1. DATA PREPROCESSING

We preprocessed the data with the pipeline described in Chapter 2. We excluded drug names in the FDA database of drugs[13] from spelling correction to prevent uncommon drug names from being replaced by more common, similar drug names. Removing empty messages and messages in a language other than English left 125,161 messages. Spelling correction corrected 24,834 mistakes. We also normalized drug names to their generic forms using the FDA database.

8.3.2. ADE EXTRACTION AND DATA SELECTION

The extraction of ADE has been described elsewhere (Chapter 9). Adverse drug events were normalized to SNOMED-CT concepts in line with the OHDSI project [222]. Although some previous work has elected to use MedDRA instead of SNOMED, this work focuses predominantly on Twitter data. Annotated datasets for ADE normalization of data that is more comparable to patient forum posts, i.e., Askapatient [151, 353] and Reddit data [20], make use of SNOMED-CT.

For our pipeline, we selected each post that contains an ADE and the subsequent four posts for CS extraction ('Data Selection' in Figure 8.1). Pre-selection of posts that are likely to contain the concept of interest has been shown to aid extraction in social media data with a large signal-to-noise ratio [194]. The window of four subsequent posts was chosen to be relatively wide so as to not miss any coping strategies. The selected posts were not automatically linked to that particular ADE, but purely determined the processing scope for subsequent steps including relation extraction. If an ADE is present in the window of another ADE (e.g., in the second post), its subsequent four posts are also included for CS extraction. The data is deduplicated so any post only occurs once irrespective of the number of ADE within range.

[13] https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files

8.3.3. COPING STRATEGY EXTRACTION

We compared multi-label classification and NER+EL for the end-to-end extraction of coping strategies. These extraction methods are comparable because we know for each sentence which CS concepts it contains.

MULTI-LABEL CLASSIFICATION (MLC)

We used sentence-BERT models [247] for multi-label classification. Sentence-BERT models employ a pretraining method using Siamese networks that results in models more suitable for sentence-level tasks such as measuring semantic similarity. As social media text does not consistently conform to grammatically rules, we choose a pragmatic approach to sentence splitting based on punctuation[14]. We used three different sentence-BERT models [247]: (1) the recommended model for semantic similarity (all-MiniLM-L6-v2) which has been fine-tuned on over 1 billion sentence pairs, (2) a specific natural language inference (NLI[15] model trained on NLI data only and (3) the recommended model for semantic search (msmarco-distilbert-dot-v5) trained on the MSMARCO data set [17]. The MS MARCO data set is a large scale information retrieval corpus based on real user search queries in the Bing search engine and ranked passages for these queries. For this model, the training data consisted of a set of over 500k examples. The full MS MARCO corpus contains over 8 Million examples. The latter model was tuned for dot-product similarity. We also tried the model variant tuned for cosine similarity, but this performed similarly. For the NLI and semantic similarity models, we used the sentences as queries and the labels as retrieval items, whereas for the semantic search model all possible concepts from the ontology (i.e., all possible labels) were used as queries and the sentences as retrieval items because these models are tuned for short queries and longer retrieval documents.

These models were unsupervised and thus training data is not necessary for retrieval. As the models output a similarity (between 0 and 1), we used the training data to determine the optimal threshold (0.1 to 1, steps of 0.1) to select the set of assigned labels. We employed five-fold cross validation in which data are stratified per post.

8

NER WITH ENTITY LINKING

For Named Entity Recognition (NER), we used BERT models, specifically we compared the original BERT model [84] to one trained on English medical social media data (EnDRBERT [303]) and one trained on biomedical texts (PubmedBERT [119]). We used the same five-fold cross-validation as for multi-label classification (60% train, 20% validation, and 20% test per fold). The learning rate (0.01) was optimized on the validation data. Models were trained for 3 or 4 epochs based on validation data. To align experiments with multi-label classification, we trained NER on individual sentences.

We experimented with including ADE as a second entity type during the training of NER models. We expected that identifying ADE may be an easier task than identifying CS and coping strategies for ADE should occur in their vicinity.

We analyzed different possible entity linking methods for the extracted CS phrases. We used the state-of-the-art method for ADE entity linking, BioSyn [291]. We explored

[14] See `https://github.com/AnneDirkson/CopingStratExtract`

[15] Natural language inference is the task of predicting whether one sentence infers the other. An NLI model predicts for a premise whether the hypothesis is true, false or unrelated to the premise.

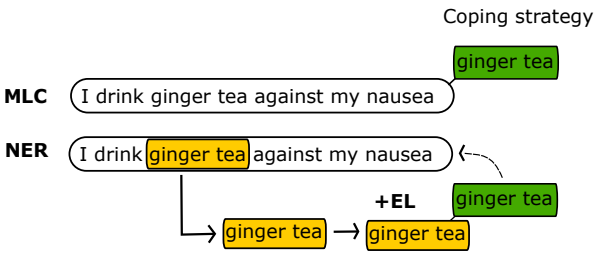


Figure 8.2: Illustration of multi-label classification (MLC) and Named Entity Recognition with Entity Linking (NER + EL). Labels resulting from EL are linked to the original sentence as shown by the dotted line to generate sentence-level results for NER+EL. The sentence-level labels from MLC are then compared with these sentence-level labels from NER+EL.

both BioBERT [174] and SapBERT [187] as base embeddings for this method. SapBERT is a recent pretraining scheme that leverages the UMLS (a biomedical ontology with 4M+ concepts). Liu et al. [187] show that SapBERT pretraining can improve entity linking performance of various BERT-based models with especially large gains for social media data. It also attained a better performance with BioSyn than BioBERT [187]. BioSyn provides a ranking of possible labels present in the phrase. Since CS phrases can have multiple labels, we applied a simple heuristic to allow for multiple labels: The second label is also added if its similarity is closer to the first label than the third label. We attempted to determine a similarity threshold, as we did for the classification approach, but because the similarity metric used in BioSyn is not normalized, this worked poorly.

We compared BioSyn with the best unsupervised multi-labeling classification approach for entity linking. Labels resulting from entity linking were linked to the original sentence to generate sentence-level results for NER+EL. The sentence-level labels from MLC were then compared with these sentence-level labels from NER+EL. Figure 8.2 visualizes this comparison. For these experiments, the same five-fold cross-validation was used.

8.3.4. NEGATION DETECTION

Coping strategies can also entail *not* doing something instead of doing something (e.g. 'I avoid salt'). We found 43 examples during annotation (i.e. labeled CS-NEG) (see Table 8.1). We used a simple heuristic negation method, relying on the Spacy [136] implementation of the Negex algorithm [60]. We used the basic English term set supplemented with additional sixteen preceding and three following heuristics for identifying negation that were manually identified in the data. If one of the heuristics is present, we considered any strategies within the five preceding or subsequent tokens (excluding punctuation) depending on the type of heuristic to be negated. We also determined the dependency relations of strategies. Strategies are negated if they have one of the following dependency relations: (1) negation, (2) no as a determiner or (3) non as a adjectival modifier. We evaluated our heuristic method using entities in the NER that should (CS-NEG) (43 entities) and should not be negated (CS) (781 entities). It attained an F_1 score of 0.810 with a recall of 0.829 and a precision of 0.790.

8.3.5. RELATION EXTRACTION

It is important to determine *which* ADE the coping strategy relates to. We applied a rule-based approach for relation extraction: If there is an ADE mentioned earlier in the message, select the closest one. Otherwise, select the ADE mentioned afterwards within the message. In the annotated data, in 134 of the 365 posts (36.7%) where the ADE is mentioned within the post, another ADE is also mentioned within the post. If there is no ADE in the message itself, select the ADE mentioned closest to the strategy earlier in the discussion within at most preceding four posts.

We evaluated our approach on the annotated data (see Table 8.2). We excluded the 232 cases (29.2%) for which the annotators could not determine which ADE the strategy related to. Manual analysis revealed these were the results of errors in the ADE annotation. Within posts, our rule-based classifier attained an accuracy of 88.4%. For all posts including those with cross-post relations, our classifier attained an accuracy of 84.7%.

8.3.6. DATA POST-PROCESSING

Further data post-processing consisted of three steps. First, we removed strategies that are not connected to any ADE (25.1%) as these are likely to be false positives. We checked a random selection of 50 cases and found that 42 of the 50 were false positives, whereas for the other eight the ADE was missed or not mentioned (e.g., for antidepressants the ADE is implied). Second, we removed labels for which the most important token is already connected to another label with a higher semantic similarity, i.e., a sentence will often be linked to >1 highly similar labels (e.g., 'ground ginger' and 'root ginger' for the token 'ginger' and 'cannabis' and 'cannabis oil' for the token 'marijuana'). We also removed labels for which the most important token is the location of an ADE. The third step was combining multi-label instances; We considered two labels as part of one multi-label instance if the locations of the key tokens are adjacent, they are connected to the same ADE and they have the same negation value. An example is 'high fiber' and 'fruit' for the term 'high fiber fruits'.

8.4. RESULTS

First, we describe our ground truth data in Section 8.4.1. Hereafter, we present the best NER method for extracting spans with coping strategies in Section 8.4.2. We compare the best NER method combined with entity linking with multi-label classification for end-to-end extraction in Section 8.4.3. Section 8.4.4 reports the coping strategies for ADE found in a case study on a patient forum for GIST patients.

8.4.1. DATA DESCRIPTION

As this task is novel, we will describe our ground truth data to explore the challenges this task presents. Table 8.1 describes the annotated data for NER and entity linking. The annotated data contains a total of 824 coping strategies, of which 5.2% were negative strategies meaning they entail not doing something (e.g., not drinking milk). Thus, negation detection will be necessary to differentiate positive from negative strategies. The median length of the annotated coping strategies was relatively short (3 words) but they could be very long (up to 29 words). In fact, 5.4% (52) of the coping strategies contained

	No ADE detection		
	Micro F1	Micro R	Micro P
BERT	0.200 ± 0.157	0.155 ± 0.146	**0.671** ± 0.188
EndrBERT	0.089 ± 0.167	0.089 ± 0.172	0.433± 0.399
PubmedBERT	**0.204** ± 0.170	**0.165** ± 0.160	0.443 ± 0.246
	With ADE detection		
	Micro F1	Micro R	Micro P
BERT	**0.380** ± 0.048	**0.331** ± 0.111	0.522 ± 0.096
EndrBERT	0.251 ± 0.182	0.224 ± 0.205	0.503 ± 0.293
PubmedBERT	0.244 ± 0.119	0.161 ± 0.082	**0.713** ± 0.149

Table 8.8: Token-level evaluation results for NER of coping strategies with or without ADE extraction as a joint task. Our metrics are lenient and ignore prefixes, i.e, it is considered correct when the model predicts the correct entity type for a token irrespective of the B- or I-tag.

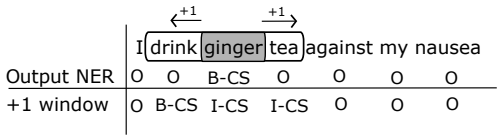


Figure 8.3: Illustration of adding a window of 1 token on both sides of CS mentions identified in NER.

more than 10 words. The data is sparse: Only 11% (481 of 4195) of the posts selected for annotation contained coping strategies. Note that the annotated 527 discussion threads were already preselected to be more likely to contain patient experiences prior to NER annotation so a full patient forum is likely to be more sparse still (See Section 8.2.2).

The ground truth for entity linking demonstrates that not all coping strategies can be captured with a single label from the ontology: 7.2% (59) of the annotated coping strategies were labeled with two labels (e.g. 'cinnamon' and 'chewing gum' for the entity 'cinnamon gum'). Moreover, our ground truth reflects the long-tailed label space. Our labeled 824 coping strategies only cover 284 unique concepts, which equals 0.6% of the ontology.

Table 8.2 describes the ground truth for Relation Extraction between ADEs and coping strategies. On average, there were two different ADEs that the strategy could be linked to within the span of six posts (the post itself and five prior). The ADE for which the CS was advised was mentioned often (an average of 7 times within the span of the post itself and five posts prior). In 31.6% of the cases, the relation was not within the same post but spanned across posts.

8.4.2. NAMED ENTITY RECOGNITION

The first approach to extraction that we evaluated consists of two steps, namely NER and entity linking. Table 8.8 shows the results for the first step of this approach: Named Entity Recognition of coping strategies. We compare models on their micro F_1 score, because it takes into account the label imbalance by aggregating the contributions of all

	Token level evaluation		
	Micro F1	Micro R	Micro P
No window	0.380 ± 0.048	0.331 ± 0.111	**0.522** ± 0.97
+1 on both sides	**0.394** ± 0.018	**0.453** ± 0.108	0.376 ± 0.068
	Entity level evaluation		
	Missed (%)	Correct (%)	Partially correct (%)
No window	**39.1** ± 1.2	27.8 ± 10.9	**33.1** ± 4.1
+1 on both sides	37.2 ± 11.1	**40.0** ± 11.2	22.7 ± 2.5

Table 8.9: Results for adding a window (+1 token) on either side of the extracted CS in NER.

classes and is standard in evaluating multi-label classification tasks. The best performing model was the standard BERT model that was trained to identify both ADE and CS entities ($F_1 = 0.380$). Adding ADE as an additional entity type[16] doubled its performance (+0.180) (See Table 8.8). Without the addition of ADE entities, PubmedBERT, which is trained on biomedical text, outperformed the other models ($F_1 = 0.204$).

Due to the complexity of the CS entities, we explored whether adding an additional token on either side of the identified strategies would benefit performance (See Figure 8.3). Table 8.9 reveals that adding a window of 1 token boosted token-level performance slightly ($F_1 = 0.394$) by increasing recall (+0.122) at a cost to precision (-0.146). On an entity level, the number of entities that are missed entirely was reduced (-1.9 % point), the number of entities that were partially correct was also reduced (-10.4% point), whereas the number of fully correct entities was increased (+12.2% point). We thus included a window of one token on each side for the extracted phrases (i.e., the input for entity linking).

8.4.3. END-TO-END EXTRACTION

Table 8.11 shows the results for end-to-end extraction of coping strategies for both approaches (NER with entity linking and MLC). Although the other multi-label classification models performed very poorly, the best performing method for end-to-end extraction was multi-label classification with the Semantic Similarity sentence-BERT model ($F_1 = 0.220$). With oracle NER (using the manually labeled NER data as input), entity linking using BioSyn based on SapBERT could outperform the classification approach ($F_1 = 0.241$). This higher performance was mainly driven by a higher precision (0.271). Yet, with the addition of NER as an intermediate step the performance dropped below that of multi-label classification. Moreover, multi-label classification outperformed even oracle NER in terms of recall (0.306 compared to 0.283). Macro F_1 scores are computed by averaging the F_1 scores for each class, thus treating all classes equally irrespective of their prevalence. Table 8.11 shows that the macro F_1 scores were far lower than the micro F_1 scores, indicating that across the board the models performed worse on less frequent coping strategies in the annotated data.

As the ontology is hierarchical, we also investigated how far off the predictions of

[16] On a token level, this means adding B-ADE and I-ADE tags

Prediction	Ground truth	Shared higher level
Lip balm	Lotion	Personal care product
Take whole dosage at once	Split dosage	Dosage
Rice	Bread	Starchy food
Therapeutic bed	Assistive bed	Sleeping aid

Table 8.10: Examples of cases where the predicted label and true label are not the same but do fall under the same direct hierarchical category (+1 level)

the best model were by investigating the performance at coarser levels of the concept hierarchy. The results are shown in Table 8.12. The performance was increased to $F_1 = 0.318$ when we considered if the target and predicted labels fell directly under the same direct category in the hierarchy (i.e. '+1 level (strict)') (see Table 8.10 for examples). Also the precision was increased (0.172 to 0.304). The macro F_1 showed a similar increase (from 0.105 to 0.320) which may indicate that it is mostly the infrequent coping strategies that are predicted incorrectly on the detailed level but correctly on the coarser level.

This is rather restrictive measure however, as the target and predicted labels need to fall directly under the same category. There may also be cases where the predicted label is equal to the category directly above the target label (e.g. the predicted label is chocolate and the target label is dark chocolate) or cases where the predicted label does fall under the category directly above the target but not directly (e.g. the predicted label is brown rice (+1 is rice) and the target label is bread (+1 is starchy food) in Table 8.6). When we consider whether the predicted label is equal to or falls under the category directly above the target ('+1 level (lenient)' in Table 8.12), the micro F_1 increases further to 0.498 and the precision increases drastically to 0.861.

When we considered if both target and predicted labels fell under the same overarching category in the hierarchy (i.e. 'Top Category'), we saw another increase in performance to $F_1 = 0.556$. An example would be if the model predicted another food that is not a starchy food such as dairy (See Table 8.6). Although this results in a very general categorization, it may nonetheless be useful to medical researchers, practitioners, and patients interested, for instance, in all edible substances or all lifestyle interventions that patients recommend for a certain ADE.

8.4.4. CASE STUDY ON GIST ADE COPING

For the case study on the entire GIST patient forum, we employed multi-label classification using semantic similarity sentence-BERT as it was the best performing method. Negation detection and relation extraction rely on knowing where in the sentence entities occur, but multi-label classification does not provide this information. Thus, we identified the approximate location of each CS (i.e., each assigned label) as the token in the sentence with the highest similarity to the assigned label.

This resulted in a total of 32,643 strategies of which 3% (1,017) are negated and 4% (1,375) are multi-label strategies. Figure 8.4a shows the ten most prevalent coping strategies mentioned on the forum. Manual analysis indicated that a large portion of these were false positives: They either refer to primary medication (e.g. imatinib); surgery

NER	Entity linking	Micro F1	Micro R	Micro P	Macro F1
None	SemSearch SBERT	0.001 ± 0.001	0.093 ± 0.180	0.001 ± 0.001	0.001 ± 0.001
	NLI SBERT	0.016 ± 0.013	0.018 ± 0.014	0.014 ± 0.012	0.008 ± 0.007
	SemSim SBERT	**0.220** ± 0.011	**0.306** ± 0.010	**0.172** ± 0.014	**0.105** ± 0.010
Oracle NER	+ SemSim SBERT	0.142 ± 0.043	**0.410** ± 0.089	0.086 ± 0.028	0.038 ± 0.015
	+ BioSyn (B)	0.236 ± 0.040	0.258 ± 0.039	**0.217** ± 0.040	**0.084** ± 0.018
	+ BioSyn (S)	**0.241** ± 0.029	0.283 ± 0.030	0.210 ± 0.028	0.083 ± 0.011
NER	+ SemSim SBERT	0.130 ± 0.021	**0.202** ± 0.039	0.097 ± 0.017	0.037 ± 0.008
	+ BioSyn (B)	**0.155** ± 0.017	0.168 ± 0.032	**0.151** ± 0.037	**0.049** ± 0.013
	+ BioSyn (S)	0.144 ± 0.026	0.162 ± 0.009	0.134 ± 0.039	**0.049** ± 0.016

Table 8.11: Results for end-to-end extraction of coping strategies. SBERT: Sentence-BERT, SemSim: Semantic Similarity, SemSearch: Semantic Search, BioSyn (B): BioSyn with BioBERT, BioSyn (S): BioSyn with SapBERT.

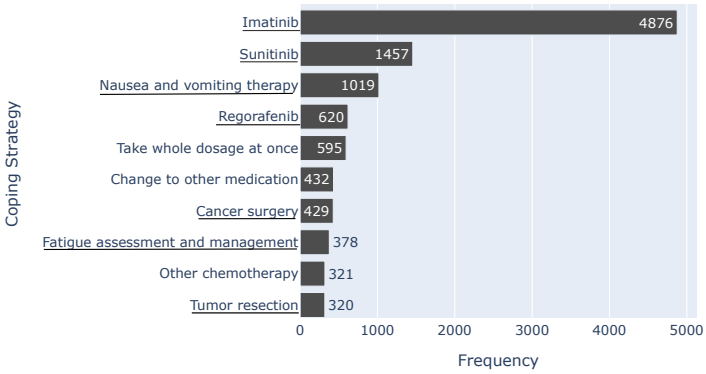
Hierarchy level	Micro F1	Micro R	Micro P	Macro F1
Baseline	0.220 ± 0.011	0.306 ± 0.010	0.172 ± 0.014	0.105 ± 0.010
+1 level (strict)	0.318 ± 0.034	0.336 ± 0.015	0.304 ± 0.048	0.320 ± 0.016
+1 level (lenient)	0.498 ± 0.020	0.350 ± 0.013	0.861 ± 0.063	0.407 ± 0.017
Top categories	0.556 ± 0.018	0.392 ± 0.017	0.952 ± 0.033	0.422 ± 0.040

Table 8.12: Hierarchical evaluation of multi-label semantic similarity SBERT

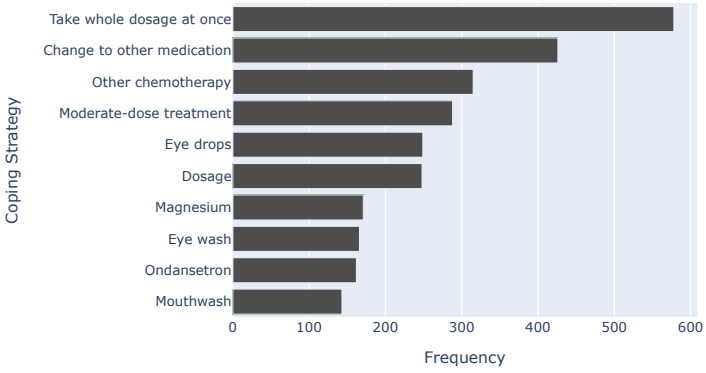
procedures (e.g. cancer surgery) for the disorder itself; side effects (e.g. nausea and vomiting therapy refers to instances of 'nausea'); person names or medical professionals (e.g. oncologist). We manually removed 44 of the 100 most prevalent coping strategies (red lines in Figure 8.4a indicate the removed items in the top-10).

After manual filtering, the total number of coping strategies mentioned was 20.238, of which 3% (694) were negated and 5.5% (1.122) were multi-label. These mentions referred to 2.917 unique coping strategies, which relate to 690 different ADEs. Figure 8.4b shows the most prevalent coping strategies after filtering. Figure 8.5 shows all the coping strategies divided by the highest categories of the ontology (after manual filtering). It appears advice on therapeutic, surgical, or alternative medical procedures ('interventions or procedures' e.g., 'thyroid hormone treatment' or 'moderate-dose treatment') was most prevalent, followed by recommendations to consume medication or supplements and strategies relating to what or how to eat or drink ('eating and drinking').

Figure 8.6 presents the ADEs for which the most coping strategies were provided (See Figure 8.6). The side effect for which the most advice was given was nausea followed by fatigue. In the top 10, various side effects relate to different types of pain (i.e., pain, cramp, painful Mouth) or edema (i.e., edema or periorbital edema). We explored in further detail the most prevalent coping strategies for each of these ADEs. Here we show the results for nausea and cramp, as they most clearly reveal how our semi-automated pipeline can lead to knowledge discovery. We also present results for diarrhea and edema to highlight the problems with negation detection. More analysis for these side effects and the most prevalent coping strategies for the other six side effects are included in Appendix B.



(a) Before manual filtering. Underlined strategies have been manually selected for removal.



(b) After manual filtering.

Figure 8.4: Ten most prevalent coping strategies on the GIST patient forum.

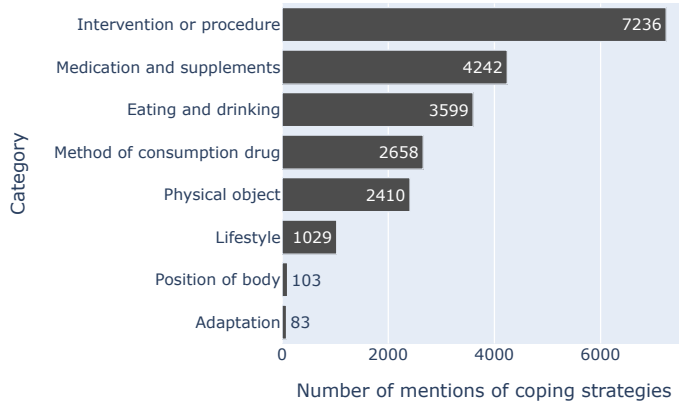


Figure 8.5: Mentions of coping strategies per top category of the ontology (after manual filtering)

Figure 8.7a shows the top 10 coping strategies recommended on the patient forum for nausea. Manual examination of underlying messages reveals that eating and drinking different forms of ginger is recommended, as well as drinking herbal tea (both ginger and peppermint). Patients also recommend taking anti-nausea medication ondansetron and splitting the dosage ('split dosage'). The other categories which relate to how you consume medication (e.g., 'half to one hour before food') do relate to this broader topic, but the specific labels are incorrect. Amongst others, patient recommend to avoid taking medication on an empty stomach and to take it after dinner or just before bed.

Figure 8.7b shows the top 10 coping strategies mentioned on the patient forum for cramps. Manual examination of the underlying messages shows that patients recommend supplements like magnesium, calcium, and potassium ('medication and supplements', 'magnesium', and 'potassium '), food that is high in potassium, tonic water, pickle (juice), and drinking a lot of water ('hydration therapy'). Some patients also recommend exercise ('exercise pain management') although others say it triggers cramps. This is also an example of a case where a coping strategy (exercising) is consistently provided with an incorrect (but semantically similar) label.

Despite decent performance ($F_1 = 0.810$) on our annotated data, qualitative checks revealed that negation detection performed poorly. For instance, manual examination of the underlying messages showed that patients recommend avoiding dairy foods[17] and lactose to reduce diarrhea. However, in Figure 8.8a, only few instances have been negated (red bar) for dairy foods and none for lactose. Another example can be seen in Figure 8.8b, where patients appear divided over whether to avoid or use salt in food ('sodium' and 'low salt food') to reduce edema. The underlying messages, however, are consistent: Patients recommend avoiding salt (blue bar for 'low salt food' and red bar for 'sodium').

[17] The SNOMED concept for dairy is 'dairy foods'

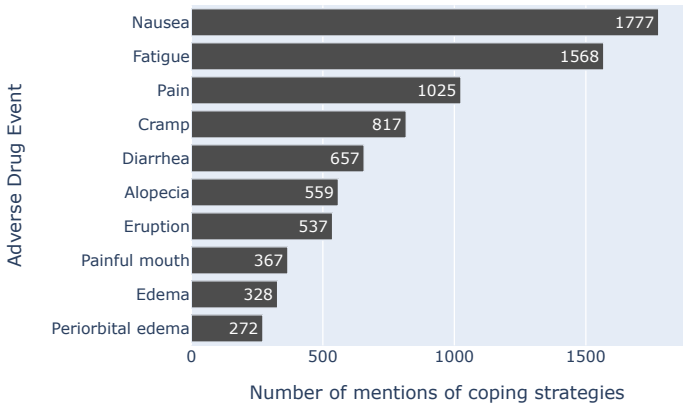


Figure 8.6: The top 10 side effects with the highest number of linked coping strategy mentions (after manual filtering). Alopecia is another term for hair loss, and eruption is another term for rash.

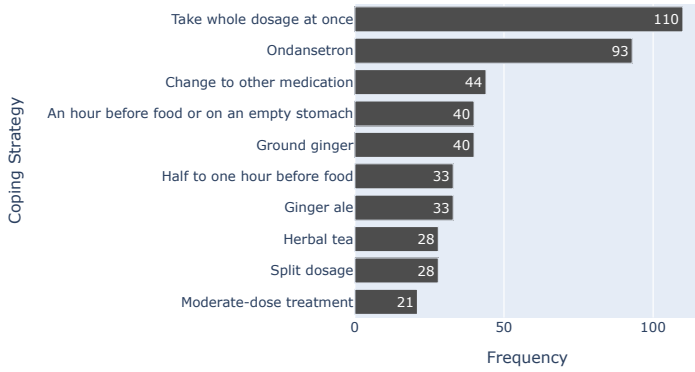
8.5. DISCUSSION

8.5.1. COMPARISON OF APPROACHES

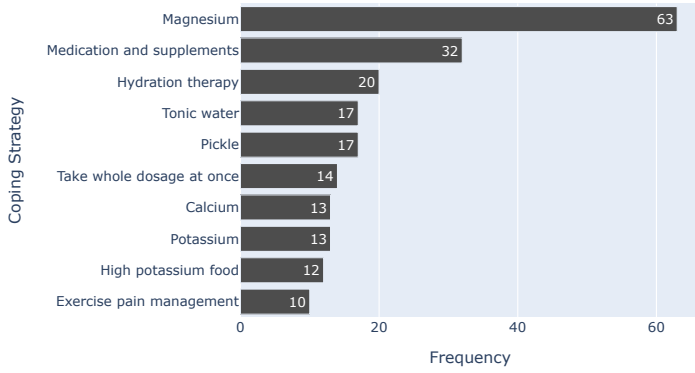
For the extraction of coping strategies for side effects, multi-label classification ($F1$ = 0.220) outperforms named entity recognition (NER) with entity linking (EL) ($F1 = 0.155$). Specifically, Sentence-BERT based on semantic similarity attains the best end-to-end performance, although the quality of the model is still low. Named entity recognition appears to be the bottleneck for the alternative approach, as oracle NER with EL performs even better than multi-label classification ($F1$ = 0.241). This is reflected by the poor token-level NER performance ($F_1 = 0.380$). We found that it is beneficial to include ADE as an additional entity type for NER; This roughly doubled performance (F_1 = 0.200 to $F_1 = 0.380$). Adding a window of one token on each side of the entities further improved performance (to $F_1 = 0.394$), driven by a shift from partially to now fully correct entities. Also, we found that a courser level of ontology matching is considered, the F_1 scores are considerably higher. Overall, we can conclude that multi-label classification is the recommended approach for extracting coping strategies, unless named entity recognition can be improved. One challenge that will remain is the large variety of coping strategy mentions in user-generated text. Increasing the training data will only solve this partly, because there will always be unseen coping strategies in newly seen data.

8.5.2. RELEVANCE OF OUR FINDINGS

These results are also relevant for related tasks, such as the extraction of adverse drug events (ADEs) from social media. Previous work has found that for this task NER is also the bottleneck [159, 193, 194, 335]. Thus, it is worth investigating if multi-label classification is more suited to this task. Moreover, coping strategies for side effects are

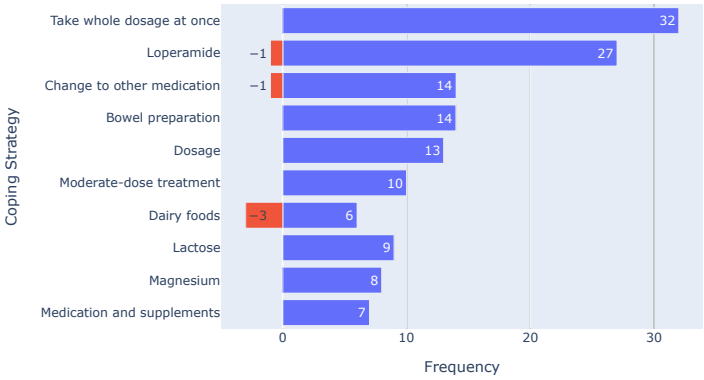


(a) Top 10 coping strategies for nausea

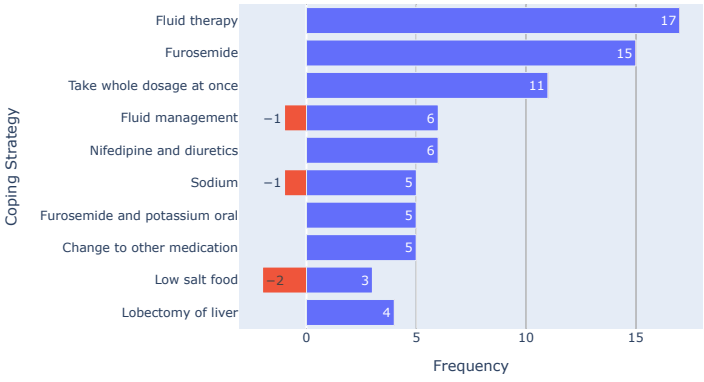


(b) Top 10 coping strategies for cramp

Figure 8.7: Top 10 coping strategies (after manual filtering) without negation



(a) Top 10 coping strategies for diarrhea



(b) Top 10 coping strategies for edema

Figure 8.8: Top 10 coping strategies with negation (after manual filtering). Blue bars indicate that patients recommend taking this strategy and red bars indicate patients that recommend avoiding it (i.e. strategy is negated)

but one type of biomedical complex entity. Unlike named entities, complex entities are often not proper nouns, they tend to be long, and may contain non-entity words (i.e., are discontinuous). Other valuable entities to extract from social media may be advice on psychological coping or coping with the disease in daily life situations e.g. work and childcare. Complex biomedical entities may require different approaches than named entities, and future research is necessary to elucidate whether multi-label classification is consistently preferable to NER with entity linking.

8.5.3. POTENTIAL APPLICATION SETTINGS

Although the quality of our extraction pipeline is insufficient for fully automated knowledge discovery, semi-automated discovery with additional manual qualitative checks can uncover coping strategies for side effects that patients mention online. These can, in turn, be used as input for hypothesis generation. Some examples that we found are drinking ginger tea or taking ondansetron against nausea, and drinking pickle juice or eating potassium-rich food (e.g. bananas) against cramps. Manual examination of the messages underlying a detected strategy can identify cases where the specific label is incorrect (e.g., 'hydration therapy' in Figure 8.7b refers to drinking enough water), as well as cases where it concerns various strategies around a certain topic (e.g., labels referring to how medication should be consumed in Figure 8.7a). These cases likely contribute to the higher performance ($F_1 = 0.498$) when we consider whether the predicted and target labels fall under the same higher order ontological concept.

Expert knowledge is necessary for the manual qualitative checks of the output from the automatic pipeline. Future work could include user studies to estimate the extent of the manual work as well as the extent of the domain knowledge necessary to complete this task. As our work describes the first attempt to tackle this problem, the amount of manual work may also decrease with further improvements to the automatic pipeline. Currently, end-to-end automatic extraction of coping strategies results in a high false positive rate for both MLC and NER+EL. Although recall is more important than precision in a semi-automated system, a high false positive rate is likely to increase the manual work required from experts.

Although we are unable to share our data, we provide the code to visualize and inspect extracted coping strategies[18] in one's own data set. We also share a demonstration of what the visualization would look like.[19] This demonstrates how medical researchers could be aided to conduct adequate qualitative checks and inspect the underlying messages manually using an interface.

Although certain strategies may be self-evident or well known, such as taking anti-nausea medication (e.g., ondansetron) against nausea, others have not been documented previously. Systematic extraction of coping strategies has substantial potential for empowering patients and for generating hypotheses on why these strategies are effective. The coping strategies that are advised should be considered carefully by medical professionals for possible risks before disseminating them amongst patients.

[18] https://github.com/AnneDirkson/CopingStratExtract

[19] https://www.loom.com/share/dda9794a0d354589b95e5b01b5ab23a5

8.5.4. LIMITATIONS

Our work also has a number of limitations. First, the categories included in the ontology are limited to the experiences of GIST patients we collaborated with and the types of coping strategies we encountered on the forum. Although at present our ontology is sufficient to facilitate knowledge discovery, it should be further refined and expanded, for instance through examination of patient forums for other disorders. Furthermore, it would be worthwhile to expand the ontology with categories presented in previous theoretical or qualitative work on coping strategies.

Second, our evaluation of coping strategy extraction is restricted to the labels present in our ground truth data, which cover only 0.6% of the ontology. The performance could thus be overestimated compared to real data if these labels were relatively easy. We preselected discussion threads for annotation based on a high number of patient experiences and at least one drug name using a machine learning model (Chapter 3). Although the performance of this model was good (F_1 = 0.815), discussions around straightforward coping strategies may be easier to identify and thus more likely to be included in the annotated data.

A third limitation is that not all forum posts were subject to coping strategy extraction in the case study. Prior to CS extraction, we selected all posts that contain an ADE and the subsequent 4 posts (see Figure 8.1). Errors in ADE extraction[20] may exclude posts containing coping strategies. Although it may restrict the detected coping strategies, we include this step because previous work has shown that it is beneficial to reduce the data imbalance ratio for extraction [194]. Moreover, our models were trained on similar data. Errors in ADE extraction may also result in the inclusion of posts containing false positives such as symptoms of the disease, resulting in coping strategies that are not directed at resolving adverse drug events.

8.5.5. FUTURE WORK

Aside from further refining our ontology, future work could be directed at exploiting the hierarchical structure of the label space to improve coping strategy extraction, as was done by Rios and Kavuluru [250] and Song et al. [284]. The hierarchical evaluation could also be expanded with more complex hierarchical evaluation metrics such as hierarchical precision and recall [330]. Another possibility would be to include synonyms of the target labels sourced from the UMLS or from the BioPortal term search function. It would also be worthwhile to improve upon our method for ADE–CS relation extraction. Manual error analysis showed that most errors were cases where patients did not explicitly mention which ADE was the target of the coping strategy because it was self-evident to them (e.g. blood pressure medication). Such common sense reasoning appears to often rely on the textual similarity between the ADE and the CS. Thus, relation extraction may be improved by incorporating a similarity metric. Although the performance of negation detection seemed decent (F_1 = 0.810), manual examination of the output revealed negation was not aiding knowledge discovery due to many false positives and negatives. Our heuristics appear insufficient and we recommend future research into improving this module.

Future work could also be directed at researching the low performance of NER for coping strategies, which we expect is due to the descriptive and fuzzy nature of the

[20] ADE extraction has a token-level performance of F_1 0.626 and an entity-level performance of 0.716

entities. We found that the longest correctly identified entity was 9 tokens long, whereas the maximum length of our annotated entities was 29 tokens (see Table 2.3). On average, correctly identified entities were a median of 2 tokens long (± 1 token), partially correctly identified entities were a median of 4 tokens long (± 3 tokens) and missed entities were a median of 2 tokens long (± 3 tokens). It thus appears that missed entities are not on average far longer than correctly identified entities. In contrast, entities that are only partially detected correctly tend to be longer on average. A further investigation of the robustness of NER (e.g. for length and variety of the entities and size of training data) would be insightful for improving the NER model further. Such investigations would also be of interest for other complex entities.

In addition to improving separate modules of the pipeline, future work could include improving their integration. In our current pipeline, the integration of multi-label classification with negation detection and relation extraction was complicated by the need of these modules to know the location of the entity within the sentence. We resolved this by determining the most important token per label that the sentence was labeled with. However, future work could look towards using the attention mechanism of the BERT model underlying multi-label classification, following work on explainable ICD code assignment by Mullenbach et al. [212]. However, this will not be trivial as the Sentence-BERT model is geared towards embedding the entire sentence and does not provide token-specific embeddings. An attention-based approach would also help with differentiating multiple coping strategies (e.g., 'Gatorade, bananas') from a single coping strategy with multiple labels (e.g., 'ginger tea' has the labels 'ginger' and 'herbal tea'). In this work, we defined a coping strategy with two labels as one where the important words were adjacent. Although this is not conventional in related fields such as ICD code detection or ADE extraction, we allow for multiple labels per strategy to curb the exponential growth of the ontology by addition of combined labels.

8.6. CONCLUSION

In this chapter, we have presented a new task, the extraction of coping strategies for side effects from online patient discussions. We developed an ontology for coping strategies, initially tailored to our case of Gastrointestinal Stromal Tumors (GIST), and presented the results for automated extraction method. Moreover, we developed the first pipeline for coping strategy extraction which we use in a case study in which we analyzed an online forum for GIST patients. We showed that automatic extraction of coping strategies for side effects is challenging, with F_1 scores of 0.220 for exact matching to the correct ontology item. We therefore recommend the use of our analysis methods in a semi-automatic fashion in interaction with a human expert to enable the generation of new hypotheses for medical research. Another use would be to discover potentially harmful strategies in the patient-to-patient advice for the purpose of interventions by medical experts.

PART IV:

GIST AS A CASE STUDY

Knowledge, like air, is vital to life.
Like air, no one should be denied it.

Alan Moore, *V for Vendetta*

9

PATIENT FORUMS AS A COMPLEMENTARY DATA SOURCE

Edited from: **Anne Dirkson**, Suzan Verberne, Gerard van Oortmerssen, Hans Gelderblom & Wessel Kraaij. Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers. Major revisions at Scientific Reports.

Current methods for monitoring side effects for a drug after its release onto the market (i.e., pharmacovigilance) result in severe under-reporting of adverse drug events (ADEs). Patient forums have the potential to complement current pharmacovigilance practices by providing real-time uncensored and unsolicited information.

In this chapter, we conduct a case study on a patient forum for Gastrointestinal Stromal Tumor (GIST) patients. We present algorithms that can automatically find the side effects posted on a patient forum and determine automatically for which medication the side effect is being reported. We show that patient forum data can provide suggestions for which ADEs impact quality of life the most: For many side effects, the relative reporting rate differs decidedly from that of the registration trials, including for example cognitive impairment and alopecia as side effects of avapritinib. We also show that our method can provide real-world data for long-term ADEs, such as osteoporosis and tremors for imatinib, and novel ADEs not found in registration trials, such as dry eyes and muscle cramping for imatinib. We thus posit that automated pharmacovigilance from patient forums can provide real-world data for ADEs and should be employed as input for medical hypotheses for rare cancers.

9.1. INTRODUCTION

Adverse Drug Events (ADEs), harmful reactions that result from the intake of medication, pose a major health concern [340] and can have a great impact on the quality of life of a patient [253]. Clinical trials are unable to fully assess the ADEs of a drug due to their limited duration and relatively small sample size, which precludes the discovery of long-term ADEs and rarer ADEs. Furthermore, clinical trials focus on patients in relatively good condition. They mostly exclude elderly, patients with comorbidities, pregnant women, and children [274, 289] and thereby are unable to assess the ADEs that may occur within these patient groups.

Despite post-market surveillance systems, ADEs remain severely under-reported with on average over 90% of ADEs remaining undiscovered [130]. Especially non-serious ADEs are under-reported despite the strong influence they might have on patient adherence and quality of life (QoL) [252].There is an increased recognition that information sources that are more representative of the everyday "real world" are necessary to supplement clinical trials [160, 236]. In recent years, both the FDA and EMA have started to investigate how they can make use of such real world evidence to strengthen their post-market surveillance of drugs (i.e., pharmacovigilance) [244]. One promising resource for the semi-automatic discovery of real-world evidence is social media data [13, 115, 154].

The main advantage of using social media for pharmacovigilance is that it is uncensored and spontaneous. Previous studies have shown that the attitudes of medical professionals cause bias in ADE reporting. Surveys show that medical professionals may not report an adverse drug event for various reasons including lack of time, uncertainty about whether the drug causes the ADE or because the ADE is either trivial or well-known [98, 128]. Social media data has several other distinct advantages compared to other potential information sources. First, the sheer volume of information is not easily obtainable by other means [266]. Second, it has been found that users more often share information with peers than with physicians or at clinical trials [75]. A third advantage is that social media is able to provide near-instantaneous information which allows for real-time monitoring and early signal detection [276]. Yet, some concerns of representativeness of users and data quality have also been put forward [40, 58] which we will address in the discussion.

Patient forums, online communities where patients gather to exchange information and experiences, are a type of social media that could be especially valuable as a resource for ADE detection. It has been estimated that 8% of posts in specific online forums for patients are reports of adverse drug events [114]. Nonetheless, most research at present has focused on generic social media [171, 266]. In this chapter, we present the first empirical case study investigating the value of automated pharmacovigilance from patient forums for a rare cancer. In collaboration with patient organizations, we have collected and extracted ADEs from a large forum of patients with Gastrointestinal Stromal Tumors (GIST). Although it is the most common of the sarcomas, it is a rare disease with an incidence of 10-15 per million per year [285].

9.2. MATERIALS AND METHODS

9.2.1. DATA COLLECTION

In agreement with the GIST International Support Organization, we collected data from their at the time public Facebook group using the Facebook API. The data ranges from 24 Oct 2009 until 1 Nov 2020 and includes 125,161 English messages in 14,631 conversational threads. The 1,493 non-English messages (1.2%) on the forum were removed. On 1 Nov 2020, the forum had 5,555 members and 1,567 users were active on that day.

Our study design and data management plan were approved by the Leiden University privacy officer. We did not collect usernames to protect user privacy in line with data minimization practices. The collected messages were stored securely, and access was restricted to the involved researchers and annotators. For the labeling of data, we did not use commercial tools but set up private servers that were only accessible to the annotators. In accordance with the GDPR (Article 9.2), we did not obtain consent from each user as the GDPR allows for the use of data from publicly accessible forums with justified cause without individual consent. The necessity to take informed consent was formally waived by the Leiden University privacy officer. Nonetheless, we are unable to share the data according to the GDPR, because access to the forum has become restricted to members since our data collection (i.e., it is no longer publicly accessible).

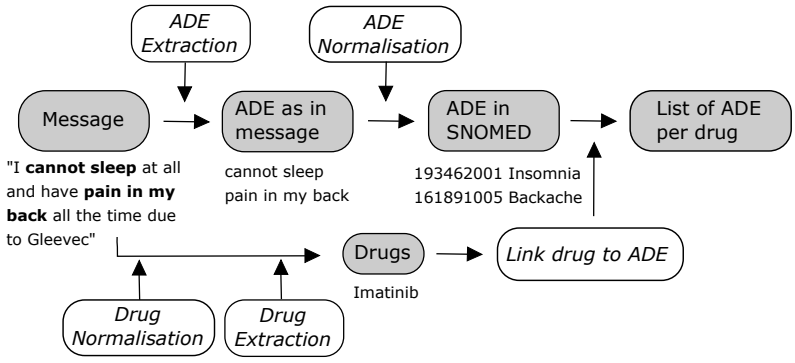


Figure 9.1: An overview of the software pipeline we developed for automatically determining which adverse drug effects (ADE) are mentioned on a patient forum. All italicized parts indicate modules we developed. An example message is provided to clarify each step. ADE: adverse drug events

9.2.2. MACHINE LEARNING PIPELINE

We developed a software pipeline to automatically extract the ADEs from the messages on the patient forum using state-of-the-art methods. As shown in Figure 9.1, we first extract (i.e., ADE Extraction) the words that contain an ADE (e.g., “cannot sleep”) from each message using a specialized information extraction model. This model is trained on forum messages that are manually labeled for ADEs by human annotators. For such tasks where words that contain a certain concept (like an ADE) are extracted (also called Named Entity recognition tasks), predictions are done for each individual word in the sentence. Therefore, the data for training this model is also labeled per word. Specifically, words are labeled for if they are at the Beginning of an entity (B), Inside an entity (I), or Outside an

entity (O) [245]. This is the most common format for sequence labeling tasks, or tasks in which predictions are made per word. Forum messages can contain multiple ADE, which may also span across sentences.

Since posts that contain ADE are a small subset of the data, we wanted to select posts that had a high likelihood to contain an ADE to reduce the time the annotators needed to spend on labeling the data before we had sufficient manually labeled examples to train our model. To create our data selection for manual labeling, we selected all discussions that contained at least one drug name (i.e., at least one exact match with a drug in RxNORM [313]. Prior to data selection, drug names were normalized to their generic variants (e.g., Gleevec to imatinib) and spelling correction was applied to correct misspelt drug names (see Appendix A.1 for more details on preprocessing). From the discussion threads with at least one drug name, we selected the discussions with the highest percentage of posts in which authors shared experiences (such as that you experienced an ADE). In order to estimate which percentage of posts in a thread included patient experiences, we used a previously developed model (Chapter 3). In short, the model was a linear SVC classifier based on trigrams (i.e., sequences of three letters) that could identify experiences with an overall performance (F_1 score) of 0.815.

In total, 4,195 messages (527 discussions) from the GIST forum were labeled by three GIST patients and the first author using an annotation guideline[1]. Subsets of the data (30 threads, between 179 to 211 posts total) were annotated by two annotators to be able to measure to what extent they would label the data the same. Each annotator would label two such overlapping sets. We choose to not have all annotators label the same overlapping data to decrease their workload. For our data, the average agreement between two human annotators was substantial (mean Cohen's $\kappa = 0.71$). A small sample of the annotated data is available in the Appendix A (Table A.4) as an example.

We use 80% of our annotated data and an additional 1,250 messages from a publicly available data set [151] to train our model. Another 10% of our annotated data is used to determine how we can best train our model (i.e., the development data). See Section A.0.2 for the technical details on how we trained our extraction model and Section A.0.1 for details on how the data was preprocessed (i.e., transformed from raw data to input for a machine learning model) before ADE extraction. The remaining 10% of the annotated data is used to evaluate how well our model works on data it has not seen before (i.e., the test data).

We find that on this test data our model has a sensitivity (also called recall) of 0.739: it can retrieve 52.3% of entities fully and 16.6% partially. If it retrieves an entity partially, it has managed to label some of the words of the entity correctly but not all. The specificity of the model is 0.998, meaning that it can correctly identify 99.8% of the true negatives. Its precision of the model is 0.695, meaning that 69.5% of all retrieved entities are true positives. Our model thereby outperforms state-of-the-art models on this task [337]. Yet, its overall performance ($F_1 = 0.72$) is still slightly lower than that of humans (average pairwise $F_1 = 0.80$). Moreover, we find that our model is able to find new adverse drug events for which there were no manually labeled examples (see Section A.0.2 for more details).

We use a specialized machine learning model to link the extracted phrases containing

[1]Available at: `https://github.com/AnneDirkson/ConversationAwareFiltering/tree/master/guideline`

ADE (e.g., "cannot sleep") to concepts in SNOMED-CT (e.g., Insomnia) (i.e., ADE Normalization in Figure 9.1). This allows us to aggregate instances where the same ADE is expressed in different ways. In general terms, this model compares the extracted ADE to all synonyms of concepts in a selected subset of SNOMED to find the best match by ranking how similar each synonym is to the extracted ADE. We train this model using three external data sets [20, 151, 353]. On average, this model can correctly label 64.5% of the ADEs. For an additional 14.6% of the cases, the correct label was included in the top 5. See Section A.0.3 for more details on the training and evaluation of the normalisation model.

We also extract the medication mentioned in the forum message. We first change all medication names to their generic forms (e.g., Gleevec to Imatinib) during Drug Normalization. For this step, we use the RxNORM database [313]. We then extract all generic drug names (e.g., Imatinib) during Drug Extraction using a list of generic drug names from the RxNORM. Finally, we determine which drug the ADE mentioned in the message is most likely to belong to, based on the message and the conversational thread (i.e., Link drug to ADE in Figure 9.1). We designed a simple set of rules (see Section A.4) that select the correct drug 93% of the time if we restrict the possible choices to a list of possible GIST medications (i.e., Imatinib, Sunitinib, Regorafenib, Avapritinib, Ripretinib, Nilotinib, Pazopanib, Ponatinib, Sorafenib) to prevent drugs that resolve the ADE (e.g., "ondansetron" for nausea) from being not chosen. An ADE is linked to no drug ("Unknown") if no drug is mentioned in the message nor in the conversational thread prior to the message.

We describe all technical details of our pipeline in the Appendix A, and we have made our code open-source[2]. Our pipeline for ADE extraction from patient forums is the first that is both publicly available and targeted at English data. van Stekelenborg et al. [321] employed proprietary software and the work by Audeh et al. [13] is on French data. Although we are unable to share the original forum messages, we provide an output file of all extracted ADEs (including which drug they are linked to) for each discussion thread and post as a Supplementary File[3].

9.2.3. DATA ANALYSIS

We investigate the ADEs reported online for all medication that is standard treatment for GIST patients: the first-line treatment imatinib, the second-line treatment sunitinib, the third-line treatment regorafenib, and two recently approved drugs, namely ripretinib, now fourth line treatment, and avapritinib, which was specifically approved for PDGFRA exon 18 mutations. Both were approved in 2020 [103, 311]. All analyses were conducted in Python.

We first identify the 20 most prevalent ADEs for each drug. It is important to note that if an ADE was mentioned twice in one message, it was counted only once. Due to privacy considerations, we do not have access to data on who posted which message and consequently, we are unable to remove cases where the same person posts about an ADE multiple times in different messages. We aggregate ADEs into categories based on the SNOMED-CT hierarchy and the medical expertise of Prof. Dr. Gelderblom.

We also inspect long-term ADEs for GIST medication that has been on the market for

[2] `https://github.com/AnneDirkson/CHyMer`
[3] https://github.com/AnneDirkson/CHyMer/tree/main/suppl

more than five years (i.e., imatinib, sunitinib, and regorafenib). We define long-term ADEs as ADEs that have their first mention on the forum after more than five years of ADE reports concerning that particular drug on the forum. We thereby assume that short-term ADEs will be mentioned at least once in the first five years of ADE reports for a particular drug. Note that we use this proxy because we do not have information on how long patients posting on the forum have been taking a drug as we do not know who posted a message. A limitation of our approach is that rare (but not necessarily long-term) ADEs may not be filtered out. However, by considering how frequently long-term ADEs are reported, we can partially mitigate this issue. We do not aggregate ADEs into larger categories for this analysis because we found that this favored categories that contain very many infrequently occurring ADEs over more relevant ADEs. For the 20 most prevalent long-term ADEs, we manually checked whether there were erroneous categories of ADE that were the result of errors during the extraction step (e.g., "elevated mood" was assigned to any case in which only "elevated" was extracted instead of the full ADE).

Finally, we investigate which ADEs mentioned on the forum are novel (i.e., not reported in the registration trial). We compare our findings to the registration trials for GIST patients instead of the general Summary of Product Characteristics (SmPC) of the drug because the SmPC is not specific to our patient population whereas the registration trials are. For imatinib, we included one phase II trial [78], two phase III trials [36, 331] for Gastrointestinal Stromal Tumor patients based on the approval summary [71] and the work by Reichardt [246]. We also include the ADEs mentioned for GIST in the FDA report for imatinib [308]. For sunitinib, we include one phase III trial for GIST [79] and ADEs mentioned for GIST in the FDA report [309]. For regorafenib, we include one phase III trial for GIST [81] and the ADEs for GIST in the FDA report [310]. We provide supplementary files[4] describing which specific ADEs (with their manually assigned SNOMED CT identifier) were included for each medication.

For this analysis, we set a threshold of 5 as a minimum frequency (i.e., the ADE needs to be mentioned on the forum at least 5 times). We first automatically filtered out any ADEs that were mentioned in the registration trial using their SNOMED-CT identifier. We also filtered out all SNOMED concepts that occurred below these concepts in the SNOMED hierarchy (e.g., leg edema falls under edema and should also be filtered out). Prof. Dr. Gelderblom then manually verified the most prevalent novel ADEs for each drug by comparing them to the ADEs mentioned in the registration trial. We also manually removed any ADE categories from the top 20 that were fully the result of extraction errors.

9.3. RESULTS

Table 9.1 reports the number of ADEs found for each medication type on the GIST patient forum. The number of ADEs reported increases with the number of patients that have been prescribed a certain medication. Manual analysis revealed that most of the "Unknown" cases are in fact not ADEs but symptoms of GIST or side effects of surgery.

For each medication, we can analyze how often ADEs are reported. For example, Figure 9.2 shows the most often reported ADEs reported for avapritinib. Impaired cognition is the most reported ADE followed by fatigue, nausea, edema, and loss of hair.

[4] https://github.com/AnneDirkson/CHyMer/tree/main/suppl

Treatment type	Drug	# of ADE found	# of ADE types
First-line	Imatinib	13,376	685
Second-line	Sunitinib	2,335	324
Third-line	Regorafenib	319	226
Fourth-line	Ripretinib	319	90
PDGFRA exon 18 mutations	Avapritinib	297	112
Off-label	Nilotinib	59	40
Off-label	Pazopanib	51	27
Off-label	Sorafenib	47	32
Off-label	Ponatinib	17	13
	Unknown	2,948	497
	Total	21,051	1,086

Table 9.1: The number of ADEs and ADE types reported on the patient forum for each GIST medication. ADE: adverse drug events

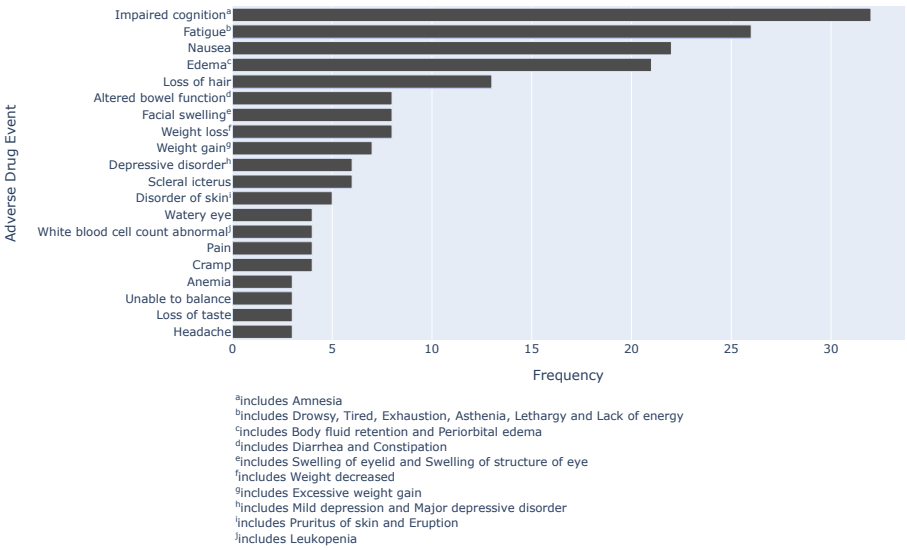


Figure 9.2: The 20 most prevalent adverse drug events reported for avapritinib (formerly BLU-285) on the GIST patient forum

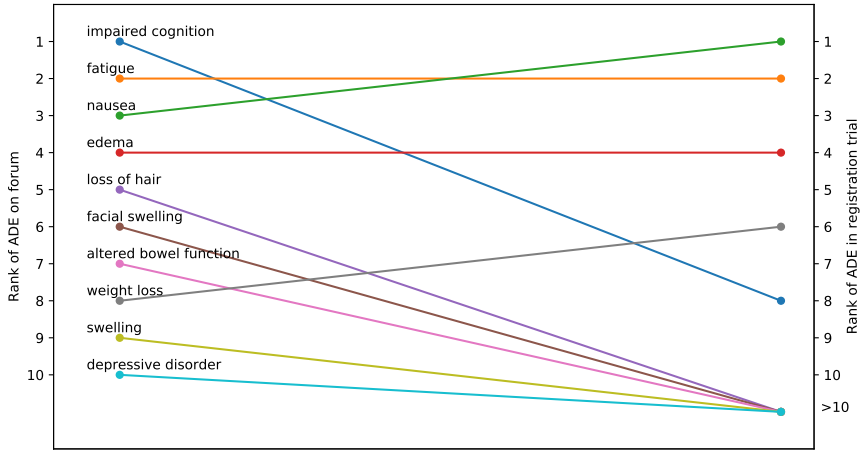


Figure 9.3: The change in rank in terms of prevalence of reporting of the top 10 adverse drug events found for avapritinib on the forum (left) compared to the registration trial (right). ADE: adverse drug events

These ADEs were all reported in the registration trial albeit in the different order as can be seen in Figure 3 (e.g., cognitive impairment was the 8th most prevalent ADE in the registration trial). Incidence rates of ADEs from the clinical trials cannot be compared to the relative reporting rates of ADEs on the forum directly, as nonclinical social media data does not allow us to infer who does not have an ADE. Users that do not report an ADE might still experience it. Thus, reporting rates of ADEs from forum data are only interpretable in a relative sense (i.e., nausea is reported more than fatigue). Nonetheless, relative differences between ADE reporting on a forum and incidence from the registration trial can provide insight into which ADEs are perceived by patients as having the most negative impact on their quality of life; ADEs that are reported relatively more often than expected based on incidence are more salient to patients. Aside from cognitive impairment, we find that, for example, loss of hair (i.e., alopecia) is reported more often than one would expect based on the prevalence in the clinical trial. It was in fact the 23rd or least prevalent ADE at 13% of all patients.

We also analyze ADEs that occur after long-term use of a drug. Figure 9.4 shows the most prevalent long-term ADEs reported for Imatinib on the GIST patient forum. The most reported are dyspnea, toothache, tremor, vertigo and excessive weight gain. It appears that patients suffer from problems with their teeth (i.e., toothache and tooth disorder), muscles (i.e., tremor, muscle atrophy and muscle fatigue), and skeletal system (i.e., osteoporosis). We acknowledge that these ADEs might be related to other factors such as age, and no definitive causality can be deduced from patient reports. Nonetheless, analysis of long-term ADEs on patient forums can provide valuable hypotheses for future research.

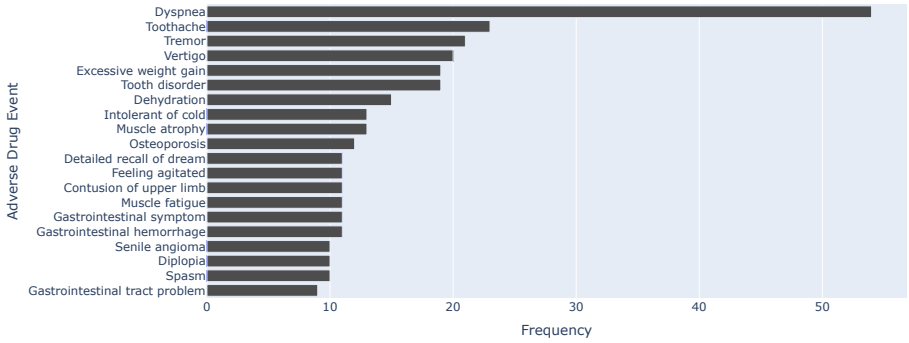


Figure 9.4: The 20 most prevalent long-term adverse drug events reported for imatinib on the forum

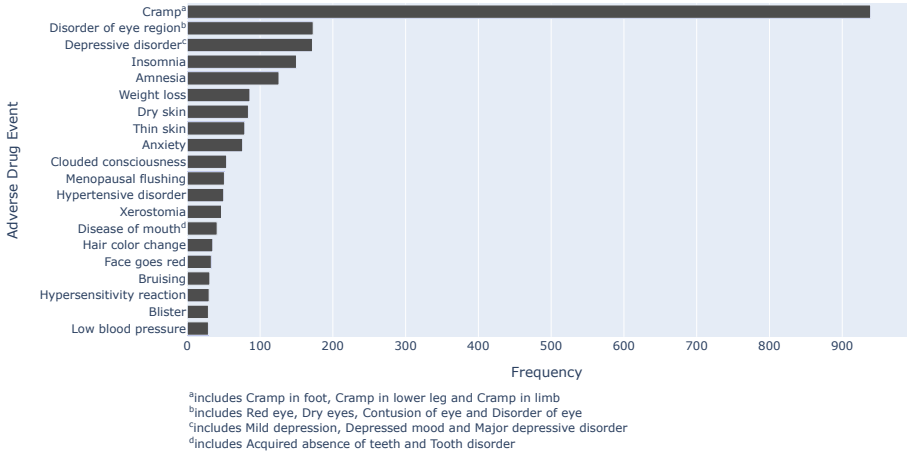


Figure 9.5: The 20 most prevalent adverse drug events for imatinib that were not found in the registration trials.

Finally, we compare the ADEs found in registration trials to those reported on the GIST patient forum to uncover novel ADEs for GIST patients. In contrast to generic social media, disease-specific forums have the unique benefit of providing ADEs for a specific patient population, e.g., GIST patients. In turn, this enables the comparison to known ADEs for that specific patient population through comparison with the relevant clinical trials. For imatinib, we initially found 214 novel ADEs that were reported at least 5 times. Figure 9.5 shows the 20 most prevalent ADEs reported for imatinib that were not reported in the registration trials (the list was curated by an oncologist specialized in sarcomas). Muscle cramp, problems with the eyes, depression, insomnia and amnesia are reported most often. Patients also report novel skin problems (i.e., dry skin, thin skin, bruising and blisters), mouth problems (i.e., xerostomia and tooth problems) and problems with too high or low blood pressure.

Although these ADEs had not been reported during the registration trials for use of imatinib for GIST, many are included in the general Summary of Product Characteristics (or SmPC) of imatinib [101], which means that they have either been found for another disorder (e.g., imatinib is also used by patients with chronic myelogenous leukemia (CML)) or that they were found in the post-marketing phase. Overlap between the SmPC and the 20 most prevalent ADEs that were not reported in the registration trials includes muscle cramps, eye disorders, depression, insomnia, amnesia, weight loss, dry skin, anxiety, high and low blood pressure, xerostomia (dry mouth), bruising and blisters. For ADEs found for other disorders, forum data can provide an indication that these ADEs also occur amongst GIST patients. A high degree of overlap with other patient populations taking imatinib is not surprising, as many ADEs may not be disease-specific. Adverse drug events may also have been added to the SmPC as a result of post-marketing reports by GIST patients. Overlap with these ADEs is promising, as it underscores that forum data may pose an alternative for obtaining such information after release of a drug onto the market.

Forum data can also indicate ADEs that are novel for all imatinib users. Thin skin, clouded consciousness, menopausal flushing, change in hair color, and tooth problems are examples of adverse drug events found on the forum that were not reported in either registration trials for GIST or in the general SmPC.

For more detailed investigations, we provide an interactive demo: `https://dashboard-gist-adr.herokuapp.com/`

9.4. DISCUSSION

In this chapter, we showcase the potential of patient forums as a complementary source of knowledge for pharmacovigilance for rare cancers with a case study. Although ADEs mentioned on a patient forum provide valuable information, causality assessment is necessary before this information can be used as real-world evidence. Similar to spontaneous reporting through official channels, the causality of an adverse drug event needs to be determined before it can be coined an adverse drug response. Whereas an adverse drug event is “any untoward (i.e., unexpected and negative) medical occurrence that may appear during treatment with a pharmaceutical product but which does not necessarily have a causal relationship with the treatment”, an adverse drug response infers a causality relation between drug and effect [102, 341].

Our work differs from previous studies [13, 321] in a number of important aspects. First, in contrast to previous work, we assess ADEs in the context of a specific disease. This enables us to compare our results to registration trials specific to that patient population. We believe that this approach is far more promising than previous approaches which assess ADEs irrespective of which patients are taking the drug, as our approach allows for an investigation of the value of pharmacovigilance from patient forums for specific diseases, including rare and orphan diseases. The focus on rare disorders is the second major difference with previous work. Semi-automatic discovery of ADEs from patient forums is particularly promising for patients with rare diseases, because clinical research into these disorders is scarce. This lack of research is due to a combination of low funding, low interest from pharmaceutical companies, and dispersed patient communities [15, 131, 305]. In fact, according to Aymé et al. [15] online forums could enable the coordinated, trans-geographic effort that is necessary to attain progress for rare diseases. We assessed which ADEs are novel in comparison to those found in the registration trial prior to market release. Thus, we did not take into account which ADEs are discovered by official post-marketing systems, such as by the FDA or EMA, for GIST patients. These systems do not share with researchers which patients reported which ADE and thus all ADEs for a drug are aggregated irrespective of disorder. Comparisons to a specific patient population are thus not possible at this time, although such comparisons would be valuable. There are promising initiatives such as OHDSI[5] that are attempting to make such detailed analysis possible in the future.

Moreover, we are the first study to investigate automatic extraction of long-term side effects from online forums. Some GIST patients take imatinib for longer than 5 or 10 years due to its efficacy [52, 226]. Although post-market clinical studies have evaluated the long-term efficacy of imatinib [52, 226], only one study [226] recorded adverse events and only if they were the reason patients reduced their dosage. The ADEs reported were edema, fatigue, rash, and diarrhea. These ADEs were also reported in the original registration trial and are consequently not specific to long-term usage.

Despite the promise of patient forums as a resource for real-world data, two sources of concern have also been expressed in the literature. A first concern is that the patients that post on the patient forum are not representative for the general patient population [40, 58]. Some patients may lack the skills, access or desire to post on social media [242]. Generally speaking, young people, women and those of higher socioeconomic class are more highly represented on social media [58]. To address this concern, our future work will include a survey amongst GIST patients to investigate the representativity bias on patient forums. Furthermore, this concern is not in fact unique to social media as a potential resource for pharmacovigilance; Clinical trials, surveys and spontaneous reports are also subject to representativity bias. A second concern that has been posited is that the quality of the ADE reports from social media may be inferior. However, studies have shown that reports from patients can be similar in quality compared to those of healthcare professionals [37]. This is also the case for reports on patient forums [322].

Nonetheless, our method does have some limitations due to three sources of noise. Automatic extraction using machine learning methods enables the processing of large volumes of forum messages but also introduces errors into the data as methods do not

[5] https://ohdsi.org/

attain perfect performance e.g., reports may be missed, false positives may be included, or ADEs may be linked to the wrong concept (see Appendix A.0.3 for a more detailed evaluation of errors). A second possible source of noise is negated ADEs, i.e., when a user indicates they do not have a certain ADE. We do not separately identify whether an ADE is negated, because our model is only trained to recognize cases where the ADE is not negated using labeled data in which only non-negated ADE are annotated. However, our model may erroneously extract negated ADE, as they are textually similar to true positives. Furthermore, duplicate records in the data may also introduce noise. Patients may post multiple times about the same ADE and since we do not have access to (anonymized) usernames of posters, we cannot remove these duplicates. Consequently, the real-world data provided by patient forums is noisier overall than the data obtained from spontaneous reports or clinical trials. Automatically extracted ADEs from patient forums should be interpreted in this light; Individual reports may be less reliable but on an aggregate level these reports can provide valuable indications of ADEs and issues that patients are facing. Further clinical research or surveys could be used to validate these hypotheses.

9.5. CONCLUSION

We have shown with a case study of an online forum for GIST patients that patient forums can provide real-world data for both long-term ADEs, such as osteoporosis and tremors for imatinib, as well as for ADEs that were not found in the original registration trials, such as dry eyes and muscle cramping for imatinib. Patient forums are also able to reveal a patient-centric perspective of ADEs by showing which ADEs affect quality of life the most. We find that the relative reporting rate of an ADE often differs decidedly from that of the registration trials. For example, alopecia and cognitive impairment were both reported far more often for avapritinib than would have been expected based on the prevalence in the registration trial. Thus, despite its limitations and noisy nature, automated extraction of ADEs from patient forums can help combat current under-reporting of ADEs by providing much needed real-world data that can function as input for new medical hypotheses and research.

10

COMPARING QUESTIONNAIRE AND FORUM DATA

Edited from: Dide den Hollander, **Anne Dirkson**, Suzan Verberne, Wessel Kraaij, Gerard van Oortmerssen, Hans Gelderblom, Astrid Oosten, Anna K.L. Reyners, Neeltje Steeghs, Winette T.A. van der Graaf, Ingrid Desar and Olga Husson (2022). *Symptoms reported by Gastrointestinal Stromal Tumour (GIST) patients on imatinib treatment: combining questionnaire and forum data.* Supportive Care in Cancer.

In this chapter, we compare the most frequently reported adverse drug effects (ADEs) for imatinib on the GIST patient forum to those reported by Dutch GIST patients for imatinib in a cross-sectional survey study. This survey was conducted amongst 328 patients and consists of items from the EORTC QLQ-C30 and Symptom Based Questionnaire (SBQ).

We find that both the symptoms reported in the survey and those reported on the forum mirror the side effect profiles of imatinib in the registration trials, whereas the relative reporting rates of ADEs differ. The coverage of the more specific EORTC Symptom Based Questionnaire (EORTC-SBQ) is higher (9 of 10) than that of the cancer-generic EORTC QLQ-C30 (4 of 10). One of the most frequently mentioned ADEs on the forum, namely alopecia, was not included in any of the questionnaires.

In conclusion, we find a large degree of overlap between the ADEs reported on the GIST forum and those reported in a cross-sectional survey. Furthermore, the automatically extracted ADEs from the forum data can be used to select the most appropriate questionnaire for patient-reported outcomes, as well as update the questionnaires to include side effects that are relevant to patients.

CRediT author statement

DdH, AD, SV, OH, and GvO conceptualized this study. DdH, AD, OH, and SV contributed to the methodology design. Survey data was collected by HG, AO, AR, NS, WvdG and ID (investigation). The experiments for the forum study were carried out by AD (investigation and software) under the supervision of SV and WK. DdH did the formal analysis that compared the survey and the forum results. DdH and AD wrote the original draft. All authors did review and editing work to finalize the manuscript.

10.1. INTRODUCTION

GISTs represent a rare (10-20 cases per 1,000,000/year) family of mesenchymal tumors arising anywhere along the gastrointestinal tract [285]. Treatment with tyrosine kinase inhibitors (TKIs) improves survival for patients with a gastrointestinal stromal tumor (GIST), both in the adjuvant and palliative setting, but is not without side effects [51, 78, 80]. TKIs are the only effective systemic treatment for high-risk localized and advanced GISTs [73]. Specifically, imatinib has significantly changed the prognosis of non-resectable advanced or metastatic GIST patients: from a median overall survival of 14-18 months up to 57 months [35]. TKIs are taken orally on a daily basis until progressive disease. Especially imatinib is considered to be moderately to well tolerated, at least when compared to conventional chemotherapy [53]. Side effects are seen in virtually all patients, with the most frequent being (periorbital) edema, diarrhea, fatigue, myalgia/musculoskeletal pain, and nausea [145].

Treatment-related side effects or symptoms have significant impact on health-related quality of life (HRQoL) and are an important aspect of HRQoL assessment. HRQoL and symptoms can be assessed using patient-reported outcome measures (PROMs), providing subjective assessments coming directly from the patient, without interpretation by health care professionals or anyone else [306]. The patient perspective is needed to create a more complete overview of treatment-related symptoms, as previous research has shown a gap between the reporting by clinicians and by patients, with clinicians under-reporting symptoms [12, 85]. Another resource of patient-reported data are social media, including patient forums, i.e. online communities where patients exchange information and experiences. Social media are increasingly recognized as sources for reports of patient experiences including symptoms [270]. The reports from social media are unselected, unsolicited, and unbiased, and indicate which symptoms have impact on their health or daily life [251] without the burden of completing questionnaires. Furthermore, its data can also detect emerging issues that may not be mentioned in registration trials or are not covered in existing PROMs [114, 221].

Few studies have investigated patient-reported symptoms in patients with GIST using TKIs. In a qualitative study, 77 different symptoms were reported by GIST patients using TKIs [283]. In another interview study [105], GIST patients with metastatic disease who used imatinib, subjectively described most frequent symptoms as being periorbital edema, nausea, fatigue, exhaustion, cognitive impairment, muscle pain and cramps, and joint pain. Patients also described the considerable impact of these symptoms on their daily lives, again pointing out the gap between physician-reported side effects and the lived experiences of patients. Quantitative data are scarce: one study reported severe fatigue in one third of GIST patients on TKI [239], while another study reported diarrhea, fatigue, and insomnia [69].

To date, interventional studies in GIST patients often use generic (e.g. Short Form Health Survey (SF-36) [333]) or cancer-generic (e.g. EORTC QLQ-C30 [1]) PROMs, that do not assess symptoms specific to TKIs. To incorporate TKI-related symptoms in patient reported outcome measures for GIST patients in future research, more detailed insight into symptom prevalence, relevance, and priority of issues is needed. In the current study, we use two different data sources for patient-reported symptoms, i.e. survey data and data extracted from an online GIST patient forum to examine: (1) the prevalence of symptoms

reported by patients; (2) to what extent the issues reported on a patient forum are covered by existing PROMs (i.e. EORTC QLQ-C30 and items from the EORTC Symptom Based Questionnaire [283]); and (3) the issues that should be prioritized for incorporation in future HRQoL assessment based on the top 10 most prevalent issues.

10.2. METHODS

10.2.1. STUDY DESIGN AND PARTICIPANTS

A cross-sectional population-based survey study was conducted among patients aged ≥ 18 years at diagnosis registered in the Netherlands Cancer Registry (NCR) and who had been diagnosed with GIST (according to the ICD-10-GM codes C15-20, C26, C48, and C80), between January 1, 2008 and December 31, 2018. Only patients diagnosed within one of the GIST expertise centers (Radboud University Medical Center [Nijmegen], Erasmus MC Cancer Institute [Rotterdam], Leiden University Medical Center, The Netherlands Cancer Institute [Amsterdam], and University Medical Center Groningen) were selected. Patients who had cognitive impairment or were too ill at time of the study, according to the advice from their (former) treating specialist, or died prior to the start of the study (according to data from the hospital of diagnosis and/or data from the Dutch municipal personal records database) were excluded. The NCR is a population-based registry which is maintained by the Netherlands Comprehensive Cancer Organization (IKNL) and collects records, including patient, tumor, and treatment characteristics, on all newly diagnosed cancer patients in the Netherlands based on data from the Nationwide Network and Registry of Histo- and Cytopathology (PALGA) in the Netherlands [54].

Data from the (at the time) public Facebook group of GIST Support International (GSI) was used to automatically extract symptoms from the messages on the patient forum. GSI is a United States-based non-profit corporation founded in 2002[1]. The main aims of the organization are to connect GIST patients and their families and friends, to provide information, and to stimulate research. Members are encouraged to interact and share ideas and experiences in the online community. The forum was moderated by assigned, experienced GSI members.

Ethical approval for the cross-sectional study was provided by the medical ethical committee of the Radboud University Medical Center (2019-5888). According to the Dutch law, approval of one ethical committee for questionnaire research is valid for all participating centers. Permission to use data from the Facebook group was given by GSI. Discussions were pseudonymised and messages could not be traced back to individual members. No formal approval was needed for the use of data from the public Facebook group, as the General Data Protection Regulation (GDPR) allows use of data from publicly accessible forums with justified cause.

10.2.2. RECRUITMENT AND DATA COLLECTION

Survey study Eligible patients received an invitation letter from their (ex-)treating physician explaining the goals and procedure of the study. Participants provided informed consent, including permission to link survey data with data from the NCR. Data was collected from September 2020 through June 2021. Survey administration was

[1] https://www.gistsupportorg

done within the Patient Reported Outcomes Following Initial treatment and Long term Evaluation of Survivorship (PROFILES) registry [317]. PROFILES is a data management system set up in 2009 in the Netherlands for the study of the physical and psychosocial impact of cancer and its treatment. PROFILES contains a large web-based component and is linked directly to clinical data from the NCR. Participants could complete the survey online or on paper upon request.

Forum study The English messages from the patient forum were collected on November 1, 2020 and ranged from October 24, 2009 to November 1, 2020. The number of messages was 125,161 in 14,631 conversational threads. A software pipeline was developed to first extract words containing side effects from each forum message and then to automatically determine which side effect is being mentioned. These algorithms were trained on data hand-labeled by human annotators. The sensitivity or recall of the extraction of side effects is 0.739 meaning 73.9% of the side effects reported on the forum can be found by the algorithm. The precision is 0.695, which means that 69.5% of the side effects identified by the algorithm are side effects. The remaining 30.5% are false positives. The accuracy of automatic labeling of side effects with SNOMED-CT concepts is 0.645 (i.e. 64.5% of the side effects are automatically linked to the correct concept in SNOMED-CT) (See Chapter 9). Text about imatinib was extracted from the forum data as well and then linked to the symptom mentioned in the message that it was most likely associated with. The methods of sensitivity and accuracy analysis, text extraction, and linkage of the symptom to imatinib are described in Appendix A.

10.2.3. STUDY MEASURES

Questionnaires and individual items from the EORTC Quality of Life Group (QLG) portfolio were selected as they belong to the most frequently used cancer-specific PROMs worldwide and were developed following well-established guidelines [147]. From the 30-item questionnaire EORTC QLQ-C30, version 3.0 [1], 11 symptom-specific items were evaluated (i.e. dyspnea, pain, feeling weak, appetite loss, nausea, vomiting, constipation, diarrhea, fatigue, problems with concentrating and problems with remembering things). Other symptoms related to TKI use were assessed by 8 additional items from the EORTC Symptom Based Questionnaire (EORTC-SBQ), an 61-item set that was recently developed for patients receiving targeted therapy [1] (i.e. swelling of the face or around the eyes, swelling in any part of the body, muscle aches, pains, or cramps, aches or pains in joints, food and drink tasting different from usual, pain or soreness in mouth, indigestion or heartburn, skin problems). Furthermore, one item about hand-foot syndrome was added from the EORTC Item Library. The items were selected based on prevalence reported in a systematic review of the symptoms associated with TKIs used in the treatment of GIST [282]. One item of own design about the impact of changed physical appearance was added as this was an issue that physicians frequently heard from patients, based on symptoms such as periorbital edema and hair discoloration.

10.2.4. STATISTICAL ANALYSIS

Survey study For analysis, only patients using TKI at time of study participation were selected. In case of low numbers of patients using a specific TKI, the results were only

exploratively compared and presented separately in Appendix C. Prevalence scores for symptoms were determined based on a score of 2 or higher on the 4-point Likert scale being 1- "not at all", 2- "a little", 3- "quite a bit" and 4- "very much", and represented by numbers and percentages out of the total number of patients taking the specific TKI. All analyses were conducted using SPSS version 25.0 (Statistical Package for Social Sciences, Chicago, IL, USA).

Forum study To reduce noise, only side effects that are mentioned at least five times are included, duplicate side effects from the same forum message were excluded, and false positives are reduced by excluding cases where no drug is mentioned in the conversational thread (see Chapter 9). Prevalence of symptoms in the patient forum data was based on how often the symptom was mentioned.

As a secondary analysis, the 10 most prevalent symptoms for each TKI in the survey study and the forum study were compared based on relative reporting rate. Comparison based on absolute prevalence in the two studies was not possible, because of the difference in how prevalence was calculated.

10.3. RESULTS

10.3.1. PARTICIPANTS

In the cross-sectional survey study, a total of 521 (former) GIST patients were invited to participate and 328 (response rate 63%) consented and completed the survey. 107 GIST patients used TKI at time of study participation: 92 used imatinib, 6 sunitinib, 6 regorafenib, and 3 ripretinib. Based on these numbers, we focused on imatinib treatment for this analysis, and results of the explorative analysis for the other TKIs are included in Appendix C. Characteristics of patients using imatinib are shown in Table 10.1. No patient characteristics are available from the forum study.

10.3.2. PREVALENCE SCORES

Prevalence scores for symptoms related to imatinib are shown in Table 10.2. In the survey study, three most prevalent patient-reported symptoms for imatinib were fatigue (73%), muscle pain or cramps (73%), and swelling in the face or around the eyes (59%). In the forum study, for imatinib, the three most prevalent symptoms were fatigue (8.6%), nausea (7.8%), and cramp (6.9%).

10.3.3. RELATION BETWEEN QUESTIONNAIRE AND FORUM SYMPTOMS

Table 10.3 shows the coverage of the 10 most reported symptoms related to imatinib on the online forum in the EORTC QLQ-C30, the EORTC-SBQ, and the EORTC item library. The EORTC QLQ-C30 includes 4 out of 10 most prevalent symptoms on the online forum. The EORTC-SBQ and EORTC item library cover 9 and 10 symptoms, respectively.

Finally, the 10 most prevalent symptoms in the survey study and the forum study were compared based on relative reporting rate, indicated as in descending values in Table 10.4. For imatinib, 7 symptoms overlapped between the two studies. Symptoms from the forum study that were not in the top 10 for imatinib in the survey study were nausea, pain, and alopecia (Table 10.4). Fatigue was the most prevalent symptom both in the survey study

	Imatinib (n = 92)
Age (mean ± SD (range))	66.5 ± 10.0 (28-87)
Time since diagnosis in years (mean ± SD (range))	6.0 ± 2.9 (1.9-12.6)
Sex	
– Male	50
– Female	42
Highest formal education	
– Primary school only	4
– High school	20
– College or university	67
– Missing	1
Relationship status	
– Single	6
– Married/relationship	73
– Seperated/divorced	6
– Widowed	7
Comorbidities	
– None	28
– One	17
– Two or more	47
Comorbidities (specified) *	
– Heart disease	9
– Stroke	2
– Hypertension	21
– Lung disease	7
– Diabetes	7
– Ulcer or stomach disease	3
– Kidney disease	5
– Liver disease	7
– Anemia or other blood disease	13
– Thyroid disease	4
– Depression	8
– Osteoarthritis	26
– Back pain	26
– Rheumatoid arthritis or other joint inflammation	6
– Other cancer	4

Table 10.1: Patient characteristics from the survey study. *Assessed using the Self-Administered Co-morbidity Questionnaire [259]

SURVEY STUDY (n=92)	
Symptoms	Prevalence* (%)
Fatigue	66 (73)
Muscle aches, pains, or cramps	66 (73)
Swelling of the face or around the eyes	54 (59)
Aches or pains in joints	48 (52)
Problems with remembering things	47 (52)
Skin problems (e.g. itchy skin, dry skin, skin discoloration)	46 (50)
Diarrhea	46 (50)
Feeling weak	38 (41)
Indigestion or heartburn	37 (40)
Swelling in any part of the body	35 (38)
Shortness of breath	31 (37)
Food and drink tasting different from usual	33 (36)
Pain	31 (34)
Problems with concentrating	29 (32)
Problems because of changed appearance	28 (30)
Appetite loss	21 (23)
Nausea	21 (23)
Hand-foot syndrome	20 (22)
Pain or soreness in mouth	16 (17)
Constipation	11 (12)
Vomiting	5 (5)
FORUM STUDY (10 most prevalent symptoms**)	
Fatigue	1181 (8.6)
Nausea	1062 (7.8)
Cramp	939 (6.9)
Disorder of skin	680 (5.0)
Oedema	544 (4.0)
Pain[a]	524 (3.8)
Alopecia	466 (3.4)
Altered bowel function[b]	433 (3.2)
Pain in limb[c]	325 (2.4)
Facial swelling	235 (1.7)

Table 10.2: Prevalence scores for symptoms for imatinib. *For the survey data, prevalence is based on percentage of patients with this symptom out of the total number of patients taking imatinib. For the forum data, prevalence is based on percentages of each symptom out of the total number of symptoms for imatinib were calculated. **Adapted from: `https://dashboard-gist-adr.herokuapp.com/` accessed on July 14, 2021. [a]includes: chronic pain and generalized aches and pains [b]includes: constipation and diarrhea [c]includes: any pain in upper or lower limb, excludes: cramp, muscle pain, hand-foot syndrome

Symptoms from forum	EORTC QLQ-C30	EORTC-SBQ	EORTC item library
Fatigue	X	X	X
Nausea	X	X	X
Cramp		X	X
Disorder of skin		X	X
Oedema		X	X
Pain	X		X
Alopecia			X
Altered bowel function	X [a]	X	X
Pain in limb		X	X
Facial swelling		X	X

Table 10.3: Coverage of symptoms from online forum in questionnaires. [a] (diarrhea, constipation)

Rank	Survey	Rank	Forum
1.	Fatigue	1.	Fatigue
	Muscle aches, pains or cramps	2.	Nausea
3.	Swelling of face or around the eyes	3.	Cramp
4.	Aches or pains in joints*	4.	Disorder of skin
	Problems remembering things*	5.	Edema
6.	Skin problems#	6.	Pain
	Diarrhea#	7.	Alopecia
8.	Feeling weak	8.	Altered bowel function
9.	Indigestion or heart burn	9.	Pain in limb
10.	Swelling in any part of body (Edema)	10.	Facial swelling

Table 10.4: Ranking of prevalence of symptoms related to imatinib in survey study and forum study. *same prevalence (52%) # same prevalence (50%)

and the forum study, but the relative reporting rates for the other symptoms differed. Due to the very low number of patients taking sunitinib, regorafenib, or ripretinib in the survey studies, no formal comparison was made. However, explorative analysis showed a similar pattern of overlap between the 10 most prevalent symptoms of the two studies (Appendix C).

10.4. DISCUSSION

This chapter describes the use of two sources for patient-reported symptom rates outside trials in GIST-patients treated with imatinib: surveys and messages from an online patient forum. The most prevalent symptoms in both studies were fatigue and muscle pain or cramps. The EORTC-SBQ and EORTC item library cover the majority of symptoms out of the top 10 most prevalent symptoms on the online forum, but coverage by the EORTC QLQ-C30 was limited. More than half of the 10 most prevalent symptoms were shared between the two sources, but the relative reporting rate of symptoms differed. The prevalent symptom from the online forum that was not covered by the EORTC-SBQ was alopecia. A similar pattern was found for other TKIs prescribed for GIST in the explorative

analysis.

The symptoms found in the survey and the forum study mirror the side effect profiles of imatinib reported in the registration trials, but relative reporting rates differ, for example for muscle cramps [145]. These symptoms occur more frequently over time and may therefore be registered less, or not recognized as adverse drug effects during the initial registration trials. Furthermore, previous work has shown that patients report symptoms earlier and more frequently with worse symptom severity than clinicians [21], and this was particularly the case for muscle cramps and musculoskeletal pain in chronic myeloid leukemia (CML) patients using imatinib [96]. Studies investigating prevalence of patient-reported symptoms in patients with GIST using TKIs are scarce. Previous studies showed that, similar to our results, severe fatigue is common in GIST patients, especially in those taking TKI [48, 239]. Consequently, fatigue had a negative impact on overall quality of life, functional, psychological, and physical well-being [239]. A study investigating symptom burden with the MD Anderson Symptom Inventory for GISTs (MDASI-GIST) identified the most severe symptoms in GIST patients, including muscle soreness and cramping, fatigue, and general weakness [338], matching the most prevalent symptoms found in our data. Unfortunately, the MDASI-GIST is not validated outside the United States. Symptoms that were most prevalent in our study are also the same as the self-reported side effects in a qualitative study, such as muscle pain, cramps, and edema for imatinib [105].

This chapter demonstrates that the EORTC portfolio adequately captures what is important to patients on TKI treatment regarding symptoms and HRQoL, although the cancer-generic EORTC QLQ-C30 on its own lacks most treatment-specific symptoms that were reported on the forum. The forum data also reveals side effects that are not routinely included in PRO-assessment for TKIs, i.e., alopecia. Although it is usually less extensive than in chemotherapy, alopecia is a known adverse effect of TKIs [192, 198] and is more prolonged given the continuous daily dosing schedule. The fact that the reporting rate of alopecia is high on the patient forum indicates that it is an important symptom for patients taking TKIs nonetheless, and can be considered for inclusion HRQoL assessment in future studies.

Differences in relative reporting rate between the two data sources are difficult to interpret, because details on patient characteristics and clinical information were lacking. For example, nausea was ranked higher in the forum study for imatinib treatment than in the survey study. Nausea most frequently occurs in the beginning of TKI-treatment, and declines over time, e.g., with the use of anti-emetics or changes in dosing schedules [145]. As the survey study included patients who were at least 2.5 years since diagnosis at time of participation, we hypothesize that the presence of nausea may have already declined whereas patients posting on the forum about nausea may just have started treatment. Furthermore, one might hypothesize that patients who post messages or complete questionnaires experience more symptoms or higher impact on HRQoL than those that do not, however data on the symptom burden or HRQoL of patients causing them to be active in online cancer communities is scarce. Ector et al. [93] reported that TKI-treatment itself and QoL were not associated with a need for more or less information in chronic myeloid leukemia patients. One study found no differences in use of online support groups for arthritis, fibromyalgia, and breast cancer between patients who post messages and patients who only read messages in case they experienced many or new

symptoms [323]. Comparison with a population that was not active on online support groups is not available. The currently used survey study in Dutch GIST patients included an evaluation of social media use to investigate differences between patients that use social media to converse with other patients and those that do not. Analysis of these data is currently ongoing.

Some limitations need to be taken into consideration. First, online forum data and questionnaire data are unavoidably subject to sample bias [34, 127] and responder bias, respectively. However, as no background information is available for the posters on the online forum, we cannot assess bias in the current analysis. Furthermore, we have no data on which and how many symptoms were reported by family members of GIST patients who also had access to the forum. In recent years, use of online support groups by family members was not significantly different from cancer survivors [104], which could also apply to our forum data. Assessment of responder bias in the survey study was also not possible for the subgroup of patients using TKI included in the current analysis, because information about TKI-treatment was not available for the non-responder population. Second, a formal comparison of symptom prevalence and prioritization between the two datasets was not possible because of the difference in measurement. The survey study only assessed a limited number of predefined symptoms, whereas the forum study used uncensored, unsolicited reports resulting in a larger number of different symptoms (see Chapter 9). Prevalence rates were also calculated differently from the two sources, in which methods for extraction of symptoms and linkage to TKI from the online forum could also have induced false positives, e.g. by extracting text that in fact did not refer to a symptom or linkage of a symptom to the wrong TKI. Additionally, patients might post about the same symptom more than once, which could not be assessed without assessing user names and breaching privacy, causing a skewed distribution in the actual frequency and relative reporting rate of the symptoms. Third, it remains challenging to distinguish for patients, and therefore for researchers as well, if symptoms are solely related to treatment, or to tumor burden or comorbidities [161]. This could be clarified in future studies by asking patients to consider time of onset or improvement after dose modification. Fourth, the number of patients taking other TKIs than imatinib was low in the survey study, limiting generalisability. This is probably due to including patients who were at least 2.5 years since diagnosis, selecting patients with a favorable course of disease and/or response to imatinib. Lastly, insufficient information was available in this study to prioritize symptoms for specific subgroups based on clinical characteristics such time since start of TKI treatment and treatment setting (adjuvant or palliative).

This chapter presents an innovative approach to gain more insight in patient-reported symptoms in GIST patients using TKI. Using automatic extraction of symptoms from an online patient forum and linking them to specific TKIs offers a valuable complementary resource for PRO-data. In addition to interviews with patients and health care professionals that are the primary sources for HRQoL issues in PROMs, forum data may include the perspective of patients who would not be invited or not willing to participate in such interviews. It provides insight into which symptoms are relevant in a large group of patients, which is uncommon for rare cancers, which may help prioritize the selection of HRQoL issues for evaluation (e.g. the high prevalence of muscle cramps in this study). Lastly, forum data raises symptoms or side effects that are not part of existing

PROMs (i.e., alopecia in this study), prompting further investigation whether or not they can be included in PROMs and keeping PROMs up to date. This approach is compatible with the novel flexible strategy for HRQoL assessment by the EORTC QLG, combining existing EORTC questionnaires with add-on symptom questions from the EORTC Item Library [39, 165]. In studies investigating GIST (and possibly other cancer) patients using TKIs, we recommend combining the EORTC QLQ-C30 (to facilitate comparison of cancer-generic HRQoL issues between studies and other (cancer-)populations) with a selection of symptoms from the EORTC-SBQ and individual items from the EORTC Item Library (for symptoms that are missing in the EORTC-SBQ). In studies where only symptoms or adverse events are of interest, the Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) can also be used [22]. In clinical practice, symptoms can be selected based on known side effects from registration trials and clinical experience. Hierarchy in relevance may be based on data from patient forums. More sensitive detection and measurement of symptoms and their impact on HRQoL will help improve assessment of treatment outcomes in research and shared-decision making about (dis-)continuation of treatment in clinical practice. In conclusion, this chapter shows the prevalence of TKI treatment-related symptoms reported by GIST-patients in a survey and on an online patient forum in a real-life setting. Frequently-reported symptoms were not fully covered by cancer-generic measures, and additional issues were reported on the patient forum. Combining these sources of patient-reported data creates a more comprehensive overview of symptom experience and treatment side effects in GIST-patients and helps improve future HRQoL assessment in care and research.

11

ASSESSING SAMPLE BIAS

Edited from: **Anne Dirkson**, Dide den Hollander, Suzan Verberne, Ingrid Desar, Olga Husson, Winette T.A. van der Graaf, Astrid Oosten, An Reyners, Neeltje Steeghs, Wouter van Loon, Hans Gelderblom and Wessel Kraaij. Sample bias in online patient-reported outcomes of Gastrointestinal Stromal Tumor patients: Survey study.

Although representativeness of the online patient population is an often noted as a concern, studies in this field are limited. In this chapter, we investigate the sample bias of patient-centered social media in Dutch Gastrointestinal Stromal Tumor (GIST) patients through a population-based survey amongst 328 patients. We specifically examine peer-to-peer digital communication. We use logistic regression analysis to analyze clinical and demographic differences between forum users and non-users.

Eighteen percent of survey respondents report having contact with fellow patients via social media. 78% of forum users made use of GIST patient forums. We found no statistically significant differences for age, sex, socioeconomic status and time since diagnosis between forum users (n=46) and non-users (n=273). Patient forum users did differ significantly in (self-reported) treatment phase from non-users (P = .001). The odds of being on a patient forum were 2.8 times as high for a patient that is being monitored, compared with a patient that is considered cured. The odds of being on a patient forum were 1.9 times as high for patients that were on curative (adjuvant) treatment and 10 times as high for patients that were in the palliative phase compared to patients that were considered cured.. Forum users also reported a lower level of social functioning (84.8 of 100) than non-users (93.8 of 100) (P = .008).

In conclusion, forum users amongst Dutch GIST patients show no particular bias on the most important demographic variables of age, sex, socio-economic status and time since diagnosis. Nonetheless, our results warrant further investigation of the sample bias in other online patient populations as well as research into methods for bias mitigation

11.1. INTRODUCTION

Online patient forums provide patients with both emotional and informational support [324]. In recent years, social media has also been investigated as a potential complementary information source for patient generated health data, for example for pharmacovigilance [13, 55, 115, 171, 266]. The main advantage of social media is that it offers uncensored information [128] in large quantities [42]. Moreover, patients are more likely to share information with fellow patients than with their physicians [75]. Thus, social media may contain information that is not collected in clinical trials or reported in spontaneous reporting systems.

Post-market surveillance is necessary as clinical trials are of limited duration and suffer from sample bias; they often exclude elderly, patients with comorbidities, and pregnant women [274, 289]. Current post-market medication surveillance systems rely mostly on spontaneous reports of adverse events, medical literature, and observational databases. The majority of these spontaneous reports are made by health professionals. In fact, in the Dutch surveillance system Lareb, only 26.3% of all reports between 2010 and 2015 were made by patients [320].

Reliance on spontaneous reports alone results in a severe under-reporting of adverse drug responses (ADRs) [130]. According to work by Lopez-Gonzalez et al. [189], the under-reporting is associated with reporting of severe ADRs only, fear of ridicule for reporting suspected ADRs, lethargy, and indifference and complacency by professionals (i.e. the idea that only safe drugs are allowed onto the market). Although previous work has shown that the ADRs reported on social media are often less serious than those reported via official channels, they do affect the quality of life of the patient [13]. In fact, social media would be able to provide a more patient-centric view of which ADRs are most salient to patients on a day-to-day basis [197].

Yet, researchers as well as patients have expressed concern about sample bias on social media [13, 23, 32, 44, 58, 276, 287, 301]. Previous research on social media usage in general shows that young people, women, and people of a higher socioeconomic class are generally highly represented [34, 125, 126, 162]. Although there has been some work that shows that these differences persist over time [127, 162], other work indicates that some factors such as age are becoming less influential as the overall adoption of social media is growing. According to a recent report of the Pew Research Center, in 2021 72% of all Americans were using social media including 45% of adults over 65 [10].

Based on studies of the general population of social media users [34, 125–127, 162], it appears that those demographic groups that consume more medication (i.e. the elderly, people of low socioeconomic status, and patients with chronic conditions) are generally not highly represented on social media platforms [189]. However, it remains unclear whether these findings generalize to the specific case of online patient-to-patient communication.

Although there is a large literature base on patient communication forums and the extraction of adverse drug effects, to date the work on sample bias in online patient-to-patient communication is limited to two studies. Prior work on American breast cancer patients [122, 123] using action logs of forum activity in an artificial setting, has shown that users are relatively more likely to be Caucasian than African American. No other significant demographic differences were found between users and non-users. A more

comprehensive overview of literature on patient communication forums for GIST patients on broader topics than bias can be found in the recent work of den Hollander et al. [82] and our own prior work presented in Chapter 9.

Other studies addressed another bias that is relevant when mining social media for patient generated health data: so-called activity bias [323] or the fact that only some users actively post messages. We will use the term "passive users" for forum users that do not post messages and "active users" for forum users that do post messages. Passive users are also commonly referred to as "lurkers" in previous research. Amongst breast cancer patients, Han et al. [123] found that active users were more likely to be younger, Caucasian, living alone and have a greater information need than passive users. Another study [323] specifically compared passive to active community members to posters for breast cancer, arthritis and fibromyalgia and corroborated that posters are younger on average. They also found that active users had a longer disease history and a higher self-reported mental well-being than passive users. In this article, we do not compare active and passive users due to the small sample size.

As Baeza-Yates [16] noted "any remedy of bias starts with awareness of its existence" (p. 54). Thus, to provide a starting point for mitigating bias for the use of patient generated health data from social media in the future, we conducted a survey to investigate sample bias in social media usage amongst Gastrointestinal Stromal Tumor (GIST) patients in the Netherlands relative to the survey sample. GIST is a rare form of cancer which often has a long palliative care trajectory in which patients are treated with chronic, oral medication (tyrosine kinase inhibitors or TKIs) for many years. If caught early, GIST can be cured. Treatment with TKIs can improve survival for GIST patients both in adjuvant and palliative setting, but often also lead to adverse drug events [82]. Patient reports from social media may be especially valuable for rare disorders where patients are sparse and spread out geographically.

In this chapter, we investigated (1) what proportion of patients have contact with fellow patients on social media, (2) why patients abstain from engaging with online patient communities, and (3) to what extent there are significant demographic and clinical differences between those that use social media to converse with patients and those that do not. This study did not assess general social media usage but focused specifically on online communication with other patients. We defined social media as an online communication channel where information and messages are exchanged. When referring to 'online patient communities', we mean online groups on social media where the main purpose of the group is for (certain) patients (e.g., breast cancer patients) to communicate with one another. We use the term online patient communities and patient forums interchangeably.

Based on general social media, we hypothesized that forum users will differ in demographic factors including age, sex and socioeconomic status from non-users. We also hypothesized that forum users will differ in marital status and have a lower level of social functioning than non-users, in line with the social compensation model [199] (i.e. those who have less real life (offline) social support make more use of online digital communities). We also expect that forum users will differ from non-users in their treatment status and that their symptom burden may be higher while their global health scale may be lower. Overall, we expect patients with worse outcomes to be online more

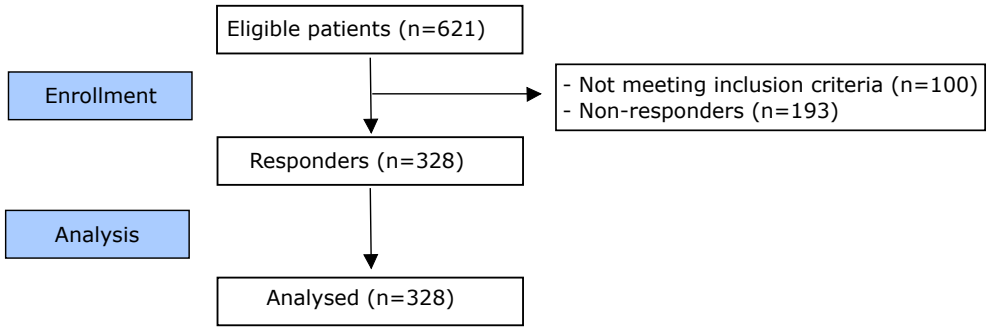


Figure 11.1: CONSORT flow diagram of response rate

often to ask for and receive advice than their peers with better health outcomes.

11.2. METHODS

11.2.1. STUDY DESIGN & PARTICIPANTS

A cross-sectional study was conducted among Dutch GIST patients aged ≥ 18 years at diagnosis, diagnosed between January 1, 2008 and December 31, 2018 in five GIST reference centers. Patients were selected from the Netherlands Cancer Registry (NCR), a population-based registry which is maintained by the Netherlands Comprehensive Cancer Organization (IKNL) and collects patient and tumor characteristics on all newly diagnosed cancer patients in the Netherlands. Exclusion criteria were: cognitive impairment or being too ill at time of the study, according to advice from the (former) treating specialist. Eligible patients were invited by their (ex-)treating physician by letter. Upon consent of the patient, including permission to link the survey data with National Cancer Registry (NCR) data, patients could complete the survey online or on paper upon request. Survey administration was done within the Patient Reported Outcomes Following Initial treatment and Long term Evaluation of Survivorship (PROFILES) registry [317], a data management system set up for the study of the physical and psycho-social impact of cancer and its treatment. PROFILES contains a large web-based component and is linked directly to clinical data from the NCR. Data was collected from September 2020 through June 2021. Ethical approval for the cross-sectional study was provided by the medical ethical committee of the Radboud University Medical Center (2019-5888). According to the Dutch law, approval of one ethical committee for questionnaire research is valid for all participating centers.

11.2.2. SURVEY

Participants completed questions regarding their participation in social media and online patient communities. These questions were developed by the authors. Respondents were asked whether and how patients use digital platforms to have contact with other patients. Possible answers (translated to English) were: “Generic social media (like Facebook or Twitter)”, “General forum or discussion group”, “Specific online patient forum”, “Other, namely” or “I do not use digital communication”. Patients were

provided with the following definition for a digital medium (translated to English): an online communication channel where information and messages are exchanged between participants. Patients were allowed to give multiple answers.

Respondents having contact with other patients online were subsequently asked about their motivations for going online and about their frequency of posting messages. Both questions were adapted from a Dutch survey designed by van Uden-Kraan et al. [323] in collaboration with medical experts and patient representatives. Survey respondents were allowed to provide multiple reasons for engaging with online forums as well as additional reasons in an open text field. Respondents that did not have contact with other patients on specific online patient forums were asked for their reasons for not doing so. Survey respondents were allowed to provide multiple reasons for abstaining from forum use as well as additional reasons in an open text field.

Demographic variables (i.e., age, sex, and socioeconomic status) as well as clinical variables (i.e., tumor type, tumor stage, time since diagnosis, and whether surgery and/or targeted therapy was part of treatment) of survey respondents were collected from the NCR. Survey respondents were additionally asked about their marital status, their current treatment phase, whether they presently use medication, their most recent medication (if any), and the presence of the fourteen possible comorbid conditions measured in the Charlson comorbidity index [61] (heart condition, stroke, high blood pressure, asthma, chronic bronchitis, COPD, diabetes, stomach ulcer, liver disorder, blood disorder, thyroid disease, depression, arthritis, and back pain). Patients were allowed to fill in "Other" for the most recent targeted medication received for treating GIST. This option was intended for new or experimental TKIs, but because patients frequently used this option for other type of medication such as antacids, it was removed for post hoc analysis.

The options patients can choose for self-reported treatment phase are defined as follows: "Cured and not monitored" ("I am cured and no longer need to be monitored") refers to patients that are considered cured after surgery with or without adjuvant imatinib; "On curative treatment" ("I am being treated and can still be cured") refers to patients that are undergoing adjuvant imatinib treatment; "Follow-up after treatment" ("I am not being treated but am only being monitored") refers to patients that are being monitored after surgery with or without adjuvant imatinib and are not undergoing treatment at this time; "On palliative treatment" ("I am being treated but cannot be cured") refers to patients undergoing palliative treatment with thyroid kinase inhibitors and "Best supportive care" ("I cannot be cured but am not being treated") refers to patients that are palliative but are not receiving thyroid kinase inhibitors.

To measure overall health-related quality of life (HRQoL), social functioning and symptom burden, participants completed the EORTC QLQ-C30 version 3.0 [1, 106]. HRQoL was measured with 2 items on a scale of 1 to 7 (from "Very poor" to "Excellent"). Social functioning was measured with 2 items on a scale of 1 to 4 (1- "not at all", 2- "a little", 3- "quite a bit" and 4- "very much"). Eight symptom-specific items were evaluated on the same scale (i.e. dyspnea, pain, insomnia, appetite loss, nausea, constipation, diarrhea, fatigue). Each symptom was measured with 1 to 3 items. The scores for a single symptom from multiple items were averaged. Symptom burden was measured by averaging the eight symptom scales. For 17 respondents symptom burden was not assessed, as there was missing data for at least one symptom. All scales were linearly transformed to a "0-

100" scale in line with the standard scoring manual [100]. A higher score on global QoL or on the scales measuring the level of functioning translates to a higher level of functioning and QoL, whereas a higher score on the symptom scales means the patient experiences more complaints.

Any questions that were not previously validated were pre-tested with patients and changed according to their feedback (cognitive debriefing). The questionnaires cannot be shared due to copyright restrictions.

11.2.3. DATA ANALYSIS

Reasons for abstaining and engaging with online patient-to-patient communication were analyzed manually by the first author. Fifty-two cases (16%) contain missing data. As none of these cases are forum users, the data is not missing completely at random (MCAR). Since we do not observe any other patterns in the missing data that cannot be explained by the variables on which we have full information, the data is missing at random (MAR). Since the missing data occurs in multiple variables, we used Multivariate Imputation by Chained Equations (MICE) [186, 316] to impute these values, which is valid under assumption of MAR.We generated 20 imputed data sets that include all survey respondents (N=328).

We aimed to analyze whether there were statistically significant differences in demographic and clinical characteristics as well as quality of life measures between forum users and non-users. For each imputed data set, a multiple logistic regression analysis was performed with forum use as the dependent and demographic and clinical factors are independent variables (see Section 11.2.2). The effects of one variable on forum use are thus conditional on the other variables in the model. We report the average and standard deviation of the 20 imputed data sets, since this provides a more reliable result than a single run. We use the mean as the average for all variables except the *P*-value where we use the median [94].

For this analysis, the number of variables was restricted by the small size of the user population. We checked for multicollinearity using Variance Inflation Factor (VIF) tests. If the VIF value was larger than 3, we removed one of the collinear explanatory variables. In total, we removed two variables accordingly: the most recent medication, and whether the patient is on systemic treatment currently ("On systemic treatment currently"). Note that whether the patient received targeted therapy at some point in time ("Targeted therapy") is included. Moreover, two categories of self-reported treatment phase, namely on palliative treatment and on best supportive care needed to be merged into one palliative category, as only one patient was receiving best supportive care. Benjamini-Hochberg correction [29] was used to adjust for multiple testing (controlling the false discovery rate or Type I errors at 0.05). Analyses were conducted using statsmodels (v 0.12.2) and scipy (v 1.4.1) in Python 3.7. Graphs were created with plotly (v 5.3.1) in Python.

11.3. RESULTS

11.3.1. PARTICIPANTS

In total, 328 GIST patients responded to the survey (response rate 64%). The median age of the participants was 67 years (range 28 to 91 years), and 53.8% were male (see Table

11.1). On average, they had been diagnosed with GIST for 5 years ranging from 1 to 12 years since diagnosis. One hundred and sixty-two participants (49%) are in follow-up after treatment with curative intent, 61 (19%) were considered cured and are not in follow-up, and 100 receive systematic treatment, either with curative (34) or palliative intent (67). One patient received best supportive care only.

Nine of the patients did not answer the question about forum usage and their forum use is thus unknown. Consequently, the sum of the reported numbers under forum usage (Y and N) does not equal the number reported for all respondents. The percentages were calculated based on the counts per category, i.e., 55% of non-users are male (150 of the 273 non-users).

			FORUM USER*	
		ALL	N	Y
Count		328	273	46
Age	Median (Range)	67 (28-91)	68 (28- 91)	65 (47 – 83)
Sex	Count (%)			
– Male		174 (53%)	150 (55%)	21 (45%)
– Female		154 (47%)	123 (45%)	25 (54%)
Socio-economic status	Count (%)			
– Low (1-3)		90 (28%)	74 (27%)	13 (28%)
– Intermediate (4-7)		132 (40%)	113 (41%)	16 (35%)
– High (8-10)		106 (32%)	86 (32%)	17 (37%)
Marital status	Count (%)			
– Married or living together		246 (75%)	202 (74%)	38 (83%)
– Single		79 (24%)	68 (25%)	8 (17%)
– Missing		4 (1%)	3 (1%)	0
Time since diagnosis (in years)	Median (Range)	5 (1- 12)	5 (1-12)	5 (2 -11)
Tumor stage	Count (%)			
– I		121 (37%)	109 (40%)	8 (17%)
– II		61 (19%)	51 (19%)	10 (22%)
– III		66 (20%)	53 (19%)	10 (22%)
– IV		55 (17%)	38 (14%)	16 (35%)
– Missing		25 (8%)	22 (8%)	2 (4%)
Surgery	Count (%)			
– Yes		287 (88%)	244 (89%)	36 (78%)
– No		41 (12%)	29 (11%)	10 (22%)
Targeted therapy	Count (%)			
– Yes		214 (65%)	170 (62%)	39 (85%)
– No		114 (35%)	103 (38%)	7 (15%)
Self-reported current treatment status	Count (%)			
– Cured and not monitored		61 (19%)	56 (21%)	2 (4%)
– On curative treatment		34 (10%)	31 (11%)	3 (7%)
– Follow-up after treatment		162 (49%)	139 (51%)	19 (41%)

– On palliative treatment		66 (20%)	42 (15%)	22 (48%)
– Best supportive care		1 (0.3%)	1 (0.4%)	0
– Missing		4 (1%)	4 (1%)	0
On systemic treatment currently	Count (%)			
– Yes		208 (63%) **	181 (66%)	25 (54%)
– No		108 (33%)	83 (30%)	21 (46%)
– Missing		12 (4%)	9 (3%)	0
Most recent medication	Count (%)			
– Imatinib		178 (54%)	140 (51%)	31 (67%)
– Sunitinib		9 (3%)	7 (3%)	2 (4%)
– Regorafenib		6 (2%)	4 (1%)	2 (4%)
– Other		15 (5%)	8 (3%)	4 (9%)
– No therapy		114 (35%)	103 (38%)	7 (15%)
– Missing		14 (4%)	11 (4%)	0
Number of comorbid conditions	Count (%)			
– 0		109 (33%)	92 (34%)	14 (30%)
– 1		71 (22%)	59 (22%)	10 (22%)
– 2 +		146 (45%)	120 (44%)	22 (48%)
– Missing		2 (0.6%)	2 (0.7%)	0
Global health scale (0-100)	Mean (SD)	78.6 (18.1)	79.0 (17.7)	76.1 (20.1)
Symptom burden (0-100)	Mean (SD)	12.1 (12.8)	11.4 (12.6)	15.6 (13.0)
Social functioning (0-100)	Mean (SD)	92.4 (18.9)	93.8 (17.1)	84.8 (26.0)

Table 11.1: Demographic characteristics of survey respondents. *Nine participants did not answer this question. **It appears patients that are currently being monitored may have misunderstood this question, inflating the number of patients that are currently on targeted medication for GIST

11.3.2. SOCIAL MEDIA USAGE

As shown in Table 11.2, 81% of GIST patients do not have contact with other patients via any social media platform. We distinguished between specific social media, such as patient forums, and general social media such as Twitter or Facebook[1]. Of the patients to communicate with peers via social media, , the majority (46 of 59) make use of specific online patient forums focused on GIST. Only 6 respondents make use of general social media platforms to communicate with other GIST patients and only 7 use more general cancer-related forums or discussion groups for this purpose.

11.3.3. REASONS FOR ABSTAINING FROM ONLINE COMMUNICATION WITH PEERS

Table 11.3 shows the reasons the 265 non-users report for not using any digital medium to communicate with fellow patients. Patients were allowed to report multiple reasons.

[1] Although it is possible for patient communities to exists as groups on general social media platforms (in fact: the biggest GIST forum is a Facebook group), general social media refers to communication with peers outside of GIST-specific communities on these general social media platforms.

Which of the following digital media do you use to have contact with other patients? (Indicate all that apply)	Frequency
General social media (like Facebook or Twitter)	6 (2%)
General cancer-related forum or discussion group	7 (2%)
GIST specific online patient forum	46 (14%)
Any social medium	59 (18%)
None or via another medium than social media	265 (81%)
Missing	4 (1%)
TOTAL	328

Table 11.2: Descriptive statistics for usage of social media to have contact with other patients. Respondents can give multiple answers to this question.

Twenty patients did not fill in the question. The most common reason reported for abstaining from using a digital medium to communicate with peers was that they felt no need to do so (31.8%), followed by finding it too confronting (13.5%) and not knowing where to find online communities (12.2%). Only eight participants reported not using social media to communicate with other patients because they lack the skills or access to do so.

11.3.4. REASONS FOR ENGAGING WITH PATIENT FORUMS

Survey respondents most frequently used patient forums to communicate with other patients. The number of survey respondents that use other online platforms was too small to analyze how they compare to non-users. Thus, we will focus on analyzing the sample bias of GIST-specific patient forums. Hereafter, when we refer to 'forum users we mean users of GIST-specific patient forums.

Table 11.4 shows the reasons users reported for engaging with a disease-specific patient forum. The most prevalent reasons were having a question on or having heard new information about their illness (both 40%) and being curious how the other members are doing (36%). Another prevalent trigger was experiencing new symptoms (31%).

11.3.5. CHARACTERISTICS OF THE PATIENT FORUM USERS

In total, 85.8% (273) of the participants were not making use of specialized GIST patient forums (see Table 1). The difference in model fit between the multiple logistic regression model and the null model was found to be statistically significant in all 20 imputed data sets ($LR = 47.0 \pm 1.48$, $df = 20$, $P < .001 \pm 0.0004$). Likelihood ratio tests between the full model and the full model without the variable were used to test the significance of individual variables.

Table 11.5 reports the average results of twenty runs of multiple logistic regression models of which factors influence forum use. Our analysis shows that self-reported treatment status differs significantly between forum users and non-users for each run ($LR = 10.6$, $P = .001$). The odds of being on a patient forum were 2.8 times as high for a patient that is being monitored, compared with a patient that is considered cured. The odds of being on a patient forum were 1.9 times as high for patients that were on curative

Self-reported reason	Frequency
Feel no need to communicate (digitally) with other patients	78 (29.4%)
I find it too confronting or burdensome	33 (12.5%)
I would not know where to find online communities	30 (11.3%)
There are too many negative comments	26 (9.8%)
I do not have the time	23 (8.7%)
The information shared is useless or less valuable	20 (7.5%)
I communicate with enough patients personally or via another non-digital medium	18 (6.8%)
I do not use social media, lack a computer or digital skills or do not like obtaining information digitally	8 (3.0%)
I obtain sufficient information via my medical specialist or searching online	7 (2.6%)
I no longer have symptoms or do not like to consider myself a patient	5 (1.9%)
I have privacy concerns	3 (1.1%)
They do not exist in my language	2 (0.8%)
No particular reason	1 (0.4%)
Missing	20 (7.5%)'
Total number of users that do not use any digital communication with other patients	265

Table 11.3: The reasons non-users report for not using social media to communicate with other patients. Multiple answers were possible.

Self-reported reason	Frequency
When I have a question about my illness	18 (40%)
When I have heard new information about my illness	18 (40%)
When I am curious how other members are doing	16 (36%)
When I get new symptoms	14 (31%)
When I have a lot of symptoms	6 (13%)
When I feel insecure	5 (11%)
Before making a medical choice	4 (9%)
For the company	4 (9%)
Because other members expect me to be there	2 (4%)
When I feel lonely	1 (2%)
It is part of my daily routine	1 (2%)
I never use the forum anymore	1 (2%)
Total number of users that do communicate via patient forums with other patients	45

Table 11.4: The reasons users report for visiting the patient forum. Multiple answers were possible

(adjuvant) treatment and 10 times as high for patients that were in the palliative phase compared to patients that were considered cured.

We did not find significant differences between forum users and non-users for other disease-related characteristics when they were adjusted for covariates. We also did not find significant differences in key demographic variables such as age, sex, socioeconomic status, and marital status. Yet, we did find a significant difference in the level of social functioning in seven of twenty runs ($LR = 6.8$, $P = .008$). Forum users on average reported a lower level of social functioning than non-users (84.8 vs 93.8 of 100). These scores were normalized according to the scoring manual[100]. Converting the normalized values back to the mean raw score gives a 1.19 for forum users and a 1.46 for non-users, where 1 translates to the highest possible value for self-reported social functioning on the survey items.

	COEF	SE	*df*	LR	*P*	ODDS RATIO		
						5%		95%
Intercept	-2.795 (0.541)	2.081 (0.034)	-	-	-	-	-	-
Age	-0.020 (0.004)	0.018 (0.0002)	1	1.318 (0.535)	0.264 (0.100)	0.945 (0.004)	0.980 (0.004)	1.015 (0.004)
Sex	0.622 (0.039)	0.371 (0.004)	1	2.858 (0.348)	0.088 (0.021)	0.900 (0.034)	1.863 (0.072)	3.860 (0.161)
Socio-economic status	-	-	2	1.365 (0.485)	0.249 (0.081)	-	-	-
– Low (1-3)	-	-		-	-	-	-	-
– Intermediate (4-7)	-0.386 (0.102)	0.441 (0.006)		-	-	0.288 (0.027)	0.683 (0.066)	1.622 (0.163)
– High (8-10)	0.048 (0.101)	0.440 (0.005)		-	-	0.445 (0.041)	1.055 (0.102)	2.499 (0.255)
Marital status	-0.321 (0.085)	0.468 (0.006)	1	0.517 (0.251)	0.467 (0.114)	0.291 (0.038)	0.728 (0.062)	1.820 (0.062)
Time since diagnosis	0.016 (0.019)	0.073 (0.001)	1	0.118 (0.159)	0.847 (0.152)	0.880 (0.018)	1.017 (0.019)	1.174 (0.022)
Tumor type	0.567 (0.063)	0.377 (0.003)	1	2.292 (0.519)	0.129 (0.042)	0.843 (0.054)	1.766 (0.112)	3.699 (0.237)
Tumor stage	-	-	3	2.602 (0.920)	0.116 (0.071)	-	-	-
– I	-	-		-	-	-	-	-
– II	0.506 (0.126)	0.547 (0.009)		-	-	0.572 (0.072)	1.671 (0.211)	4.886 (0.632)
– III	0.212 (0.214)	0.6262 (0.013)		-	-	0.372 (0.090)	1.266 (0.290)	4.309 (0.942)
– IV	0.863 (0.170)	0.663 (0.013)		-	-	0.655 (0.119)	2.405 (0.433)	8.834 (1.613)
Surgery	0.039 (0.124)	0.574 (0.012)	1	0.053 (0.103)	0.887 (0.103)	0.339 (0.036)	1.048 (0.123)	3.237 (0.421)
Targeted therapy	0.120 (0.099)	0.573 (0.010)	1	0.073 (0.080)	0.826 (0.097)	0.368 (0.032)	1.133 (0.111)	3.490 (0.383)
Self-reported current treatment status	-	-	3	10.673 (1.096)	**0.001**** (0.0006)	-	-	-
– Cured and not monitored	-	-		-	-	-	-	-
– On curative treatment	0.590 (0.264)	1.071 (0.050)		-	-	0.225 (0.040)	1.863 (0.446)	15.559 (4.651)

– Follow-up after treatment	1.026 (0.262)	0.865 (0.061)		-	-	0.519 (0.080)	2.881 (0.685)	16.179 (5.095)
– Palliative	2.288 (0.232)	0.965 (0.057)		-	-	1.503 (0.229)	10.111 (2.208)	68.678 (19.838)
Number of comorbid conditions	-	-	2	0.419 (0.259)	0.532 (0.144)	-	-	-
– 0	-	-		-	-	-	-	-
– 1	0.275 (0.108)	0.497 (0.007)		-	-	0.501 (0.057)	1.325 (0.143)	3.505 (0.362)
– 2 +	0.207 (0.077)	0.451 (0.005)		-	-	0.510 (0.036)	1.234 (0.093)	2.987 (0.240)
Global health scale/ QoL	0.029 (0.002)	0.014 (0.0001)	1	4.382 (0.686)	0.036 (0.016)	1.001 (0.002)	1.039 (0.002)	1.057 (0.002)
Symptom burden	-0.0003 (0.005)	0.018 (0.0004)	1	0.088 (0.096)	0.830 (0.114)	0.964 (0.006)	1.000 (0.005)	1.036 (0.005)
Social functioning	-0.025 (0.002)	0.009 (0.0002)	1	6.865 (0.900)	**0.008*** (0.005)	0.958 (0.002)	0.975 (0.001)	0.994 (0.002)

Table 11.5: Average results (with SD) of a logistic regression of demographic and clinical characteristics of patient forum users and non-users using MICE with 20 runs. For the p-value, the median is reported. *Significant after Benjamini-Hochberg correction in some runs **Significant after Benjamini-Hochberg correction in all runs.

11.4. DISCUSSION

11.4.1. SUMMARY OF FINDINGS

A survey was conducted amongst 328 GIST patients in the Netherlands. Our results show that the majority of survey respondents do not have contact with other patients via social media. They indicate a large heterogeneity of reasons of why they abstain from doing so with the most prevalent being they feel no need, find it too confronting, or do not know where to find such online communities. Of the minority that do use social media for this purpose, most use disease-specific patient forums. The most prevalent reasons for accessing a patient forum are i) having a question about their illness, ii) having heard new information, iii) experiencing new symptoms, or iv) wondering how other patients are doing. Patient forum users differ significantly in their (self-reported) treatment phase from non-users. Patients in the palliative phase are 10 times more likely to be forum users than patients that are cured. Patients that are monitored approximately 3 times and patients undergoing curative treatment approximately 2 times more likely to be users than cured patients. For seven of the twenty data imputations, forum users also have a significantly lower level of social functioning.

11.4.2. COMPARISON WITH EXISTING LITERATURE

In contrast to the general population of social media users, patient forum users do not appear to differ in age, sex and socioeconomic status from non-users. On the one hand,

this may be an effect of the increasingly more widespread adoption of social media. This idea is supported by the small number of patients that indicate they lack the skills or access to be on social media (3.3%). On the other hand, it is also possible that there is less demographic bias on patient forums than in general social media. This may be related to the widely different goals that users have with their participation. Although a feeling of community and social support may overlap, patients report motivations such as questions around their illness, and the experience of new symptoms that normal social media users are unlikely to share.

Prior work [122] on forum usage amongst breast cancer patients did not find significant differences between forum users and non-users in terms of clinical characteristics, i.e. stage of cancer and quality of life. We similarly did not find any significant differences for these characteristics, although we did find significant differences for clinical characteristics that prior work did not investigate i.e. treatment phase. Prior work also found that amongst breast cancer patients, non-users and passive users had greater offline social support than posters. Their results supported the social compensation model [199] i.e. those who have less real life (offline) social support use and engage online with digital communities. The lower offline support of forum users compared to non-users in our data also supports this theory. However, passive users appear to have a lower offline support than active users amongst GIST patients. This would support the competing theory: the social engagement model [163] i.e. those that have more social resources will use and benefit from online social communities more. Consequently, our data offers support for the social compensation model for those who use a forum (i.e. those with less real-life support are more likely to be using a forum) and social engagement theory for those who actually actively engage with the forum community (i.e. users with sufficient social resources will be active and benefit more). Demographic differences in terms of age, marital status (i.e. living alone or not) and disease duration between passive and active users that were found in previous work were not evident from our data.

11.4.3. LIMITATIONS

First and foremost, we only studied a specific patient population in a single country and thus further research is needed to elucidate to what extent our results are generalizable. Patients in other countries may have lower digital access or skills or may not wish to use social media for patient-to-patient communication for other reasons (e.g. other privacy laws or country-specific customs).

Our choice of GIST patients as a target population may also impact to which disorders our results generalize to. Patients with GIST have a median age of mid 60s [285], meaning that it is on average an older population than the general population that is often studied for social media usage. Our results may consequently also generalize better to conditions that are prevalent in an older population. GIST is also characterized by a long palliative phase in which patients receive treatment. Thus, our results may also generalize better to conditions that similarly have a long treatment duration (e.g. metastasized breast cancer). As GIST is a rare type of cancer, our results may also generalize better to rare than common conditions. Further research into other patient populations should be able to provide more insight into the differences in forum usage between rare and common conditions. The fact that GIST is a rare condition makes it an interesting first case. Patient generated

health data from social media are particularly promising for rare conditions due to their dispersed patient communities and the scarcity of research [15].

A second limitation of this study is the small sample size. Amongst the 328 respondents, only 46 indicate that they use patient forums. Nonetheless, given the low incidence of GIST at 12.7 per million [113], this is a substantial number of participants. A third limitation is the sample bias of the survey itself. There may be two underlying factors, namely selection bias and responder bias. Patients who were too ill or had cognitive impairment were excluded, leading to selection bias. A non-responder analysis was conducted using the database of the Netherlands Cancer Registry to assess the extent of the responder bias. After correcting for multiple testing, no significant differences were found in terms of age, sex, socioeconomic status, time since diagnosis, tumor stage, and primary treatment between respondents and non-respondents. Moreover, it was possible to fill in the survey on paper, which prevents the exclusion of less digitally adept patients on these grounds.

11.4.4. FUTURE WORK AND RECOMMENDATIONS

Based on this work, a number of recommendations can be made. First, out of the possible digital resources that can be used to source complementary real-world evidence, patient forums should be preferred over other social media. Our results reveal that GIST patients strongly prefer disease-specific patient forums over general social media for communicating with fellow patients. Yet, most research in this field currently focuses on general social media such as Twitter [171, 266]. Our results are in line with previous work that estimates ADR reports to be more prevalent in patient forums than on Twitter [114].

Although we find that there is sample bias in patient forum users and thus the sample is not wholly representative for the patient population, sample bias is also a concern for other sources of patient reports. Understanding which patients are over- and under-represented on online forums is the first step to using online patient reports as a complementary resource, for instance for pharmacovigilance. For pharmacovigilance specifically, it is not of great concern that patients that are considered cured and not undergoing treatment currently are under-represented. Future work on comparing the sample bias of clinical trials to that of online patient forums would be beneficial to further explore its complementary value in detail. It would also be valuable to gain more insight into the different types of forum users.

Secondly, it may be beneficial to create awareness amongst medical professionals that patients are more likely to search for information in online patient communities when they have questions, have been given new information, or have new symptoms. Medical professionals could try to aid patients in their information needs by pointing them towards such resources in these cases. This may also take away the barrier mentioned by patients that they do not know where to find such online communities.

Thirdly, future work into the sample bias of patient forums for other patient populations is necessary as this study was limited to a single population in a single country. Nonetheless, our work is a stepping stone towards dissuading the concerns that researchers have expressed regarding the sample bias of social media [13, 23, 32, 44, 58, 276, 287, 301] by unravelling on which characteristics users differ significantly from the overall patient population. Future work could also investigate how compensatory

measures can be implemented to statistically correct for sample bias. As these factors may not be known for the participants of a forum, it would also be worthwhile to consider to what extent correcting for sample bias is possible without this information.

11.5. CONCLUSION

In this chapter, we investigated how representative participants in patient forums are for the general patient population by conducting a survey amongst GIST patients in the Netherlands. We found statistically significant differences in terms of treatment phase and offline social support between forum users and non-users. The consequent over- and under-representation of certain types of patients should be considered when sourcing patient forums for patient generated health data. As our study was limited to a single patient population, a further investigation of sample and activity bias in other online patient populations is warranted as well as research into methods for bias mitigation. Sample bias is inherent to any information source and only through awareness of these biases can these resources be used as a source for complementary real-world evidence in the future.

PART V:

DISCUSSION

I know where I want to be, but I can't possibly be sure we're anywhere near it

Fantastic Mr. Fox, Roald Dahl

12

DISCUSSION

In Fantastic Mr Fox, Roald Dahl manages to capture the uncertainty of progress: Despite knowing where we want to go, we do not know how long the road will be. In this thesis, we have made small steps toward the end goal of integrating patient-reported experiences from social media into the medical knowledge base. We explored how to extract patient-reported experiences from patient forums and to what extent and under which conditions they can lead to knowledge discovery and generate hypotheses.

In this chapter, we present and reflect upon our main findings for each research question in Section 12.1. We then answer our main research question in Section 12.2. We conclude with ideas for future research and recommendations in Sections 12.3 and 12.4.

12.1. MAIN FINDINGS

1. To what extent can corpus-driven spelling correction reduce the out-of-vocabulary (OOV) rate in medical social media text and improve the accuracy of subsequent classification tasks?

In Chapter 2, we aimed to correct spelling errors in domain-specific data without losing information due to false positives: domain-specific terms that disappear because they are "corrected" to other words. This challenge has been largely overlooked, although it can hinder downstream tasks. During the extraction of adverse drug events (ADE) in Chapter 7, spelling errors in the original PsyTAR data hindered automatic alignment to human-annotated ADE phrases. We created an additional corpus from these spelling mistakes that we have made publicly available. [1]

In this chapter, we experimented with unsupervised corpus-driven spelling correction. Our method combines edit-based similarity with cosine similarity based on a static (or context independent) word2vec language model. However, in recent years, context-aware language models (e.g., BERT) have entered the stage. We expect that context-aware embeddings will improve upon the static word2vec embeddings in our method based on

[1] Available at: `https://github.com/AnneDirkson/SpellingCorpus`

recent work on spelling correction of user-generated text [43, 140, 213]. Muller et al. [213] found that fine-tuning BERT with a small amount (3,000) of training sentences outperformed MoNoise [318] which relies on static embeddings. Bucur et al. [43] framed lexical normalization as a machine translation task from noisy to normalized text. They used the multilingual BART model [178] which outperforms other transfer learning models for sequence-to-sequence tasks such as machine translation. BART is trained by corrupting text with noise and then learning to reconstruct the original text. Both these methods were supervised. There has also been one study which combined BERT with edit distance in an unsupervised manner. Hu et al. [140] found that using edit distance to find candidate words for correction and then using BERT to check whether the candidate fits well within the sentence works better than the reverse: using BERT to select candidates and then finding similar words using edit distance in the candidate list. Their work, however, focused only on spelling *correction* and presumed misspellings were already detected. In contrast, our method can both detect and correct spelling mistakes in an unsupervised manner. Similar to our work however, the work by Hu et al. [140] supports the notion that it is advantageous to combine language models with edit distance for unsupervised spelling correction.

We would even argue that unsupervised spelling correction in niche domains and user-generated data cannot be resolved by improved language models alone. Language in general and slang in particular is dynamic and thus would require constant updating of these models. Moreover, to date, methods that rely solely on language models have all been supervised, as they require training data to be fine-tuned for detecting and correcting spelling mistakes.

We found that our unsupervised method can reduce out-of-vocabulary terms in two cancer-related medical forums and that it targets misspelled medical terms. Many of the remaining OOV-terms are not spelling errors but rather real words, slang, names, and abbreviations. Our method is not dependent on corpus size and works for noisy corpora (up to a noise ratio of 8%). Yet, the benefit to downstream tasks is marginal: our method can significantly improve accuracy on only two of the six classification tasks. We expect that tasks that rely more strongly on individual terms, such as extraction tasks, may benefit more.

2. Which features distinguish patient narratives from other social media text and how can they best be identified?

In Chapter 3, we analyzed the characteristics of patient narratives on a disease-specific forum. Patient narratives were characterized by past tense, first-person pronouns (i.e., talking about oneself), and health topics. In contrast, non-narrative posts were associated with future tense, second-person pronouns (i.e., talking to others) and emotional support. We found that character 3-grams were more effective for identifying patient narratives (F_1=0.815) than psycho-linguistic features or document embeddings. Their strength appears to lie in their ability to cluster relevant word types, such as tyrosine kinase cancer medication which ends in *'nib'*. These results underscore that simple methods should not be disregarded.

Our work also shows that narrative detection is a difficult task for annotators. Despite a substantial inter-annotator agreement (κ = 0.69), a significant proportion of model

errors were due to incorrect annotation (36.9% of the false positives and 36.2% of the false negatives). In hindsight, we should have provided our annotators with the conversational context of posts they were annotating. Human annotators and algorithms alike could not classify posts that lacked context, which were often answers to questions earlier in the conversation. Furthermore, it appears that an exact definition of when someone is sharing an experience is challenging and it would be beneficial for the medical informatics community to further refine the definition of a patient narrative.

3. To what extent can the addition of conversational context to state-of-the-art models improve the identification of relevant posts?

In Chapter 4, we incorporate conversational structure into BERT models using two different approaches: adding a sequential model or manually engineered features. We investigate the benefit of conversational structure to the identification of relevant posts in health-related social media discussions. We use the only publicly available medical relevance classification data set that includes the conversational structure [158] as a benchmark. This data focuses on identifying posts with medical misinformation. In addition, we annotated patient discussions for the presence of ADEs and coping strategies for dealing with ADEs. These are the specific patient narratives that we are interested in extracting. Narrative detection from Chapter 3 was used to pre-select discussions that had a high likelihood of containing ADEs. We selected 527 discussions for annotation that contained (1) at least one drug name according to a match with RxNorm [314] and (2) a high percentage of posts in which authors shared experiences. We find that a sequential layer can improve precision for one of three data sets, whereas manually engineered features do not aid performance. Nevertheless, we find that the distribution of relevant posts across discussion threads is skewed and that within a conversational thread relevant posts cluster together.

Although conversational context did not benefit performance in two of three data sets, the conversational context of social media posts should not be ignored altogether. We recommend splitting folds per discussion thread to prevent dependencies between posts from biasing model performance. We also recommend providing conversational context to annotators during labeling, as reactions to social media posts may not be understood in isolation and relations may span across posts. This was apparent for narrative detection in Chapter 2; drug-ADE relations in Chapter 9 and relations between ADEs and coping strategies in Chapter 8.

4. How effective are default transfer learning methods for extracting and normalizing adverse drug events?

In Chapter 5, we show that transfer learning using default and recommended settings can give above average results for various NLP tasks using health-related Twitter data. For extracting ADEs, we used the FLAIR package [4] which uses a BiLSTM-CRF model for NER and allows for the stacking of different embeddings through concatenation. We found that adding a classifier for sentences containing ADEs did not benefit ADE extraction and that combining BERT with FLAIR embeddings led to the highest performance (F_1=0.625). Yet, removing the FLAIR embeddings only results in a drop in F_1 score of

0.003. It is a worthwhile consideration whether the higher computational cost of adding flair embeddings weighs up against the small absolute increase in performance. Such considerations are currently not given sufficient prominence in the NLP community where absolute performance is often the only criterion.

For the classification of personal health mentions, the model trained on a larger corpus including the DIEGO Drug Chatter corpus [263] was outperformed by a model trained on a smaller corpus of task data supplemented with labeled data from different disease domains (mean F_1=0.793). Thus, our results highlight that more data is not always better, especially when explicitly considering generalisability as was done in this task.

5. How vulnerable are BERT models for Named Entity Recognition to adversarial attack and to which variation are they most vulnerable?

In Chapter 6, we analyze which changes are able to fool BERT models to make wrong predictions for extraction tasks. These changes are crafted to deliberately try to fool the model (i.e., adversarial attack). We found that under these conditions BERT models are highly vulnerable to entities being replaced with more rare entities, as well as to words in the local context of the entity being replaced with synonyms rarely seen during training. For the latter, a single change was often sufficient. We find that the vulnerability of the model to synonym replacement in the entity context depends on the vocabulary it employs. BioBERT, which retains the BERT vocabulary, is as vulnerable to synonym replacement as the generic BERT model. In contrast, SciBERT, which has a domain-specific vocabulary, is more vulnerable to synonym replacement. Although a domain-specific vocabulary can be beneficial, it is important for researchers to recognize the drawbacks: The vocabulary of BERT is limited in size and thus a models' ability to deal with more common language may be compromised.

These results underscore the need for research into methods that make BERT models more robust. We recommend researching zero-shot learning and masking strategies for entities in the training data to improve robustness to emergent entities. We also suggest investigating alternative pre-training schemes such as curriculum learning to combat vulnerability to rare synonyms.

Our conclusions are underscored by more recent work by Lin et al. [185]. Their work is methodologically similar to our own; The biggest difference is that Lin et al. [185] generate one perturbed data set of out-of-distribution data to measure robustness instead of targeting the weaknesses of specific models to generate adversarial examples. Their perturbation methods also differ: At the entity level they replace entities with entities from the same fine-grained semantic class according to WikiData. To perturb the context, they mask tokens in the sentence and use a pre-trained language model to generate substitutions. They select predicted tokens ranking between the 100th and 200th spot to create a more challenging context. In line with our results, they find that even the best NER models are brittle to adversarial examples with a larger drop in performance for entity-level attacks than for context-level attacks. Moreover, they find that models that perform better on in-domain data also perform better on out-of-distribution data, i.e., transfer learning models are more robust than BiLSTM-CRF models. Finally, they apply three data augmentation methods to improve robustness with limited success. Random masking

(i.e., replacing the letters of entities with random ones) appears to make RoBERTA slightly more robust to entity-level attacks.

In our experience, the interest of the NLP community for weaknesses of models that are now commonly employed is limited. Although there has been increasing interest exemplified by the creation of the BlackBoxNLP workshop, it is not proportional to the rapid development and improvement of existing models. A promising development is the compulsory responsible NLP checklist [2] which includes "security considerations" under the potential risks posed to AI models.

The limited interest from the NLP community stands in stark contrast to the recommendations made in recent years for responsible AI. Technical robustness and safety has been put forward as one of the seven requirements for trustworthy AI according to the EU High-Level Expert Group on AI (AI HLEG) [133]. The AI HLEG group states that models should be resilient against attack to prevent malicious use and that safeguards should be put in place to prevent unintended adverse impacts. This document does not stand alone. According to Fjeld et al. [111], 29 of 36 prominent documents on AI governance principles report safety and security of models as a principle, where secure generally refers to being "resistant to being compromised by unauthorized parties" (p. 5). Thus, guidelines for responsible AI highlight the need for understanding and combating a model's vulnerabilities to ensure robust models. Neglecting these limitations may have detrimental and unethical consequences [133]. In line with principles of trustworthy AI, we advise conducting more research into the vulnerabilities of context-aware models and possible mitigation strategies. In our opinion, organizers of NLP conferences should encourage and create more room for such work.

6. To what extent can a fuzzy continuous representation of discontinuous entities improve the extraction and normalization of adverse drug events?

In Chapter 7, we present an alternative, simplified representation scheme for discontinuous entities, FuzzyBIO. We find that for ADE extraction, a FuzzyBIO representation can improve recall and result in a higher percentage of correctly identified entities for two of the three data sets compared to the more complex but commonly employed BIOHD representation. Our simplified representation also improves end-to-end performance for continuous and composite entities in these two data sets, while it is detrimental to performance in the third data set. Our results lead us to conclude that a complex, more exact depiction (BIOHD) should not always be preferred over a simpler, less exact representation (FuzzyBIO) as this is not necessarily beneficial to the end goal; A more accurate representation can also make a task unnecessarily complicated. It seems that FuzzyBIO is able to simplify the extraction task for BERT models by standardizing entities into continuous sequences that always start with a B-tag and by excluding rare tags such as the H-tag.

The FuzzyBIO representation is less beneficial for end-to-end performance on *disjoint* or split entities (e.g., "*eyes* are feeling *dry*"). The most likely explanation based on our additional analysis is that normalization algorithms that normalize the extracted mention to a common entity form are not used to dealing with the additional noise: FuzzyBIO essentially makes disjoint entities continuous by including the words in between disjoint sections of the entity (i.e., labeling them with the I-tag). An example can be seen in Table

12.1, where a perfect extraction with FuzzyBIO would result in "Muscles are constantly quivering" while perfect extraction with BIOHD would result in "Muscles quivering". This raises the question whether normalization algorithms should be trained with noisier examples to make them more robust to noise. Overall, our work in this chapter exemplifies that it is important to not consider and perfect modules in isolation but in relation to the end-to-end pipeline.

	Muscles	are	constantly	**quivering**
BIOHD	DB	O	O	DI
FuzzyBIO	B	I	I	I

Table 12.1: An example of a disjoint ADE mention represented by the BIOHD and FuzzyBIO schemes.

7. To what extent can coping strategies for ADEs be extracted automatically from online patient discussions?

In Chapter 8, we introduce a new task: the extraction of coping strategies (CS) for ADE from online patient discussions. We present the first ontology for coping strategies, and compare baseline methods for its end-to-end resolution. We find that multi-label classification with Sentence-BERT (F_1 = 0.220) outperforms named entity recognition (NER) with entity linking (EL) (F_1 = 0.155). For the latter, NER appears to be the bottleneck, as oracle NER with EL (F_1 = 0.241) can outperform multi-label classification.

Despite the low performance, our end-to-end extraction pipeline works sufficiently well to enable knowledge discovery in a semi-automatic fashion. With additional manual qualitative checks, it is possible to uncover true recommended coping strategies. For example, we found that patients recommend drinking ginger or mint tea against nausea and that they recommend drinking pickle juice or eating potassium-rich food (e.g., bananas) against cramps. These manual checks are indispensable to filter out false positives due to adverse drug events, surgeries, primary medication, medical professionals, or person names being marked as coping strategies. They also are necessary to identify clusters of messages that may indeed refer to coping strategies and thus are insightful but where the predicted label is incorrect. Furthermore, there are cases where there are errors in the relation extraction, i.e., the coping strategies do not concern the ADE in question. Lastly, qualitative checks revealed that our negation detection is unable to differentiate between doing or avoiding something. For instance, patients recommend *avoiding* dairy and lactose for diarrhea (see Figure 8.8a) but these have not been negated. Another example is that patients recommend low salt food and *avoiding* salt ("sodium") for edema (see Figure 8.8b), but the latter is not negated.

Nonetheless, given the large and long-tailed label space, these results are very promising. Semi-automatically extracting coping strategies from online discussions could provide researchers with new hypotheses and facilitate medical research into why certain strategies work. Some strategies may work because they disrupt the efficacy of the primary medication, i.e., you do not experience an ADE (anymore) because the medication is not working. Although we are unable to provide the annotated data to the community, we

do provide the code to the pipeline and a dashboard for manually exploring the output[2]. A demonstration of the dashboard can be viewed at `https://www.loom.com/share/dda9794a0d354589b95e5b01b5ab23a5`.

The extraction of coping strategies could also empower patients themselves. However, given the noisy output, it is important to consider how and when the discovered coping strategies should be presented to the patients, as dissemination may unduly endorse the strategies. Medical professionals and patient representatives should be involved in considering the possible risks of dissemination and their mitigation.

It is still an open question to what extent our pipeline is able to extract coping strategies from other forums and for other conditions. Our ontology may be one of the limiting factors, as the categories that were included were determined based on the coping strategies we encountered on the forum for GIST patients and the experiences of GIST patients we collaborated with. For each category (e.g., food or interventions), we did include an entire category from another ontology so as to not bias our ontology to certain strategies within these categories. Another possible limiting factor is the efficacy of the ADE extraction pipeline (see Appendix A for details) on other forums and for other conditions. This pipeline was also primarily developed and validated on the GIST forum. In our CS extraction, we use the extracted ADE to select posts that may include coping strategies and for extracting for which ADE the coping strategy is recommended.

8. How can the automated gathering of real-world evidence of adverse drug events from online patient forums complement pharmacovigilance for rare cancers?

In Chapter 9, we demonstrate that patient forum data can reveal which ADEs impact quality of life the most: For many side effects, the relative reporting rate in forum data differs decidedly from that of the registration trials. Patient forums can also provide real-world evidence for both long-term and novel ADEs, i.e., ADEs not found during registration trials. Our pipeline is able to deal with zero-shot cases: It can extract ADEs not present in the training data.

Long term effects were assessed by subtracting ADEs mentioned in the first five years from those mentioned in later years for a certain drug. Although this proxy is able to find ADEs that clinicians recognize from the clinic (e.g., eye problems and osteoporosis), it is suboptimal. It would be preferable if long-term effects were determined based on how long the poster has been taking the drug. This could possibly be deduced by linking forum posts of the same user and checking for the first mention of drug usage. Psuedonymized usernames would be sufficient for this purpose. Unfortunately we did not have access to psuedonymized usernames, because the data was fully anonymized by Facebook. The Facebook API removes all usernames instead of psuedonymizing them. From our work in Chapter 11 we know that amongst GIST patients, palliative patients are more likely to be forum users than patients undergoing treatment with curative intent. Since the palliative phase of GIST is long and patients take medication during this phase, this result supports the idea that long-term effects of medication could be found on the patient forum.

Adverse drug events in clinical trials do not have explicit concept identifiers, although generally clinical trials use the Common Terminology of Adverse Drug Events (CTCAE)

[2] `https://github.com/AnneDirkson/CopingStratExtract`

[307] without reporting the identifiers. The lack of identifiers complicates the automatic comparison between the ADEs on the forum and the ADEs known from the trial. Moreover, even manual mapping to the CTCAE is insufficient as there is no mapping between the CTCAE and SNOMED-CT, which is the ontology we use for mapping the ADEs from the forum. We choose SNOMED-CT for its interoperability with previous research and with the OHDSI[3] project, a collaborative effort to create an overarching vocabulary for various sources of observational health data. Thus, we resorted to manually mapping the ADEs from clinical trials to SNOMED-CT identifiers to permit automatic filtering. We supplemented automatic filtering with qualitative filtering by a medical professional, because patients often tend to report the consequences of an underlying ADE (e.g., swelling) instead of the underlying cause (e.g., edema) which is reported in the clinical trial. In conclusion, we found automated filtering alone to be insufficient at present and both manual work and medical knowledge are still essential for this step.

Many of the chapters in this thesis describe work that contributed to the overall pipeline for ADE extraction described in Chapter 9 as well as to the pipeline for CS extraction described in Chapter 8. We present an overview of how the components from various chapters were employed in Figure 12.1. We did not perform a holistic end-to-end analysis of the various components we developed, and as such we do not know the impact of for instance our spelling correction (Chapter 2) on ADE extraction or the impact of ADE extraction (Chapter 9) on the extraction of coping strategies. This is a limitation of our current work and we hope that others will revisit these questions in future research.

9. To what extent are the adverse drug events reported on a GIST patient forum covered by existing patient-reported outcome measures namely the EORTC QLQ-C30 and the EORTC Symptom Based Questionnaire?

In Chapter 10, we collaborate with medical professionals to compare ADEs from the GIST forum to answers on patient-reported outcome measures (PROMs). Similar to the forum data, the symptoms reported in the survey amongst 328 Dutch GIST patients mirror the side effect profiles of imatinib in the registration trials but the relative reporting rates differ. Although most prevalent symptoms overlap between the forum and survey outcomes, forum data can help to choose the most appropriate PROM. The more specific EORTC Symptom Based Questionnaire (EORTC-SBQ) is preferable as it covers 9 of the 10 most reported symptoms on the online forum, while coverage of the cancer-generic EORTC QLQ-C30 is limited to 4 of the 10. Thus, even for the most suited PROM, forum data can reveal side effects that are not routinely included (i.e., alopecia) and can be used to update questionnaires to include side effects relevant to patients. The EORTC item library, which contains all EORTC items, does include an item on alopecia that could supplement EORTC-SBQ. Integrating ADEs from forum data into healthcare in this manner would not have surfaced without the involvement of medical professionals. We believe their involvement is key to attaining the end goal of integrating online patient-reported outcomes into healthcare. We also expect such collaborative efforts to be met with more support from the medical community.

[3] https://ohdsi.org/

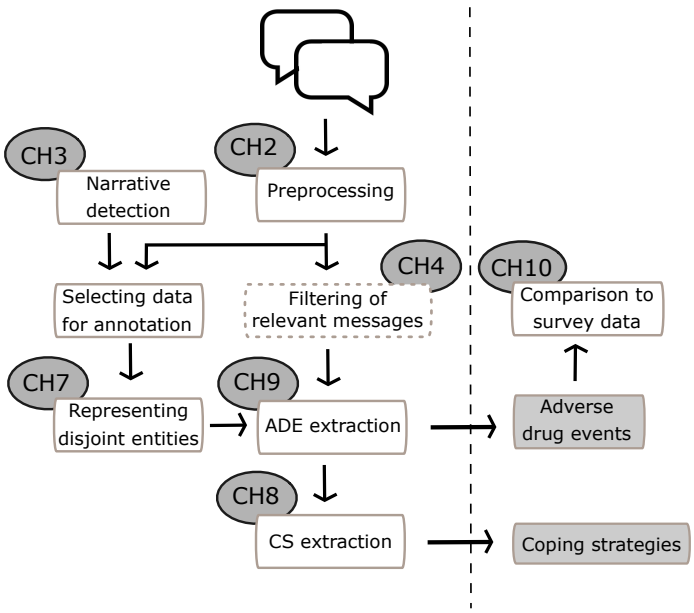


Figure 12.1: An overview of how the thesis chapters interlink and contribute to the extraction and analysis of ADEs and coping strategies. The dotted component ("Filtering of relevant messages") was not used in the overall pipeline. The output data is indicated in gray boxes to the right of the striped line. Chapter 5, 6 and 10 are excluded from this overview because they do not directly interlink with the other chapters.

The comparison of ADE prevalence from these two sources of patient-reported outcomes is challenging, because forum data does not allow us to infer who does *not* have an ADE. Users that do not report an ADE might still experience it. Surveys do offer this information by asking closed questions to respondents. Thus, prevalence rates of ADEs from the two sources cannot be compared directly because those from forum data are only interpretable in a relative sense (i.e., nausea is reported more than fatigue). Our work was therefore limited to comparing the top 10 most prevalent ADEs from each data source. Surveys would need to be conducted amongst forum users instead of the general patient population to compare prevalence in more detail.

10. To what extent are the GIST patients active on patient forums representative for the GIST population and which sample biases does this data source suffer from?

In Chapter 11, we investigate sample bias in patient forum data through a population-based survey amongst Dutch GIST patients. We find that the majority of survey respondents (82%) do not have contact with other patients via social media. This does not necessarily pose a problem as our key question is whether forum users are representative for the general population. It is important to know to what extent forum users are representative to be able to use forum data as a complementary resource for patient-reported outcomes such as adverse drug responses. Our results show that patients that use social media to contact other patients have a strong preference for disease-specific patient forums. This supports the notion that patient forums are the preferable digital resource for patient-reported outcomes despite most research in the field focusing on general social media.

We find that forum users report a lower level of social functioning and the odds of being on a patient forum are higher for patients that are monitored (2.8 times), that are on curative treatment (1.9 times) or that are palliative (10 times) than the odds for patients that are considered cured. Post-hoc analysis shows that overall GIST patients that are in relatively worse condition in terms of symptom burden and quality of life and that are on medication, especially third- or fourth-line medication, appear over-represented. Although it is vital to interpret results with these biases in mind, it is equally vital to promote awareness that sample bias is by no means unique to forum data but inherent to any source of patient-reported outcomes.

In this chapter, we studied a specific patient population in a single country that has a rare disorder characterized by a long palliative phase. It is an open question to what extent our results are generalizable, yet this is a first stepping stone in response to the strong voice of concern about sample bias of health-related social media [13, 23, 32, 58, 276]. Although we do no find significant non-responder bias, our underlying assumption that the survey respondents are representative for the general GIST population is another limitation of our work.

12.2. ANSWER TO MAIN RESEARCH QUESTION

To what extent can automated extraction of experiential knowledge from patient forum posts aid knowledge discovery to yield hypotheses for clinical research?

In this thesis, we collected experiential knowledge from forums centered around certain patient communities (i.e., disease-specific forums). We focused on two types of experiential knowledge, namely adverse drug events and coping strategies for adverse drug events. Nonetheless, patients also share other experiences on online discussion groups that have the potential to lead to knowledge discovery. These include experiences with their diagnostic process, experiences relaying how they cope emotionally and psychologically with having the disease, and advice on day-to-day coping (e.g., with work or childcare).

Of the two types of experiences we focused on, the extraction of adverse drug events is the easier task. For this task, benchmarks, state-of-the-art algorithms, relevant ontologies, and external data sets were available. Moreover, the search space is clearly delineated by possible symptoms registered in medical ontologies. Adverse drug events can be extracted from patient forum messages with sufficient success to enable the discovery of novel ADEs, long-term ADEs, and a ranking of which ADEs are most important to patients as shown by how often they are reported. This ranking differs notably from the known prevalence from clinical trials, although it mirrors the side effect profile. Although the relative importance can inform where healthcare may have the most impact, novel and long-term ADEs can yield new hypotheses for future research (Chapter 9). Moreover, they can also be used to keep patient-reported outcome measures up to date (Chapter 10).

The extraction of coping strategies is more challenging because the task is novel; resources are lacking and the search space requires delineation. At present, the quality of models for extraction is poor (Chapter 8), yet the potential for knowledge discovery is substantial, as coping strategies for adverse drug events have not been studied previously. Aside from empowering patients directly, the discovery of coping strategies can yield hypotheses on why these strategies are effective. However, the poor performance of automatic extraction may necessitate additional manual qualitative checks of the relevant forum messages.

Whether the extracted experiential knowledge can aid knowledge discovery is contingent on a number of conditions. First, the source data need to be representative of the patient population or at least, the sample bias must be sufficiently understood to allow for bias mitigation. Our results indicate that in our main use case (i.e., the GIST patient forum), patients in certain treatment phases (i.e., on curative treatment, in follow-up, and on palliative treatment) and in relatively worse condition were over-represented compared to patients that are considered cured and doing relatively well (Chapter 11). Second, the models underpinning extraction need to be able to deal with zero-shot cases and be sufficiently robust to variation in the input data. On both accounts, state-of-the-art models do not perform well (Chapter 6). Third, these models also need to be able to deal with the conversational nature of and the noise inherent to medical social media (Chapter 4, Chapter 2 and 7).

In our work, we found that semi-automated knowledge discovery is preferable to fully automated knowledge discovery from patient forums. In Chapter 9, we saw that it

was necessary for a medical professional to manually improve the filtering of ADEs with those from clinical trials. Here, medical knowledge of which causal ADE might result in other consequential ADE was essential. Without additional filtering, our results may be dismissed as not truly novel by other medical professionals. In 8, we saw that coping strategies can be extracted automatically, but for knowledge discovery to occur a domain expert needs to filter the findings (i.e., remove false positives) and inspect the underlying messages.

Yet, the complementary value of knowledge discovery from patient experiences is partly in its undirected nature; It is most beneficial if no hypotheses or paradigms guide and restrict the open-ended knowledge extraction. Although neither medical nor patient perspectives should influence knowledge discovery, they are important when interpreting findings and determining which are to be researched further. For instance, findings should be placed in an academic medical context and priority may be given to those findings that patients value most. Extracting additional information about the extracted patient-reported experiences such as the severity of the ADE or the dosage that led to the ADE would also be helpful to this process.

To be able to place any findings in context, it may be instrumental to obtain clinical information about the posters (e.g., their comorbid conditions or duration of disease). This could be done by linking to additional information sources such as the Netherlands Cancer Registry (NCR); by holding surveys amongst users or possibly by extracting these features automatically from posts. Both the technical feasibility of the latter as well as how often patients actually mention clinical features in their posts still needs to be explored. Moreover, to do so, it is essential that different posts from a user can be linked. This would have the additional benefit that it becomes possible to distinguish between one user mentioning an outcome (e.g., an ADE) multiple times from multiple users mentioning an outcome. It also would enable longitudinal knowledge discovery. In our studies based on Facebook data, it was not possible to link different posts from a single user while protecting their privacy in line with the GDPR. We will elaborate on possible GDPR-compliant alternatives in Section 12.4.2.

12.3. FUTURE RESEARCH

In this section, we will propose ideas for future research divided into three broad topics. In Section 12.3.1, we discuss future work into mining experiential knowledge from social media, including improved and more reliable mining of ADEs. In Section 12.3.2, we delve into recommended future directions for a more standardized and interpretable mapping of extracted ADEs to a medical ontology. In Section 12.3.3, we introduce overarching ideas for improved knowledge extraction from noisy real-world data of which patient forum messages are one example.

12.3.1. MINING EXPERIENTIAL KNOWLEDGE FROM SOCIAL MEDIA

The main use case of social media mining for health has been adverse drug events for pharmacovigilance. To increase the relevance of ADE mining for pharmacovigilance, future work could investigate mining the dosage of medication, the severity of ADEs and details of the impact on daily life. Adverse responses to surgical interventions or

withdrawal of medication could also be a worthwhile avenue for future work. Moreover, sourcing ADEs from a large variety of languages would expand coverage. Currently, adverse drug event detection has already been developed to some extent for Russian [193, 303], Spanish [271], French [13] and Chinese [350]. It would also be valuable to explore how the complementary value of ADE mining from patient forums for pharmacovigilance may differ between different types of disorders, such as between common and rare disorders but also between chronic and more acute disorders. We expect that patients with rare or chronic disorders will share more experiences on disease-specific forums based on prior qualitative work but this is still an open question. Expanding our end-to-end work to other disorders than GIST would have the additional benefit of further refining our methodology. In selecting which disorders to examine, we recommend considering disorders with a large efficacy-effective gap i.e., large differences in outcomes measured in randomized clinical trial (RCT) and those observed in real-world evidence. Previous studies have demonstrated such a gap for schizophrenia [294] and for systemic cancer treatment [234].

Moreover, to integrate ADE detection from disease-specific patient forums into healthcare in the long run, future research into the limitations of machine learning pipelines is important. First, a further understanding of biases in the data is necessary for accurate interpretation of evidence for ADEs. We recommend expanding upon our work on sample bias and activity bias in Chapter 11. Research into mitigation strategies would also be beneficial. Second, in line with our work in Chapter 6, we believe further research into the vulnerabilities and biases of our models is also necessary to make them more robust. For instance, it is an open question to what extent end-to-end detection may over- or under-represent certain classes of ADEs. We expect that BERT models may find some classes easier to identify than others, which would skew the relative ADE frequencies. On a similar note, we recommend researching uncertainty estimation methods in order to visualize error propagation in end-to-end ADE detection systems. This would also allow researchers to be more transparent towards laymen and medical professionals.

Future research could also move towards mining other experiential knowledge such as coping strategies. Aside from building on our work on coping strategies in Chapter 8, we would recommend investigating psychological coping or coping with the disease in daily life situations e.g., work and childcare. More work into open-ended mining of experiential knowledge would also be useful to gain an understanding of what might be gained from online experiential knowledge that is shared between patients. To date, the work on mining of patient narratives in general has been limited. Our work in Chapter 2 provides a starting point for work in this direction.

12.3.2. ONTOLOGY MAPPING AND INTEROPERABILITY

The recent recognition that models for ADE extraction need to be able to handle zero-shot cases [193, 304] is a promising development. In other words, models need to be able to recognize all possible ADEs including ones that they have not been trained on. Essentially, this means a search space of all possible ADEs must be predefined. Currently, this is done by including all ADEs from an ontology as possible target classes that the model can normalize an ADE to [193, 291].[4] Nevertheless, more emphasis on evaluating zero-shot

[4]Normalization is generally operationalized as a classification task

performance separately is still necessary as well as more research into and agreement on the use of medical ontologies.

First, the choice of an ontology is not trivial, as commonly used ontologies are not always inter-operable. Positive recent developments in this regard are the release of maps between the two major ontologies SNOMED-CT and MedDRA in April 2021 as part of the WEB-RADR 2 project [334] and the creation of the OMOP vocabulary as part of the Observational Health Data Sciences and Informatics project [222] (OHDSI). The goal of OMOP is to enable consistent content across varied observational resources. At present, it does not include social media data. In our work, we opted to be as inter-operable with OMOP as possible by choosing SNOMED-CT over MedDRA. Future work could build upon these movements to create consistent guidelines and develop maps to ontologies commonly used in clinical trials like the CTCAE.

Agreement on a standard ontology for annotation of data is also necessary to facilitate progress in the field. To date, some data sets are annotated with MedDRA (SMM4H data [193]), some with SNOMED-CT (PsyTAR [353], COMETA [20]) and some with both (CADEC [151]). The exact concept that is chosen for a certain ADE can also differ between data sets and both guidelines for future data as well as work on aligning current data sets is called for. Some of these differences in choice arise from noise inherent to the ontologies: multiple concept identifiers are possible for the same ADE. In our work in Chapter 9, we dealt with this challenge by mapping concepts from external training data sets to synonymous concepts in our selected SNOMED-CT subset. We checked for a direct mapping in the community-based BioPortal [220] and we mapped concepts to their parent concepts if the parent concept was included in our subset (e.g., "moderate anxiety" to "anxiety").

Second, future work should research to what extent the target classes for normalization can be reduced to improve performance while maintaining sufficient detail. Medical ontologies are generally very large: SNOMED-CT contains 361,555 concepts of which 119,020 are in the Clinical Findings category and MedDRA contains around 79,000 lower level term concepts (LLTs). In our collaborative work with Magge et al. [194], we opted to use the preferred terms (PTs) of MedDRA (approx. 23,000) instead of the lower level terms (LLTs) to reduce target classes. In our work in Chapter 9, we restricted target classes to the CORE Problem List subset of SNOMED-CT[5] (5,813 concepts), which is a curated subset designed to maximize interoperability. We did not compile our own corpus for normalization but relied upon existing public data, which we mapped to the CORE subset. Any data that could not be mapped was disregarded. We chose to add five additional concepts to the CORE subset(e.g., hair color change, and hand-foot syndrome), because they were known ADEs for our drugs of interest but were not included. Thus, it appears the CORE subset is also not optimal for detecting ADEs, and future work should consider how this subset can be refined.

Third, aggregation of the detected ADEs into larger categories is desirable but not trivial. In our work in Chapter 9, the involved medical professional dr. Gelderblom indicated that closely related concepts from the CORE subset like depression and mild depression should be grouped for interpretation. Such situations can arise when concepts

[5] https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html

of different hierarchical levels are included as target classes.[6] When trying to determine whether ADEs are novel by filtering with ADEs from other data sources (e.g., clinical trials), aggregation is also important. If ADEs are not aggregated, a subcategory of "depression" such as "mild depression" would falsely be considered novel. We opted to aggregate based on the SNOMED-CT hierarchy: child concepts from at least 5 levels of depth[7] were aggregated to their parent concept if the parent concept was part of the CORE subset. Chaining was allowed meaning that a concept could be aggregated until the parent concept no longer was part of the CORE or the minimum depth (5) of the SNOMED hierarchy was reached. Aggregation was further manually fine-tuned with expert knowledge from dr. Gelderblom. We also considered using System Order Classes (SOC) to aggregate ADEs[8] but these were not deemed informative as they were too general. For transparency, the ADEs that were originally detected are included as footnotes in the data visualization (see Chapter 9). Additional research into how to best perform (hierarchical) aggregation of detected ADEs is required before end-to-end systems for ADE detection from real-world evidence can be integrated in healthcare. As of yet, this challenge has been overlooked.

12.3.3. DEALING WITH REAL-WORLD DATA

There have been various developments in the NLP field towards dealing with real-world noisy data, such as work on zero-shot methods to handle large label spaces without needing training data for each label. However, in other directions the work is still limited. We believe that some promising directions are: research into extracting complex entities; domain adaptation; robustness to noisy, user-generated data; and improved computational efficiency of models to realize real-world applications.

A first interesting avenue to explore is the extraction of complex entities that are often fuzzy in nature, of which coping strategies are but one example. Unlike named entities, these entities are often long, are not proper nouns, and may contain non-entity words (i.e., are discontinuous). Therefore, complex entities may require different approaches than named entities. We found, for example, that NER of coping strategies benefits from adding a window of one token to each entity and from adding additional entity types that are related but may be easier to identify (in our case: ADE). Possible other directions could be developing methods that integrate expert knowledge (e.g., from a medical professional) or that include a human-in-the-loop. A major obstacle for end-to-end extraction of complex entities is error propagation, as the initial extraction can form a bottleneck for subsequent entity linking or disambiguation. In this regard, useful directions to pursue are multi-task learning to leverage information from other entities or the entity linking task; a stronger focus on external validation while developing methods for extraction and conceptualizing the task as a single step, e.g., we conceptualized coping strategy extraction as extreme multi-label classification which outperformed two-step NER with entity linking (see Chapter 8). Often, these fuzzier entities are not included in benchmarks for core NLP tasks, and resources are lacking to aid their extraction. Adverse drug effects are not a good example in this respect, as there are already benchmarks, public data sets,

[6] In the CORE subset, we found that concepts range from 1 to 10 levels of depth
[7] This depth was chosen to prevent ADEs becoming too vague or general
[8] SNOMED-CT is inter-operable with SOC through the OMOP vocabulary

and relevant ontologies available. Another interesting avenue for research that does not rely on such resources is open information extraction (IE), i.e., the extraction of relation tuples from plain text without needing to specify a schema in advance. Open IE bypasses the need to delineate a search space or ontology for complex entities that may be highly variable, such as coping strategies. Besides avoiding this challenging step, delineating a search space restricts detection to only those concepts included in the search space.

Domain adaptation is a second interesting avenue for future research to improve our ability to deal with real-world data. Real-world data is often small or big data may exist but may not be available for other reasons such as privacy restrictions. Disease-specific patient forums are regularly an example of the former and electronic health records are an example of the latter. Consequently, transfer learning models pretrained on the right domain may not exist, because large amounts of (unlabeled) data are necessary to pretrain them. If sufficient labeled data exists, however, transfer learning models pretrained on other comparable domains can be fine-tuned for the task at hand. Yet, prior work has shown that for the biomedical domain using a domain-specific vocabulary improves model performance significantly (SciBERT [28] and PubmedBERT [119]). Here, the work by Hong et al. [135] is worth noting: They consider the vocabulary of the BERT model as optimizable instead of static and propose a method to update the vocabulary with domain-specific terms during fine-tuning. Hong et al. [135] find consistent performance improvements on diverse domains. We believe building upon their work on domain adaptation during fine-tuning is a worthwhile direction to explore.

As it is often also difficult to obtain sufficient labeled training data, *unsupervised* domain adaptation is another relevant research direction. Unsupervised domain adaptation encompasses methods that aim to attain good performance in a target domain by relying on labeled data from another domain (called the source domain). For instance, Ma et al. [191] use a combination of curriculum learning and domain-discriminate data selection, Ryu and Lee [256] combine adversarial adaptation with knowledge distillation and more recently, Zhang et al. [349] develop a cross domain method that does not require access to the source data but relies purely on the discrepancy in distribution between source model and target data for domain adaptation, which may be beneficial for privacy-sensitive data.

A third promising research direction is research into increasing the robustness of state-of-the-art extraction models to noisy, user-generated data. Prior work by Kumar et al. [166] found that fine-tuning a BERT model (trained on clean, curated data) with noisy user-generated data led to a drop in performance. The performance appears to degrade because the wordpiece tokenizer breaks up misspelt words into sub-words as it does not include these misspelt (sub-)words in its vocabulary [166]. One possible approach to this problem is domain adaptation. Another proposed approach has been lexical normalization [97]. Our work in Chapter 2 is an example of this approach. In our opinion, to date, normalization and preprocessing in general has received insufficient attention from the medical NLP community, despite the importance of the quality of training data to the success of a model. We consider developing methods to train models using noisy data as a third possible approach. It would be worthwhile to investigate whether and how noise can be added during pretraining or fine-tuning to increase instead of degrade performance.

A fourth avenue of promising research for utilizing real-world data is research into improved computational efficiency of transfer learning models, which is important for deploying applications. The distillation of models is one option. An example is distilBERT, a distilled version of BERT, that retains 97% of performance with only half the parameters Sanh et al. [260]. We employed distilBERT in Chapter 4 and 7. Other recent developments have been made on more efficient pretraining methods, such those underlying ELECTRA [66] and the biomedical BioELECTRA [150]. ELECTRA uses replaced token detection as a pre-training task: the model is trained to distinguish between "real" and "fake" input. Instead of replacing tokens with [MASK] as done in BERT, the input is corrupted by replacing the input tokens with fakes generated by a generator model. In addition, less complex methods that do not rely on deep learning architectures like SVM are less computationally heavy. We recommend investigating under which conditions such methods may offer better or comparable performance to more complex transfer learning methods.

12.4. RECOMMENDATIONS

In this section, we first present general recommendations concerning knowledge discovery from social media regarding acceptance of social media as a valuable source of complementary knowledge by medical professionals (Section 12.4.1). Hereafter, we discuss our recommendations for ensuring privacy of patients and consequent possibilities for data re-use (Section 12.4.2); for developing annotation guidelines (Section 12.4.3) and for long-term integration of experiential knowledge from social media into healthcare (Section 12.4.4).

12.4.1. KNOWLEDGE DISCOVERY FROM SOCIAL MEDIA

Medical professionals often question the reliability of experiential knowledge on social media. For instance, they note that it is possible for patients to falsely attribute symptoms to their medication, provide false information deliberately, or, in the case of coping strategies, experience a placebo effect. Consequently, medical professionals are reluctant to accept social media as a source of valuable knowledge.

To mitigate this concern we have three recommendations. Our first recommendation is to continue to validate the reliability of adverse drug event reports from patient forums by assessing overlap with more traditional sources, such as spontaneous reports from medical professionals, survey results and medical literature, as well as by assessing to what extent clinicians recognize the reported adverse drug responses from the clinic. To date, prior work has shown ADE reports sourced from patient forums to be of similar quality to those of medical professionals [37, 322]; to have high overlap with traditional data sources and to contain novel ADEs [30, 346]. In Chapter 9 and 10, we underscore these findings with our own case study of a forum for GIST patients.

In contrast, during this PhD, a large EU project [321] found that ADE reports from social media, including patient forums, have no additional value on top of official post-marketing systems. Although we applaud such large-scale efforts to assess the value of social media for pharmacovigilance, we recommend a large-scale follow-up project that involves computer science researchers instead of commercial parties. In the previous

project, the automatic extraction of ADEs was done by commercial partners who made use of proprietary software based on out-dated methods. We agree with van Stekelenborg et al. [321] that the capability to extract ADEs is key to determining the true value of ADE reports on social media, and thus we recommend a follow-up project in which state-of-the-art methods are used. It is also important that these methods are open-source to provide transparency and allow the community to build upon their work.

Our second recommendation is to limit the use of experiential knowledge to knowledge discovery and clarifying appropriate and inappropriate use cases. To gather support in the medical domain, it is important to emphasize but not overstate the value of experiential knowledge. Experiential knowledge can offer an collective patient perspective through "wisdom of the crowds", but is not appropriate for personalized medicine. We recommend explaining both the benefits, such as reduced patient burden and uncensored reports, and the downsides, such as imperfect performance and noise, of using AI for automatic extraction. We believe that a further demystification of AI is important in the long run to give medical professionals agency in this discussion and facilitate constructive integration of experiential knowledge from social media into healthcare.

Our third recommendation is to consider all experiential knowledge as equally valid, i.e., not considering any as misinformation. Defining some of the shared experiential knowledge as misinformation would clash with open-ended knowledge discovery. Misinformation detection methods rely on a ground truth (often after the fact), which per definition is not available for novel findings. Thus, misinformation detection will not be able to differentiate between novel information and misinformation. In addition, we find it ill-advised to brand the experience of one patient as less true than that of another. They may be wrong in their conviction (e.g., that their headache is an ADE of the drug or that gemstones help them), but that does not make their experience any less real to them. Third, experiential knowledge does not produce the truth, but hypotheses. Thus, misinformation detection is not relevant as this relates to the truth value of statements.

12.4.2. PRIVACY AND ADOPTING FAIR METADATA STANDARDS

In our work, we tried to adhere to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability[9]) which aim to increase data reuse. Although we were unable to share the forum data itself for reuse under the rules of the General Data Protection Regulation (GDPR), our methods could be employed on public forums to improve the findability of forum messages by adding entities (a type of rich meta-data) (principle F2 of Findability: Data are described with rich metadata). In turn, this can improve accessibility of the meta-data (principle A2 of Accessibility: Metadata are accessible, even when the data are no longer available). We adhere to the principles of interoperability by choosing ontologies that are interoperable with the OMOP vocabulary. This vocabulary stems from the OHDSI project, which aims for more interoperability between divergent observational data sources (see Section 12.3.2 for more details). In developing our own ontology for coping strategies, we also sourced as many concepts as possible from existing ontologies (SNOMED-CT, NCIT, PACO, and RxNORM) favoring those used by the OMOP vocabulary.

[9] Available at `https://www.go-fair.org/fair-principles/`

We could have made our data more reusable and FAIR if we would have been able to share it. This was not possible because the initially public forum became private in 2021. Under the GDPR, we were then unable to share the data. To prevent similar situations in future projects, we recommend setting up forums in collaboration with patient organizations so that the ownership of the data rests with patients instead of commercial parties (Facebook in our case). Users should be asked for consent for using the data for research purposes prior to participation in these forums. Such a setup would have the additional benefit that users could be asked for personal characteristics. In our experience, medical researchers find obtaining personal information of forum users vital to the interpretation of ADE reports. In such a collaborative setup, researchers could communicate directly to the patients about research output and patients themselves can be given insight through a tool or dashboard. The Patient Forum Miner (PFM) project [76] offers a great starting point.

Alternative valuable sources of data that we recommend exploring are forums on platforms such as PatientsLikeMe[10]. These platforms often ask patients for their consent for using data for research purposes when they make an account. In this project, we tried to set up a collaboration with PatientsLikeMe to no avail yet we recommend exploring collaborations with comparable parties that may find this idea more agreeable. An advantage of this approach is that such platforms contain forums for various conditions, while a disadvantage is that these forums are often less active than forums that are administered by a patient organization. These platforms have been known to agree to collaborate with universities, but not with individual researchers, so we recommend involving faculty management in future efforts.

12.4.3. DEVELOPING ANNOTATION GUIDELINES

We encountered a number of overarching challenges when developing annotation guidelines[11]. The first challenge was that messages from forum discussions may be difficult to interpret or be interpreted differently without the context of the conversation. We therefore recommend providing annotators with the context of the message (i.e., the messages preceding it). This can be done in a number of different ways. For annotation of named entities, annotators labeled whole discussion threads, one message at a time. For annotation of ADE-CS relations, six messages prior to the message containing the CS were provided in a single view. All variants or co-referents of the correct ADE in these (at maximum) seven messages were labeled as positive. The size of this conversational window was largely arbitrary, although chosen to be relatively wide, and we recommend careful consideration of an appropriate window size in future work. For entity linking, we did not provide annotators with the conversational context, because this was not accommodated by our annotation tool and the task was already complex.

A second challenge was determining who to select as annotators. For NER, we asked GIST patients to volunteer. Although their domain expertise was an advantage, they found the annotation task challenging and did not have sufficient time. Moreover, one annotator dropped out because they did not master the English language sufficiently. Therefore, for

[10] `https://www.patientslikeme.com/`

[11] Annotation guidelines can be found at: `https://github.com/AnneDirkson/CopingStratExtract/tree/main/annotation`

the annotation of CS-ADE relations and CS normalization, we recruited master students. For the latter task, we paid our annotators because the labeling task required a high level of dedication and time.

A third overarching challenge was deciding how to handle data that was previously labeled incorrectly for NER during labeling for entity linking (that relied on labeled CS entities) or relation extraction (that relied on both labeled CS and ADE entities). For relation extraction, we decided to not correct boundaries of entities or missed entities. This does have the consequence that there may be cases where the coping strategy cannot be linked to an ADE because the ADE has not been annotated correctly. For incorrectly labeled coping strategies, no relation can be determined so they were excluded indirectly. For entity linking, false positives were labeled as with a separate label (NOT_A_STRATEGY) as it was not possible to normalize them. Messages that contained false negatives were already excluded in the pre-selection of messages with CS. There were also cases where two coping strategies were included as a single entity (e.g. "drink water and exercise"). We instructed annotators to relabel these strategies as separate entities as our annotation tool did not allow one entity to have multiple labels. Although there is not a single correct solution to handling incorrect prior annotations, we recommend considering how annotators should handle such data explicitly in the annotation guide.

For annotation of named entities specifically, we recommend providing both positive and negative examples to illustrate definitions, e.g. the definition of what constitutes an ADE. We also recommend noting how annotators should deal with disjoint entities as these are common in the biomedical domain. We recommend a continuous annotation of disjoint entities (see the FuzzyBIO representation in Chapter 7). Annotators also find it difficult to determine the boundaries of entities, especially for complex entities. We recommend taking this into account when evaluating annotator agreement and including instructions on bounding entities in the annotation guideline. Although it is not possible to flesh out all possible cases, common cases can be streamlined (e.g. does one include the definite article?). Moreover, we recommend considering possible future layers of annotation on the same data, e.g. entity linking, when deciding upon an annotation tool. We had to switch from Doccano to Inception to accommodate entity linking of coping strategies whereas NER would have also been possible in Inception.

For the annotation of entity linking, we would additionally recommend fellow researchers to develop rules for the multi-labeling of entities, as there may be entities for which there is not an exact label in the ontology but a combination of two labels would suffice (e.g. ginger toothpaste). Allowing for multi-labeling prevents the ontology size from growing exponentially.

For the annotation of relations, a major challenge was selecting an appropriate annotation tool. We chose to conceptualize this task as a classification task and use Doccano. The biggest drawback of our approach was the transformation of the data into an appropriate format. We elected to automatically create sentences where some entities were masked so that annotators could select the cases where the masked entity was indeed the correct one. However this proved challenging and thus we recommend researching whether there are more suitable options available for future work. We also recommend considering whether annotators should label only the exact entity that has a relation to the entity at hand or also its co-referents. We decided to annotate all co-referents of the

correct ADE because it was sometimes difficult for annotators to select a single correct mention amongst multiple mentions of the same ADE. Moreover, as long as the correct ADE was selected by the model, it did not matter for our task whether it was the exact correct mention of the ADE.

12.4.4. LONG-TERM INTEGRATION INTO HEALTHCARE

To attain the long-term goal of integrating online patient-reported experiences from social media into healthcare, an appropriate regulatory framework will need to be developed. In the context of pharmacovigilance, various researchers have already advocated for a regulatory legal and policy framework [176, 228]. Regulatory recommendations specifically for updating pharmacovigilance guidance were put forward by Brosch et al. [42] in the context of the WEB-RADR 2 project. According to Brosch et al. [42], key challenges include limited follow-up options for social media data; the large volume of social media data that requires more resources to manage properly; and a mismatch between what is possible on social media and current minimal criteria for a valid ADE report. However, most pharmaceutical companies believe their regulatory framework can be adapted to include social media: 71% considers social media a possible tool from a legislative and industry perspective [227]. We recommend continuing these efforts to adapt the current regulatory framework for pharmacovigilance. However, we also urge legal and policy experts to develop a larger regulatory framework for incorporating other patient-reported experiences into the healthcare system.

Aside from a regulatory framework, we also need the involvement of medical professionals to enact change in the long run. Supportive medical professionals are indispensable in determining how patient-reported experiences can best be incorporated into healthcare and advocating for the value of experiential knowledge to their colleagues. As mentioned in Section 12.4.1, we believe that medical professionals should be taught about AI to give them agency in the discussion on how to use AI in healthcare and fuel constructive debates on this topic. The same goes for patient representatives whose insights and involvement can aid decisions on which patient-reported experiences are most beneficial for healthcare and should be prioritized. A rudimentary understanding of AI will be helpful to generate more understanding of the challenges inherent to automated analysis and the slow speed at which text mining algorithms can be developed.

In the Netherlands, there have been two recent developments of interest regarding the educating of medical professionals on AI and increasing their level of trust. The Dutch Ministry of Healthcare has presented a guideline [278] on the use of predictive AI in healthcare to increase trust amongst medical professionals. This includes amongst others: transparency about possible negative consequences, thorough external validation and evaluation of the added value of the predictive algorithm for healthcare. This guideline is accompanied by an online educational course[12]. Another online course on the use of AI in healthcare called "Nationale AI-Zorg"[13] was developed by the NL AI Coalitie (a public-private coalition of Dutch AI organisations).

Overall, we recommend starting with the integration of patient-reported experiences into healthcare for rare disorders specifically before moving on to more common

[12] Available at: `https://www.leidraad-ai.nl/`
[13] Available at: `https://zorg.ai-cursus.nl/home`

disorders. Patients with rare disorders have shown an extraordinarily high level of "citizen science" through mobilization into grassroots movements that aggregate their own data in an effort to help other patients and to influence the research agenda [49, 108, 237]. They display a clear desire to translate their experiential knowledge into actionable data. Online forums of patients with rare disorders are also relatively active which increases the number of patient-reported experiences. Finally, the potential benefits of patient-generated online data are high for this subgroup due to a scarcity of research for rare disorders.

REFERENCES

[1] N. K. Aaronson, S. Ahmedzai, B. Bergman, M. Bullinger, A. Cull, N. J. Duez, A. Filiberti, H. Flechtner, S. B. Fleishman, J. C. Haes, S. Kaasa, M. Klee, D. Osoba, D. Razavi, P. B. Rofe, S. Schraub, K. Sneeuw, M. Sullivan, and F. Takeda. The European organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85(5):365–376, 1993. ISSN 00278874. doi: 10.1093/jnci/85.5.365.

[2] ACL rolling review. Guidelines for Answering Checklist Questions, 2021. URL `https://aclrollingreview.org/responsibleNLPresearch/`.

[3] A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

[4] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[5] A. Al Hamid, M. Ghaleb, H. Aljadhey, and Z. Aslanpour. A systematic review of qualitative research on the contributory factors leading to medicine-related problems from the perspectives of adult patients with cardiovascular diseases and diabetes mellitus. *BMJ open*, 4(9):e005992, sep 2014. ISSN 2044-6055. doi: 10.1136/bmjopen-2014-005992.

[6] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909.

[7] B. Alshemali and J. Kalita. Improving the reliability of deep neural networks in NLP: A review. *Knowledge-Based Systems*, 10520, 2019. doi: 10.1016/j.knosys.2019.105210.

[8] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. B. Srivastava, and K.-W. Chang. Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896. Association for Computational Linguistics, 2018.

[9] C. Anderson, J. Krska, E. Murphy, and A. Avery. The importance of direct patient reporting of suspected adverse drug reactions: A patient perspective. *British Journal of Clinical Pharmacology*, 72(5):806–822, 2011. ISSN 03065251. doi: 10.1111/j.1365-2125.2011.03990.x.

[10] M. Anderson and A. Smith. Social Media Use in 2021. Technical Report April, Pew Research Center, 2021. URL `https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/`.

[11] V. Araujo, A. Carvallo, and D. Parra. Adversarial evaluation of bert for biomedical named entity recognition. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 79–82, 2020.

[12] T. M. Atkinson, S. J. Ryan, A. V. Bennett, A. M. Stover, R. M. Saracino, L. J. Rogak, S. T. Jewell, K. Matsoukas, Y. Li, and E. Basch. The association between clinician-based common terminology criteria for adverse events (CTCAE) and patient-reported outcomes (PRO): a systematic review. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*, 24(8):3669–3676, 2016. ISSN 1433-7339. doi: 10.1007/S00520-016-3297-9.

[13] B. Audeh, F. Bellet, M. N. Beyens, A. Lillo-Le Louët, and C. Bousquet. Use of Social Media for Pharmacovigilance Activities: Key Findings and Recommendations from the Vigi4Med Project. *Drug Safety*, 43(9):835–851, 2020. ISSN 11791942. doi: 10.1007/s40264-020-00951-2.

[14] J. L. Austin. *How to do things with words*. Oxford university press, 1962.

[15] S. Aymé, A. Kole, and S. Groft. Empowerment of patients: lessons from the rare diseases community. *The Lancet*, 371(9629):2048–2051, 2008. ISSN 01406736. doi: 10.1016/S0140-6736(08)60875-2.

[16] R. Baeza-Yates. Bias on the web. *Communications of the ACM*, 2018. ISSN 00010782. doi: 10.1145/3209581.

[17] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *ArXiv*, 2016. ISSN 16130073.

[18] S. Balasubramanian, N. Jain, G. Jindal, A. Awasthi, and S. Sarawagi. What's in a Name? Are BERT Named Entity Representations just as Good for any other Name? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 205–214. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.repl4nlp-1.24.

[19] T. Baldwin, M.-C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, and W. Xu. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135. Association for Computational Linguistics, 2015. doi: 10.18653/v1/W15-4319.

[20] M. Basaldella, F. Liu, E. Shareghi, and N. Collier. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.253.

[21] E. Basch, X. Jia, G. Heller, A. Barz, L. Sit, M. Fruscione, M. Appawu, A. Iasonos, T. Atkinson, S. Goldfarb, A. Culkin, M. G. Kris, and D. Schrag. Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes. *Journal of the National Cancer Institute*, 101(23):1624–1632, 2009. doi: 10.1093/JNCI/DJP386.

[22] E. Basch, B. B. Reeve, S. A. Mitchell, S. B. Clauser, L. M. Minasian, A. C. Dueck, T. R. Mendoza, J. Hay, T. M. Atkinson, A. P. Abernethy, D. W. Bruner, C. S. Cleeland, J. A. Sloan, R. Chilukuri, P. Baumgartner, A. Denicoff, D. St. Germain, A. M. O'Mara, A. Chen, J. Kelaghan, A. V. Bennett, L. Sit, L. Rogak, A. Barz, D. B. Paul, and D. Schrag. Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *Journal of the National Cancer Institute*, 106(9), 2014. doi: 10.1093/JNCI/DJU244.

[23] A. Bate, R. F. Reynolds, and P. Caubel. The hope, hype and reality of Big Data for pharmacovigilance. *Therapeutic advances in drug safety*, 9(1):5–11, 2018. ISSN 2042-0986. doi: 10.1177/2042098617736422.

[24] R. Beckley. Bekli: A simple approach to twitter text normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 82–86, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4312.

[25] M. Beeksma, S. Verberne, A. van den Bosch, I. Hendrickx, E. Das, and S. Groenewoud. Predicting life expectancy with a recurrent neural network. *BMC Medical Informatics and Decision Making*, 19(36), 2019. doi: 10.1186/s12911-019-0775-2.

[26] Y. Belinkov and Y. Bisk. Synthetic and Natural Noise Both Break Neural Machine Translation. *ArXiv*, abs/1711.0, 2017. URL `http://arxiv.org/abs/1711.02173`.

[27] M. Belousov, W. Dixon, and G. Nenadic. Using an Ensemble of Generalised Linear and Deep Learning Models in the SMM4H 2017 Medical Concept Normalisation Task. In *Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H)*, 2017.

[28] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371.

[29] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 5(1): 289–300, 1995.

[30] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard, and J. H. Holmes. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation HHS Public Access. *J Biomed Inform*, 44(6):989–996, 2011. doi: 10.1016/j.jbi.2011.07.005.

[31] G. Berend and E. Tasnádi. Uszeged: Correction type-sensitive normalization of english tweets using efficiently indexed n-gram statistics. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 120–125, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4318.

[32] J. Berrewaerts, L. Delbecque, P. Orban, and M. Desseilles. Patient Participation and the Use of Ehealth Tools for Pharmacoviligance. *Frontiers in Pharmacology*, 7:90, 2016. ISSN 1663-9812. doi: 10.3389/fphar.2016.00090.

[33] J. Bian and F. Yu. Towards Large-scale Twitter Mining for Drug-related Adverse Events. In *SHB12*, pages 25–32, 2012. doi: 10.1145/2389707.2389713.

[34] G. Blank and C. Lutz. Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, 61(7):741–756, 2017. ISSN 15523381. doi: 10.1177/0002764217717559.

[35] C. D. Blanke, G. D. Demetri, M. von Mehren, M. C. Heinrich, B. Eisenberg, J. A. Fletcher, C. L. Corless, C. D. M. Fletcher, P. J. Roberts, D. Heinz, E. Wehre, Z. Nikolova, and H. Joensuu. Long-term results from a randomized phase II trial of standard- versus higher-dose imatinib mesylate for patients with unresectable or metastatic gastrointestinal stromal tumors expressing KIT. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 26(4):620–5, 2 2008. ISSN 1527-7755. doi: 10.1200/JCO.2007.13.4403.

[36] C. D. Blanke, C. Rankin, G. D. Demetri, C. W. Ryan, M. von Mehren, R. S. Benjamin, A. K. Raymond, V. H. C. Bramwell, L. H. Baker, R. G. Maki, M. Tanaka, J. R. Hecht, M. C. Heinrich, C. D. M. Fletcher, J. J. Crowley, and E. C. Borden. Phase III randomized, intergroup trial assessing imatinib mesylate at two dose levels in patients with unresectable or metastatic gastrointestinal stromal tumors expressing the kit receptor tyrosine kinase: S0033. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 26(4):626–32, 2008. ISSN 1527-7755. doi: 10.1200/JCO.2007.13.4452.

[37] A. Blenkinsopp, P. Wilkie, M. Wang, and P. A. Routledge. Patient reporting of suspected adverse drug reactions: A review of published literature and international experience. *British Journal of Clinical Pharmacology*, 63(2):148–156, 2007. ISSN 03065251. doi: 10.1111/j.1365-2125.2006.02746.x.

[38] T. Borkman. Experiential Knowledge: A New Concept for the Analysis of Self-Help Groups. *Social Service Review*, 50(3):445–456, 1976. ISSN 0037-7961. doi: 10.1086/643401.

[39] A. Bottomley, J. C. Reijneveld, M. Koller, H. Flechtner, K. A. Tomaszewski, E. Greimel, P. A. Ganz, J. Ringash, D. O'Connor, P. G. Kluetz, G. Tafuri, M. Grønvold, C. Snyder, C. Gotay, D. L. Fallowfield, K. Apostolidis, R. Wilson, R. Stephens, H. Schünemann, M. Calvert, B. Holzner, J. Z. Musoro, S. Wheelwright, F. Martinelli, A. C. Dueck, M. Pe, C. Coens, G. Velikova, D. Kuliś, M. J. Taphoorn, A. S. Darlington, I. Lewis, and L. van de Poll-Franse. Current state of quality of life and patient-reported outcomes research. *European journal of cancer (Oxford, England : 1990)*, 121:55–63, 2019. ISSN 1879-0852. doi: 10.1016/J.EJCA.2019.08.016.

[40] C. Bousquet, B. Dahamna, S. Guillemin-Lanne, S. J. Darmoni, C. Faviez, C. Huot, S. Katsahian, V. Leroux, S. Pereira, C. Richard, S. Schück, J. Souvignet, A. Lillo-Le Louët, and N. Texier. The Adverse Drug Reactions from Patient Reports in Social Media Project: Five Major Challenges to Overcome to Operationalize Analysis and Efficiently Support Pharmacovigilance Process. *JMIR Research Protocols*, 6(9):e179, 2017. ISSN 1929-0748. doi: 10.2196/resprot.6463.

[41] A. Brandsen, S. Verberne, M. Wansleeben, and K. Lambers. Creating a dataset for named entity recognition in the archaeology domain. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pages 4573–4577, 2020. ISBN 9791095546344.

[42] S. Brosch, A. M. de Ferran, V. Newbould, D. Farkas, M. Lengsavath, and P. Tregunno. Establishing a Framework for the Use of Social Media in Pharmacovigilance in Europe. *Drug Safety*, 42(8):921–930, 2019. ISSN 11791942. doi: 10.1007/s40264-019-00811-8.

[43] A.-M. Bucur, A. Cosma, and L. P. Dinu. Sequence-to-Sequence Lexical Normalization with Multilingual Transformers. *ArXiv*, 2021. URL `https://arxiv.org/abs/2110.02869v3`.

[44] A. Bulcock, L. Hassan, S. Giles, C. Sanders, G. Nenadic, S. Campbell, and W. Dixon. Public Perspectives of Using Social Media Data to Improve Adverse Drug Reaction Reporting: A Mixed-Methods Study. *Drug Safety 2021 44:5*, 44(5):553–564, 2021. ISSN 1179-1942. doi: 10.1007/S40264-021-01042-6.

[45] M. H. Burda, M. Van Den Akker, F. Van Der Horst, P. Lemmens, and J. A. Knottnerus. Collecting and validating experiential expertise is doable but poses methodological challenges. *Journal of Clinical Epidemiology*, 72:10–15, 2016. ISSN 18785921. doi: 10.1016/j.jclinepi.2015.10.021.

[46] G. Burnage, R. Baayen, R. Piepenbrock, and H. van Rijn. *CELEX: A Guide for Users*. Centre for Lexical Information, 1990.

[47] J. Call, C. D. Walentas, J. C. Eickhoff, and N. Scherzer. Survival of gastrointestinal stromal tumor patients in the imatinib era: life raft group observational registry. *BMC Cancer*, 12(1):90, 2012. ISSN 1471-2407. doi: 10.1186/1471-2407-12-90.

[48] E. B. Carbajal-López, D. M. Juárez-García, A. Espinoza-Velazco, and G. Calderillo-Ruiz. Internet-Delivered Cognitive Behavioral Therapy and Psychoeducation

Program for Patients with Gastrointestinal Stromal Tumors. *Journal of cancer education : the official journal of the American Association for Cancer Education*, 2020. ISSN 1543-0154. doi: 10.1007/S13187-020-01866-3.

[49] J. F. Caron-Flinterman, J. E. Broerse, and J. F. Bunders. The experiential knowledge of patients: A new resource for biomedical research? *Social Science and Medicine*, 60(11):2575–2584, 2005. ISSN 02779536. doi: 10.1016/j.socscimed.2004.11.023.

[50] P. Carter, R. Beech, D. Coxon, M. J. Thomas, and C. Jinks. Mobilising the experiential knowledge of clinicians, patients and carers for applied health-care research. *Contemporary Social Science*, 8(3):307–320, 2013. ISSN 21582041. doi: 10.1080/21582041.2013.767468.

[51] P. G. Casali, A. Le Cesne, A. P. Velasco, D. Kotasek, P. Rutkowski, P. Hohenberger, E. Fumagalli, I. R. Judson, A. Italiano, H. Gelderblom, A. Adenis, J. T. Hartmann, F. Duffaud, D. Goldstein, J. M. Broto, A. Gronchi, A. P. Dei Tos, S. Marréaud, W. T. Van Der Graaf, J. R. Zalcberg, S. Litière, and J. Y. Blay. Time to Definitive Failure to the First Tyrosine Kinase Inhibitor in Localized GI Stromal Tumors Treated With Imatinib As an Adjuvant: A European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group Intergroup Randomized Trial in Collaboration With the Australasian Gastro-Intestinal Trials Group, UNICANCER, French Sarcoma Group, Italian Sarcoma Group, and Spanish Group for Research on Sarcomas. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 33(36):4276–4283, 2015. ISSN 1527-7755. doi: 10.1200/JCO.2015.62.4304.

[52] P. G. Casali, E. Fumagalli, A. Gronchi, J. Zalcberg, D. Kotasek, A. L. Cesne, J. Y. Blay, A. Italiano, P. Reichardt, L. H. Lindner, V. Grünwald, I. R. Judson, W. Van Der Graaf, P. Schöffski, S. Litière, S. Marreaud, S. Leyvraz, A. L. Pousa, S. Sleijfer, J. Verweij, J. M. Kerst, P. Hogendoorn, W. Van Der Graaf, and P. Rutkowski. Ten-year progression-free and overall survival in patients with unresectable or metastatic GI stromal tumors: Long-term analysis of the european organisation for research and treatment of cancer, Italian sarcoma group, and Australasian gastrointestinal tr. *Journal of Clinical Oncology*, 35(15):1713–1720, 2017. ISSN 15277755. doi: 10.1200/JCO.2016.71.0228.

[53] P. G. Casali, N. Abecassis, S. Bauer, R. Biagini, S. Bielack, S. Bonvalot, I. Boukovinas, J. V. Bovee, T. Brodowicz, J. M. Broto, A. Buonadonna, E. De Álava, A. P. Dei Tos, X. G. Del Muro, P. Dileo, M. Eriksson, A. Fedenko, V. Ferraresi, A. Ferrari, S. Ferrari, A. M. Frezza, S. Gasperoni, H. Gelderblom, T. Gil, G. Grignani, A. Gronchi, R. L. Haas, A. Hannu, B. Hassan, P. Hohenberger, R. Issels, H. Joensuu, R. L. Jones, I. Judson, P. Jutte, S. Kaal, B. Kasper, K. Kopeckova, D. A. Krákorová, A. Le Cesne, I. Lugowska, O. Merimsky, M. Montemurro, M. A. Pantaleo, R. Piana, P. Picci, S. Piperno-Neumann, A. L. Pousa, P. Reichardt, M. H. Robinson, P. Rutkowski, A. A. Safwat, P. Schöffski, S. Sleijfer, S. Stacchiotti, K. Sundby Hall, M. Unk, F. Van Coevorden, W. Van Der Graaf, J. Whelan, E. Wardelmann, O. Zaikova, and J. Y. Blay. Gastrointestinal stromal tumours: ESMO-EURACAN Clinical Practice Guidelines

for diagnosis, treatment and follow-up. *Annals of oncology : official journal of the European Society for Medical Oncology*, 29(Suppl 4):iv68–iv78, 2018. ISSN 1569-8041. doi: 10.1093/ANNONC/MDY095.

[54] M. Casparie, A. T. Tiebosch, G. Burger, H. Blauwgeers, A. Van De Pol, J. H. Van Krieken, and G. A. Meijer. Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. *Cellular oncology : the official journal of the International Society for Cellular Oncology*, 29(1):19–24, 2007. ISSN 1570-5870. doi: 10.1155/2007/971816.

[55] O. Caster, J. Dietrich, M. L. Kürzinger, M. Lerch, S. Maskell, G. N. Norén, S. Tcherny-Lessenot, B. Vroman, A. Wisniewski, and J. van Stekelenborg. Assessment of the Utility of Social Media for Broad-Ranging Statistical Signal Detection in Pharmacovigilance: Results from the WEB-RADR Project. *Drug Safety*, 41(12):1355–1369, 2018. ISSN 11791942. doi: 10.1007/s40264-018-0699-2.

[56] E. M. Castro, T. Van Regenmortel, W. Sermeus, and K. Vanhaecht. Patients' experiential knowledge and expertise in health care: A hybrid concept analysis. *Social Theory and Health*, 17(3):307–330, 2019. ISSN 1477822X. doi: 10.1057/s41285-018-0081-6.

[57] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2029.

[58] N. Cesare, C. Grant, and E. O. Nsoesie. Understanding demographic bias and representation in social media health data. In *WebSci 2019 - Companion of the 11th ACM Conference on Web Science*, pages 7–9, 2019. ISBN 9781450361743. doi: 10.1145/3328413.3328415.

[59] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos. Large-Scale Multi-Label Text Classification on EU Legislation. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 6314–6322, 2019.

[60] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001. ISSN 15320464. doi: 10.1006/jbin.2001.1029.

[61] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987. ISSN 0021-9681. doi: 10.1016/0021-9681(87)90171-8. URL `https://pubmed.ncbi.nlm.nih.gov/3558716/`.

[62] H. Chen, W. Chen, C. Liu, L. Zhang, J. Su, and X. Zhou. Relational Network for Knowledge Discovery through Heterogeneous Biomedical and Clinical Features. *Nature Publishing Group*, 2016. doi: 10.1038/srep29915.

[63] G. G. Chowdhury. Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89, 2005. ISSN 00664200. doi: 10.1002/aris.1440370103.

[64] K. W. Church and W. A. Gale. Probability scoring for spelling correction. *Statistics and Computing*, 1(2):93–103, 1991. ISSN 1573-1375. doi: 10.1007/BF01889984.

[65] E. Clark and K. Araki. Text Normalization in Social Media: Progress, Problems and Applications for a Pre-processing System of Casual English. *Procedia Soc Behav Sci*, 27:2–11, 2011. doi: 10.1016/j.sbspro.2011.10.577. URL `www.urbandictionary.com`.

[66] K. Clark, M.-T. Luong, G. Brain, Q. V. Le Google Brain, and C. D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ArXiv*, 2020. URL `https://arxiv.org/abs/2003.10555v1`.

[67] A. Cocos, A. G. Fiks, and A. J. Masino. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821, 2017. doi: 10.1093/jamia/ocw180.

[68] D. A. Coll, C. A. Rosen, K. Auborn, W. P. Potsic, and H. L. Bradlow. Treatment of Recurrent Respiratory Papillomatosis With Indole-3Xarbinol. *American Journal of Otolaryngology*, 18(4):283–285, 1997.

[69] J. A. Custers, R. Tielen, J. B. Prins, J. H. De Wilt, M. F. Gielissen, and W. T. Van Der Graaf. Fear of progression in patients with gastrointestinal stromal tumors (GIST): Is extended lifetime related to the Sword of Damocles? *Acta oncologica (Stockholm, Sweden)*, 54(8):1202–1208, 2015. ISSN 1651-226X. doi: 10.3109/0284186X.2014.1003960.

[70] J. D' Souza and V. Ng. Sieve-Based Entity Linking for the Biomedical Domain. In *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 297–302, 2015.

[71] R. Dagher, M. Cohen, G. Williams, M. Rothmann, J. Gobburu, G. Robbie, A. Rahman, G. Chen, A. Staten, D. Griebel, R. Pazdur, A. D. V. d. Abbeele, E. v. Sonnenberg, and G. D. Demetri. Approval summary: imatinib mesylate in the treatment of metastatic and/or unresectable malignant gastrointestinal stromal tumors. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 8(10):3034–8, 10 2002. ISSN 1078-0432. doi: 10.1158/1078-0432.ccr-06-0858.

[72] X. Dai. Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-3006.

[73] G. D'Amato, D. M. Steinert, J. C. McAuliffe, and J. C. Trent. Update on the biology and therapy of gastrointestinal stromal tumors. *Cancer control : journal of the Moffitt Cancer Center*, 12(1):44–56, 2005. ISSN 1073-2748. doi: 10.1177/107327480501200106.

[74] H. T. Dang. Overview of DUC 2006. In *Proceedings of HLT-NAACL 2006*, pages 1–12, 2006.

[75] K. P. Davison, J. W. Pennebaker, and S. S. Dickerson. Who talks? The social psychology of illness support groups. *American Psychologist*, 55(2):205–217, 2000. ISSN 0003066X. doi: 10.1037/0003-066X.55.2.205.

[76] M. de Boer, A. Dirkson, G. van Oortmerssen, S. Verberne, and W. Kraaij. The Patient Forum Miner: Text mining for patient communities. In *Proceedings of the 17th Dutch-Belgian Information Retrieval Workshop*, pages 25–26, 2018.

[77] J. De Langen, F. Van Hunsel, A. Passier, L. De Jong-Van Den Berg, and K. Van Grootheest. Adverse drug reaction reporting by patients in the Netherlands: Three years of experience. *Drug Safety*, 31(6):515–524, 2008. ISSN 01145916. doi: 10.2165/00002018-200831060-00006.

[78] G. D. Demetri, M. von Mehren, C. D. Blanke, A. D. Van den Abbeele, B. Eisenberg, P. J. Roberts, M. C. Heinrich, D. A. Tuveson, S. Singer, M. Janicek, J. A. Fletcher, S. G. Silverman, S. L. Silberman, R. Capdeville, B. Kiese, B. Peng, S. Dimitrijevic, B. J. Druker, C. Corless, C. D. Fletcher, and H. Joensuu. Efficacy and Safety of Imatinib Mesylate in Advanced Gastrointestinal Stromal Tumors. *New England Journal of Medicine*, 347(7):472–480, 2002. ISSN 0028-4793. doi: 10.1056/NEJMoa020461.

[79] G. D. Demetri, A. T. van Oosterom, C. R. Garrett, M. E. Blackstein, M. H. Shah, J. Verweij, G. McArthur, I. R. Judson, M. C. Heinrich, J. A. Morgan, J. Desai, C. D. Fletcher, S. George, C. L. Bello, X. Huang, C. M. Baum, and P. G. Casali. Efficacy and safety of sunitinib in patients with advanced gastrointestinal stromal tumour after failure of imatinib: a randomised controlled trial. *The Lancet*, 368(9544):1329–1338, 2006. ISSN 0140-6736. doi: 10.1016/S0140-6736(06)69446-4.

[80] G. D. Demetri, P. Reichardt, Y.-K. Kang, J.-Y. Blay, H. Joensuu, R. G. Maki, P. Rutkowski, P. Hohenberger, H. Gelderblom, M. G. Leahy, M. von Mehren, P. Schoffski, M. E. Blackstein, A. Le Cesne, G. Badalamenti, J.-M. Xu, T. Nishida, D. Laurent, I. Kuss, and P. G. Casali. Randomized phase III trial of regorafenib in patients (pts) with metastatic and/or unresectable gastrointestinal stromal tumor (GIST) progressing despite prior treatment with at least imatinib (IM) and sunitinib (SU): GRID trial. *Journal of Clinical Oncology*, 30(18_suppl):LBA10008–LBA10008, 2012. ISSN 0732-183X. doi: 10.1200/jco.2012.30.18_suppl.lba10008.

[81] G. D. Demetri, P. Reichardt, Y.-K. Kang, J.-Y. Blay, P. Rutkowski, H. Gelderblom, P. Hohenberger, M. Leahy, M. von Mehren, H. Joensuu, G. Badalamenti, M. Blackstein, A. Le Cesne, P. Schöffski, R. G. Maki, S. Bauer, B. B. Nguyen, J. Xu, T. Nishida, J. Chung, C. Kappeler, I. Kuss, D. Laurent, P. G. Casali, and GRID

study investigators. Efficacy and safety of regorafenib for advanced gastrointestinal stromal tumours after failure of imatinib and sunitinib (GRID): an international, multicentre, randomised, placebo-controlled, phase 3 trial. *The Lancet*, 381(9863): 295–302, 2013. ISSN 01406736. doi: 10.1016/S0140-6736(12)61857-1.

[82] D. den Hollander, A. R. Dirkson, S. Verberne, W. Kraaij, G. van Oortmerssen, H. Gelderblom, A. Oosten, A. K. L. Reyners, N. Steeghs, W. T. A. van der Graaf, I. M. E. Desar, and O. Husson. Symptoms reported by gastrointestinal stromal tumour (GIST) patients on imatinib treatment: combining questionnaire and forum data. *Supportive Care in Cancer*, mar 2022. ISSN 0941-4355. doi: 10.1007/S00520-022-06929-3. URL https://link.springer.com/10.1007/s00520-022-06929-3.

[83] L. Derczynski, E. Nichols, and M. van Erp. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings ofthe 3rd Workshop on Noisy User-generated Text*, pages 140–147. Association for Computational Linguistics, 2017.

[84] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[85] M. Di Maio, E. Basch, J. Bryce, and F. Perrone. Patient-reported outcomes in the evaluation of toxicity of anticancer treatments. *Nature reviews. Clinical oncology*, 13(5):319–325, may 2016. ISSN 1759-4782. doi: 10.1038/NRCLINONC.2015.222.

[86] G.-A. Dima, D.-C. Cercel, and M. Dascalu. Transformer-based Multi-Task Learning for Adverse Effect Mention Analysis in Tweets. In *Proceedings of the Sixth Social Media Mining for Health Workshop 2021*, pages 44–51, 2021. doi: 10.18653/v1/2021.smm4h-1.7.

[87] A. Dirkson, S. Verberne, and W. Kraaij. Narrative Detection in Online Patient Communities. In A. Jorge, R. Campos, A. Jatowt, and S. Bhatia, editors, *Proceedings of the Text2StoryIR'19 Workshop*. CEUR-WS, 2019. URL http://ceur-ws.org/Vol-2342/paper3.pdf.

[88] A. Dirkson, S. Verberne, W. Kraaij, G. V. Oortmerssen, and H. Gelderblom. Automated gathering of real-world evidence from online patient fora can complement pharma. *Scientific Reports*, 2022. doi: 10.1038/s41598-022-13894-8.

[89] Y. Doval Mosquera, J. Vilares, and C. Gómez-Rodríguez. Lysgroup: Adapting a spanish microtext normalization system to english. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 99–105, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4315.

[90] R. I. Doğan, R. Leaman, and Z. Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014. ISSN 15320464. doi: 10.1016/j.jbi.2013.12.006.

[91] M. Dredze, D. A. Broniatowski, and K. M. Hilyard. Zika vaccine misconceptions: A social media analysis. *Vaccine*, 34(30):3441–2, 2016. ISSN 1873-2518. doi: 10.1016/j.vaccine.2016.05.008.

[92] ECSA. Ten principles of Citizen Science. Technical report, European Citizen Science Association, 2015. URL http://doi.org/10.17605/OSF.IO/XPR2N.

[93] G. I. Ector, R. P. Hermens, and N. M. Blijlevens. Filling the gaps of patient information and comprehension. *Current opinion in oncology*, 32(4):262–268, 2020. ISSN 1531-703X. doi: 10.1097/CCO.0000000000000633.

[94] I. Eekhout, M. A. Van De Wiel, and M. W. Heymans. Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: Power and applicability analysis. *BMC Medical Research Methodology*, 17(1):1–12, 2017. ISSN 14712288. doi: 10.1186/S12874-017-0404-7/TABLES/4. URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0404-7.

[95] I. Eerola. Rare Diseases remains a major unmet medical need. Technical report, European Commission, Brussels, 2017. URL https://doi.org/10.2777/76751.

[96] F. Efficace, G. Rosti, N. Aaronson, F. Cottone, E. Angelucci, S. Molica, M. Vignetti, F. Mandelli, and M. Baccarani. Patient- versus physician-reporting of symptoms and health status in chronic myeloid leukemia. *Haematologica*, 99(4):788–793, 2014. ISSN 1592-8721. doi: 10.3324/HAEMATOL.2013.093724.

[97] J. Eisenstein. What to do about bad language on the internet. In *Proceedings of NAACL-HLT 2013*, pages 9–14. Association for Computational Linguistics, 2013.

[98] I. A. Eland, K. J. Belton, A. C. Van Grootheest, A. P. Meiners, M. D. Rawlins, and B. H. Stricker. Attitudinal survey of voluntary reporting of adverse drug reactions. *British Journal of Clinical Pharmacology*, 48(4):623–627, 1999. ISSN 03065251. doi: 10.1046/j.1365-2125.1999.00060.x.

[99] J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[100] EORTC Quality of Life Group. EORTC QLQ-30 Scoring Manual, 2001. URL https://www.eortc.org/app/uploads/sites/2/2018/02/SCmanual.pdf.

[101] European Medicine Agency. Summary of Product Characteristics Imatinib. Technical report, European Medicine Agency. URL https://www.ema.europa.eu/en/documents/product-information/glivec-epar-product-information_en.pdf.

[102] European Medicine Agency. Guideline on good pharmacovigilance practices (GVP) - Annex I - Definitions (Rev 4). Technical report, European Medicine Agency, 2017. URL https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-annex-i-definitions-rev-4_en.pdf.

[103] European Medicine Agency. Ayvakyt, 2020. URL https://www.ema.europa.eu/en/medicines/human/EPAR/ayvakyt.

[104] N. Fareed, C. M. Swoboda, P. Jonnalagadda, and T. R. Huerta. Persistent digital divide in health-related internet use among cancer survivors: findings from the Health Information National Trends Survey, 2003-2018. *Journal of cancer survivorship: research and practice*, 15(1):87–98, 2021. ISSN 1932-2267. doi: 10.1007/S11764-020-00913-8.

[105] L. Fauske, I. Hompland, G. Lorem, H. Bondevik, and Ø. S. Bruland. Perspectives on treatment side effects in patients with metastatic gastrointestinal stromal tumour: a qualitative study. *Clinical Sarcoma Research 2019 9:1*, 9(1):1–8, 2019. ISSN 2045-3329. doi: 10.1186/S13569-019-0116-3.

[106] P. Fayers and A. Bottomley. Quality of life research within the EORTC - The EORTC QLQ-C30. *European Journal of Cancer*, 38:S125–133, 2002. ISSN 09598049. doi: 10.1016/s0959-8049(01)00448-8.

[107] R. Feldman, J. Sanger, and C. U. Press. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007. ISBN 9780521836579.

[108] T. Ferguson. The first generation of e-patients. *BMJ*, 328(7449):1148–1149, 2004. ISSN 0959-8138. doi: 10.1136/bmj.328.7449.1148.

[109] T. Ferguson and B. Kelly. E-patients prefer egroups to doctors for 10 of 12 aspects of health care., 1991. URL http://www.fergusonreport.com/articles/fr039905.htm.

[110] P. Fivez, S. Šuster, and W. Daelemans. Unsupervised context-sensitive spelling correction of clinical free-text with word and character n-gram embeddings. In *BioNLP 2017*, pages 143–148, Vancouver, Canada,, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2317.

[111] J. Fjeld, N. Achten, H. Hilligoss, A. C. Nagy, and M. Srikumar. Principled Artificial intelligence: Mapping consensus in Ethical and Rights-based Approaches to Principles for AI. Technical report, Berkman Klein Center for Internet & Society, 2020.

[112] R. Ginn, P. Pimpalkhute, A. Nikfarjam, A. Patki, K. O 'connor, A. Sarker, K. Smith, and G. Gonzalez. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*, 2014.

[113] W. G. Goettsch, S. D. Bos, N. Breekveldt-Postma, M. Casparie, R. M. Herings, and P. C. Hogendoorn. Incidence of gastrointestinal stromal tumours is underestimated: Results of a nation-wide study. *European Journal of Cancer*, 41(18):2868–2872, 2005. ISSN 0959-8049. doi: 10.1016/J.EJCA.2005.09.009.

[114] S. Golder, G. Norman, and Y. K. Loke. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. *British Journal of Clinical Pharmacology*, 80(4):878–888, 2015. ISSN 13652125. doi: 10.1111/bcp.12746.

[115] S. Golder, K. Smith, K. O'Connor, R. Gross, S. Hennessy, and G. Gonzalez-Hernandez. A Comparative View of Reported Adverse Effects of Statins in Social Media, Regulatory Data, Drug Information Databases and Systematic Reviews. *Drug Safety*, pages 1–13, 2020. ISSN 11791942. doi: 10.1007/s40264-020-00998-1.

[116] G. Gonzalez-Hernandez, A. Sarker, K. O 'Connor, and G. Savova. Capturing the Patient's Perspective : a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of medical informatics*, pages 214–217, 2017. doi: 10.15265/IY-2017-029.

[117] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In *ICLR 2015*, 2015.

[118] R. C. Griggs, M. Batshaw, M. Dunkle, R. Gopal-Srivastava, E. Kaye, J. Krischer, T. Nguyen, K. Paulus, and P. A. Merkel. Clinical research for rare disease: Opportunities, challenges, and solutions. *Molecular Genetics and Metabolism*, 96 (1):20–26, 2009. ISSN 10967192. doi: 10.1016/j.ymgme.2008.10.003.

[119] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3 (1):1–23, 2021. doi: 10.1145/3458754.

[120] B. Han and T. Baldwin. Lexical Normalization for Social Media Text. *ACM Trans. Intell. Syst. Technol. Article*, 4(5), 2013. doi: 10.1145/2414425.2414430.

[121] B. Han, P. Cook, and T. Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432. Association for Computational Linguistics, 2012.

[122] J. Y. Han, J.-H. Kim, H. J. Yoon, M. Shim, F. M. McTavish, and D. H. Gustafson. Social and Psychological Determinants of Levels of Engagement With an Online Breast Cancer Support Group: Posters, Lurkers, and Nonusers. *Journal of Health Communication*, 17(3):356–371, 2012. doi: 10.1080/10810730.2011.585696.

[123] J. Y. Han, J. Hou, E. Kim, and D. H. Gustafson. Lurking as an Active Participation Process: A Longitudinal Investigation of Engagement with an Online Cancer Support Group. *Health Communication*, 29(9):911–923, 2014. doi: 10.1080/10410236.2013.816911.

[124] S. Han, T. Tran, A. Rios, and R. Kavuluru. Team UKNLP: Detecting ADRs, Classifying Medication Intake Messages, and Normalizing ADR Mentions on Twitter. In *Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H)*, 2017.

[125] E. Hargittai. Whose Space? Differences Among Users and Non-Users of Social Network Sites. *Journal of Computer-Mediated Communication*, 13(1):276–297, 2007. ISSN 10836101. doi: 10.1111/j.1083-6101.2007.00396.x.

[126] E. Hargittai. Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, 659(1):63–76, 2015. ISSN 0002-7162. doi: 10.1177/0002716215570866.

[127] E. Hargittai. Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, 38(1):10–24, 2020. doi: 10.1177/0894439318788322.

[128] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics*, 91(6):1010–1021, 2012. ISSN 0009-9236. doi: 10.1038/clpt.2012.50.

[129] A. Hartzler and W. Pratt. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. *Journal of medical Internet research*, 13(3): e62, 2011. ISSN 1438-8871. doi: 10.2196/jmir.1728.

[130] L. Hazell and S. A. W. Shakir. Under-Reporting of Adverse Drug Reactions A Systematic Review. *Drug Safety*, 29(5):385–396, 2006. doi: 10.2165/00002018-200629050-00003.

[131] H. E. Heemstra, S. van Weely, H. A. Büller, H. G. Leufkens, and R. L. de Vrueh. Translation of rare disease research into orphan drug development: disease matters. *Drug Discovery Today*, 14(23-24):1166–1173, 2009. ISSN 13596446. doi: 10.1016/j.drudis.2009.09.008.

[132] G. Heigold, G. Neumann, and J. van Genabith. How Robust Are Character-Based Word Embeddings in Tagging and MT Against Wrod Scramlbing or Randdm Nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 68–80, 2018.

[133] High-Level Expert Group on AI (AI HLEG). Ethics guidelines for trustworthy AI. Technical report, European Commission, 2019. URL `https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html`.

[134] F. Hill, R. Reichart, and A. Korhonen. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695, 2015. doi: 10.1162/COLI.

[135] J. Hong, T. Kim, H. Lim, and J. Choo. AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4692–4700, 2021.

[136] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[137] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, 2018. Association for Computational Linguistics.

[138] G. Hripcsak and D. F. Heitjan. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35(2):99–110, 2002. ISSN 1532-0464. doi: 10.1016/S1532-0464(02)00500-2.

[139] Y.-L. Hsieh, M. Cheng, D.-C. Juan, W. Wei, W.-L. Hsu, and C.-J. Hsieh. On the Robustness of Self-Attentive Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529, 2019. doi: 10.18653/v1/P19-1147.

[140] Y. Hu, X. Jing, Y. Ko, and J. T. Rayz. Misspelling Correction with Pre-trained Contextual Language Model. *arXiv*, 2021. URL `https://arxiv.org/abs/2101.03204`.

[141] X. Huang, M. C. Smith, M. Paul, D. Ryzhkov, S. Quinn, D. Broniatowski, and M. Dredze. Examining Patterns of Influenza Vaccination in Social Media. In *AAAI Joint Workshop on Health Intelligence (W3PHIAI)*, 2017.

[142] Hyeoneui, J. Mentzer, and R. Taira. Developing a Physical Activity Ontology to Support the Interoperability of Physical Activity Data. *J Med Internet Res 2019;21(4):e12776 https://www.jmir.org/2019/4/e12776*, 21(4):e12776, 2019. doi: 10.2196/12776.

[143] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025, 2020. doi: 10.1609/aaai.v34i05.6311.

[144] N. Jin. NCSU-SAS-ning: Candidate generation and feature engineering for supervised lexical normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 87–92, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4313.

[145] H. Joensuu, J. C. Trent, and P. Reichardt. Practical management of tyrosine kinase inhibitor-associated side effects in GIST. *Cancer treatment reviews*, 37(1):75–88, 2011. ISSN 1532-1967. doi: 10.1016/J.CTRV.2010.04.008.

[146] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[147] C. Johnson, N. Aaronson, J. M. Blazeby, A. Bottomley, P. Fayers, M. Koller, D. Kuliś, J. Ramage, M. Sprangers, G. Velikova, and T. Young. EORTC QUALITY OF LIFE GROUP Guidelines for Developing Questionnaire Modules. Technical report, European Organisation for Research and Treatment of Cancer, 2011. URL https://www.eortc.org/app/uploads/sites/2/2018/02/guidelines_for_developing_questionnaire-_final.pdf.

[148] M. O. Johnson and T. B. Neilands. Coping with HIV treatment side effects: Conceptualization, measurement, and linkages. *AIDS and Behavior*, 11(4):575–585, 2007. doi: 10.1007/S10461-007-9229-4.

[149] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Pearson Education, 3rd edition, 2021.

[150] K. R. Kanakarajan, B. Kundumani, and M. Sankarasubbu. BioELECTRA:Pretrained Biomedical text Encoder using Discriminators. In *Proceedings of the BioNLP 2021 workshop*, pages 143–154, 2021.

[151] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81, 2015. ISSN 1532-0464. doi: 10.1016/J.JBI.2015.03.010.

[152] P. Karisani and E. Agichtein. Did you really just have a heart attack?: Towards robust detection of personal health mentions in social media. In *Proceedings of the 2018 World Wide Web Conference*, pages 137–146, 2018. ISBN 978-1-4503-5639-8. doi: 10.1145/3178876.3186055.

[153] R. Kaur, J. A. Ginige, O. Obst, and A. Ginige. A Systematic Literature Review of Automated ICD Coding and Classification Systems using Discharge Summaries. *ArXiv (preprint)*, 2021. URL https://arxiv.org/abs/2107.10652.

[154] S. Khosla, R. White, J. Medina, M. Ouwens, C. Emmas, T. Koder, G. Male, and S. Leonard. Real world evidence (RWE) - a disruptive innovation or the quiet evolution of medical evidence generation? *F1000Research*, 7:111, 2018. ISSN 20461402. doi: 10.12688/f1000research.13585.2.

[155] I. Kickbusch and D. Gleicher. Governance for Health in the 21st Century. Technical report, World Health Organization (WHO), Copenhagen, 2012. URL https://www.euro.who.int/en/publications/abstracts/governance-for-health-in-the-21st-century.

[156] T. Kikuchi, T. Suzuki, H. Uchida, K. Watanabe, and M. Mimura. Coping strategies for antidepressant side effects: An Internet survey. *Journal of Affective Disorders*, 143 (1-3):89–94, dec 2012. ISSN 0165-0327. doi: 10.1016/J.JAD.2012.04.039.

[157] N. Kingod, B. Cleal, A. Wahlberg, and G. R. Husted. Online peer-to-peer communities in the daily lives of people with chronic illness: A qualitative systematic review. *Qualitative Health Research*, 27(1):89–99, 2017. doi: 10.1177/1049732316680203.

[158] A. Kinsora, K. Barron, Q. Mei, and V. G. V. Vydiswaran. Creating a Labeled Dataset for Medical Misinformation in Health Forums. In *IEEE International Conference on Healthcare Informatics*, 2017. doi: 10.1109/ICHI.2017.93.

[159] A. Klein, A. Magge, K. O'Connor, H. Cai, D. Weissenbacher, and G. Gonzalez-Hernandez. A chronological and geographical analysis of personal reports of covid-19 on twitter. *medRxiv (preprint)*, 2020. doi: 10.1101/2020.04.19.20069948.

[160] D. C. Klonoff, A. Gutierrez, A. Fleming, and D. Kerr. Real-World Evidence Should Be Used in Regulatory Decisions About New Pharmaceutical and Medical Device Products for Diabetes. *Journal of Diabetes Science and Technology*, 13(6):995–1000, 2019. ISSN 19322968. doi: 10.1177/1932296819839996.

[161] P. G. Kluetz, A. Slagle, E. J. Papadopoulos, L. L. Johnson, M. Donoghue, V. E. Kwitkowski, W. H. Chen, R. Sridhara, A. T. Farrell, P. Keegan, G. Kim, and R. Pazdur. Focusing on Core Patient-Reported Outcomes in Cancer Clinical Trials: Symptomatic Adverse Events, Physical Function, and Disease-Related Symptoms. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 22(7):1553–1558, 2016. ISSN 1557-3265. doi: 10.1158/1078-0432.CCR-15-2035.

[162] I. Koiranen, ·. Teo Keipi, ·. A. Koivula, and P. Räsänen. Changing patterns of social media use? A population-level study of Finland. *Universal Access in the Information Society*, 19:603–617, 2020. doi: 10.1007/s10209-019-00654-1.

[163] R. Kraut, S. Kiesler, B. Boneva, J. Cummings, V. Helgeson, and A. Crawford. Internet Paradox Revisited. *Journal of Social Issues*, 58(1):49–74, 2002.

[164] K. Kukich. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24:377–439, 1992.

[165] D. Kulis, A. Bottomley, C. Whittaker, L. van de Poll-Franse, A. Darlington, B. Holzner, M. Koller, J. Reijneveld, K. Tomaszewski, and M. Grønvold. The Use of The Eortc Item Library To Supplement Eortc Quality of Life Instruments. *Value in Health*, 20 (9):A775, 2017. ISSN 10983015. doi: 10.1016/J.JVAL.2017.08.2236.

[166] A. Kumar, P. Makhija, and A. Gupta. Noisy Text Data: Achilles' Heel of BERT. In *Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text*, pages 16–21. Association for Computational Linguistics, 2020.

[167] K. Kvarnström, A. Westerholm, M. Airaksinen, and H. Liira. Factors Contributing to Medication Adherence in Patients with a Chronic Condition: A Scoping Review of Qualitative Research. *Pharmaceutics 2021, Vol. 13, Page 1100*, 13(7):1100, jul 2021. doi: 10.3390/PHARMACEUTICS13071100.

[168] K. H. Lai, M. Topaz, F. R. Goss, and L. Zhou. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55:188–195, 2015. doi: 10.1016/j.jbi.2015.04.008.

[169] A. Lamb, M. J. Paul, and M. Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Proceedings of NAACL-HLT*, pages 789–795, 2013.

[170] J. B. Lamy, A. Venot, and C. Duclos. PyMedTermino: An open-source generic API for advanced terminology services. *Studies in Health Technology and Informatics*, 210: 924–928, 2015. ISSN 18798365. doi: 10.3233/978-1-61499-512-8-924.

[171] J. Lardon, R. Abdellaoui, F. Bellet, H. Asfari, J. Souvignet, N. Texier, M. C. Jaulent, M. N. Beyens, A. Burgun, and C. Bousquet. Adverse drug reaction identification and extraction in social media: A scoping review. *Journal of Medical Internet Research*, 17(7):1–16, 2015. ISSN 14388871. doi: 10.2196/jmir.4304.

[172] Q. Le and T. Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st international conference on machine learning*, 2014.

[173] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards Internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010*, pages 117–125, 2010.

[174] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. doi: 10.1093/bioinformatics/btz682.

[175] S. Leeman-Munk, J. Lester, and J. Cox. NCSU_SAS_SAM: Deep encoding and reconstruction for normalization of noisy text. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 154–161, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4323.

[176] M. Lengsavath, A. Dal Pra, A. M. de Ferran, S. Brosch, L. Härmark, V. Newbould, and S. Goncalves. Social Media Monitoring and Adverse Drug Reaction Reporting in Pharmacovigilance: An Overview of the Regulatory Landscape. *Therapeutic Innovation and Regulatory Science*, 51(1):125–131, 2017. ISSN 21684804. doi: 10.1177/2168479016663264.

[177] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[178] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, and F. Ai. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics (ACL), jul 2020. doi: 10.18653/V1/2020.ACL-MAIN.703.

[179] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:68, 2016. doi: 10.1093/database/baw068.

[180] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu. Bert-attack: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6193–6202, 2020. doi: 10.18653/v1/2020.emnlp-main.500.

[181] Q. Li, Q. Zhang, and L. Si. eventAI at SemEval-2019 Task 7: Rumor Detection on Social Media by Exploiting Content, User Credibility and Propagation Information. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 855–859, 2019. doi: 10.18653/v1/s19-2148.

[182] X. Li, W. Gao, S. Feng, Y. Zhang, and D. Wang. Boundary detection with BERT for span-level emotion cause analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 676–682. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.60.

[183] Z. Li, Z. Yang, L. Luo, Y. Xiang, and H. Lin. Exploiting adversarial transfer learning for adverse drug reaction detection from texts. *Journal of Biomedical Informatics*, page 103431, 2020. doi: 10.1016/j.jbi.2020.103431.

[184] N. Limsopatham and N. Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1096.

[185] B. Y. Lin, W. Gao, J. Yan, R. Moreno, and X. Ren. RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models. In *Proceedings ofthe 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737, 2021.

[186] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Ltd, New York City, USA, 1986.

[187] F. Liu, Z. Meng, M. Basaldella, and N. Collier. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference ofthe North American Chapter ofthe Association for Computational Linguistics*, pages 4228–4238, 2021.

[188] X. Liu and H. Chen. Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums. In *International conference on smart health*, pages 134–150. Springer, 2013. doi: 10.1007/978-3-642-39844-5_16.

[189] E. Lopez-Gonzalez, M. T. Herdeiro, and A. Figueiras. Determinants of under-reporting of adverse drug reactions: A systematic review. *Drug Safety*, 32(1):19–31, 2009. ISSN 01145916. doi: 10.2165/00002018-200932010-00002.

[190] M. Lui and T. Baldwin. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th annual meeting of the association of computational linguistics*, pages 25–30, 2012.

[191] X. Ma, P. Xu, Z. Wang, R. Nallapati, and B. Xiang. Domain Adaptation with BERT-based Domain Classification and Data Selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83. Association for Computational Linguistics (ACL), 2019. ISBN 9781950737789. doi: 10.18653/V1/D19-6109.

[192] J. B. Macdonald, B. Macdonald, L. E. Golitz, P. LoRusso, and A. Sekulic. Cutaneous adverse effects of targeted therapies: Part I: Inhibitors of the cellular membrane. *Journal of the American Academy of Dermatology*, 72(2):203–218, 2015. ISSN 1097-6787. doi: 10.1016/J.JAAD.2014.07.032.

[193] A. Magge, A. Z. Klein, A. Miranda-Escalada, M. A. Al-Garadi, I. Alimova, Z. Miftahutdinov, S. Lima López, I. Flores, K. O'connor, D. Weissenbacher, E. Tutubalina, J. M. Banda, M. Krallinger, and G. Gonzalez-Hernandez. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health Workshop 2021*, pages 21–32, 2021. doi: 10.18653/v1/2021.smm4h-1.4.

[194] A. Magge, E. Tutubalina, Z. Miftahutdinov, I. Alimova, A. Dirkson, S. Verberne, D. Weissenbacher, and G. Gonzalez-Hernandez. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 2021. doi: 10.1093/jamia/ocab114.

[195] A. S. Maiya. ktrain: A low-code library for augmented machine learning. *arXiv*, 2020. URL `https://arxiv.org/pdf/2004.10703.pdf`.

[196] V. Malykh, V. Logacheva, and T. Khakhulin. Robust Word Vectors: Context-Informed Embeddings for Noisy Texts. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 54–63, 2018.

[197] J. J. Mao, A. Chung, A. Benton, S. Hill, L. Ungar, C. E. Leonard, S. Hennessy, and J. H. Holmes. Online discussion of drug side effects and discontinuation among breast cancer survivors. *Pharmacoepidemiology and drug safety*, 22(3):256–62, 2013. ISSN 1099-1557. doi: 10.1002/pds.3365.

[198] S. Mariani, E. Abruzzese, S. Basciani, D. Fiore, A. Persichetti, M. Watanabe, G. Spera, and L. Gnessi. Reversible hair depigmentation in a patient treated with imatinib. *Leukemia research*, 35(6), 2011. ISSN 1873-5835. doi: 10.1016/J.LEUKRES.2010.11.028.

[199] K. Y. McKenna and J. A. Bargh. Coming Out in the Age of the Internet: Identity "Demarginalization" Through Virtual Group Participation. *Journal of Personality and Social Psychology*, 75(3):681–694, 1998. doi: 10.1037/0022-3514.75.3.681.

[200] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer Sentinel Mixture Models. In *ICLR 2017*, 2017.

[201] S. Merity, N. S. Keskar, and R. Socher. Regularizing and Optimizing LSTM Language Models. In *ICLR 2018*, 2018.

[202] M. Merolli, K. Gray, and F. Martin-Sanchez. Health outcomes and related effects of using social media in chronic disease management: A literature review and analysis of affordances. *Journal of Biomedical Informatics*, 46(6):957–969, 12 2013. ISSN 1532-0464. doi: 10.1016/J.JBI.2013.04.010.

[203] A. Metke-Jimenez and S. Karimi. Concept Extraction to Identify Adverse Drug Reactions in Medical Forums: A Comparison of Algorithms. *ArXiv*, 2015. URL `http://arxiv.org/abs/1504.06936`.

[204] A. Metke-Jimenez, S. Karimi, and C. Paris. Evaluation of Text-Processing Algorithms for Adverse Drug Event Extraction from Social Media. In *SoMeRA '14 Proceedings of the first international workshop on Social media retrieval and analysis*, pages 15–20, 2014. doi: 10.1145/2632188.2632200.

[205] P. Michael E. Porter. Perspective - What Is Value in Health Care? *The New England Journal of Medicine*, 363(1):1–3, 2010. ISSN 15334406. doi: 10.1056/NEJMp1011024.

[206] Z. Miftahutdinov and E. Tutubalina. Deep Neural Models for Medical Concept Normalization in User-Generated Texts. In *Student Research Workshop ACL 2019*, 2019. doi: 10.18653/v1/P19-2055.

[207] Z. S. Miftahutdinov, E. V. Tutubalina, and A. E. Tropsha. Identifying disease-related expressions in reviews using conditional random fields. In *Proceedings of the International Conference Dialogue 2017*, 2017.

[208] W. Min and B. Mott. NCSU_SAS_WOOKHEE: A deep contextual long-short term memory model for text normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 111–119, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4317.

[209] I. Mondal. BBAEG: Towards BERT-based biomedical adversarial example generation for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5378–5384, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.423.

[210] D. L. Mowery, B. R. South, L. Christensen, J. Leng, L.-M. Peltonen, S. Salanterä, H. Suominen, D. Martinez, S. Velupillai, N. Elhadad, G. Savova, S. Pradhan, and W. W. Chapman. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2. *Journal of biomedical semantics*, 7:43, 2016. ISSN 2041-1480. doi: 10.1186/s13326-016-0084-y.

[211] N. Mrkšić, D. Ó Séaghdha, B. Thomson, M. Gašíc, L. Rojas-Barahona, P.-H. Su, D. Vandyke, T.-H. Wen, and S. Young. Counter-fitting Word Vectors to Linguistic Constraints. In *Proceedings of NAACL-HLT 2016*, pages 142–148, 2016.

[212] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings ofNAACL-HLT 2018*, pages 1101–1111, 2018.

[213] B. Muller, B. Sagot, and D. Seddah. Enhancing BERT for Lexical Normalization. In *The 5th Workshop on Noisy User-generated Text (W- NUT)*, 2019.

[214] N. Nakashole. Commonsense about human senses: Labeled data collection processes. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 43–52. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-6005.

[215] National Cancer Institute. NCI Dictionary of Cancer Terms. URL `https://www.cancer.gov/publications/dictionaries/cancer-terms`.

[216] A. Nikfarjam and G. H. Gonzalez. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. In *AMIA Annual Symposium proceedings*, pages 1019–1026. American Medical Informatics Association, 2011.

[217] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association : JAMIA*, 22(3):671–81, 5 2015. ISSN 1527-974X. doi: 10.1093/jamia/ocu041.

[218] J. Niu, Y. Yang, S. Zhang, Z. Sun, and W. Zhang. Multi-task Character-Level Attentional Networks for Medical Concept Normalization. *Neural Processing Letters*, 49(3):1239–1256, 2019. ISSN 1370-4621. doi: 10.1007/s11063-018-9873-x.

[219] P. Norvig. Natural Language Corpus Data. In J. H. Toby Segaran, editor, *Beautiful Data: The Stories Behind Elegant Data Solutions*, pages 219–242. O'Reilly Media, 2009. ISBN 978-0-596-15711-1.

[220] N. F. Noy, N. Griffith, and M. A. Musen. Collecting community-based mappings in an ontology repository. In *Proceedings of the 7th International Conference on The Semantic Web*, page 371–386, 2008. ISBN 9783540885634. doi: 10.1007/978-3-540-88564-1_24.

[221] M. D. T. Nzali, S. Bringay, C. Lavergne, C. Mollevi, and T. Opitz. What patients can tell us: Topic analysis for social media on breast cancer. *JMIR Medical Informatics*, 5(3):1–17, 2017. ISSN 22919694. doi: 10.2196/medinform.7779.

[222] Observational Health Data Sciences and Informatics project. *The book of OHDSI*. Observational Health Data Sciences and Informatics, 2021. ISBN 978-1088855195. URL `https://ohdsi.github.io/TheBookOfOhdsi/`.

[223] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42 (13):5645–5657, 2015. ISSN 0957-4174. doi: 10.1016/J.ESWA.2015.02.055.

[224] K. O'Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, and G. Gonzalez. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. *AMIA Annual Symposium proceedings*, pages 924–933, 2014. ISSN 1942-597X.

[225] B. O'Donovan, R. M. Rodgers, A. R. Cox, and J. Krska. Use of information sources regarding medicine side effects among the general population: a cross-sectional survey. *Primary Health Care Research & Development*, 20(e153):1–8, 2019. ISSN 1463-4236. doi: 10.1017/S1463423619000574.

[226] K. Ogata, A. Kimura, N. Nakazawa, M. Suzuki, T. Yanoma, Y. Ubukata, K. Iwamatsu, N. Kogure, M. Yanai, and H. Kuwano. Long-Term Imatinib Treatment for Patients with Unresectable or Recurrent Gastrointestinal Stromal Tumors. *Digestion*, 97(1): 20–25, 2018. ISSN 14219867. doi: 10.1159/000484102.

[227] I. Omar and E. Harris. The Use of Social Media in ADR Monitoring and Reporting. *Journal of Pharmacovigilance*, 4(6), 2016. ISSN 2329-6887. doi: 10.4172/2329-6887.1000223.

[228] D. Pappa and L. K. Stergioulas. Harnessing social media data for pharmacovigilance: a review of current state of the art, challenges and future directions. *International Journal of Data Science and Analytics*, 8(2):113–135, 2019. ISSN 23644168. doi: 10.1007/s41060-019-00175-3.

[229] A. Park, A. L. Hartzler, J. Huh, D. W. Mcdonald, and W. Pratt. Automatically Detecting Failures in Natural Language Processing Tools for Online Community Text. *J Med Internet Res*, 17(8), 2015. doi: 10.2196/jmir.4612.

[230] J. Patrick, M. Sabbagh, S. Jain, and H. Zheng. Spelling correction in clinical notes with emphasis on first suggestion accuracy. In *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 2–8, 2010.

[231] M. J. Paul and M. Dredze. A Model for Mining Public Health Topics from Twitter. Technical report, Johns Hopkins University, 2011. URL `http://www.michaeljpaul.com/files/2011.tech.twitter_health.pdf`.

[232] Y. Peng, S. Yan, and Z. Lu. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the BioNLP 2019 workshop*, pages 58–65, 2019.

[233] L. Philips. The double metaphone search algorithm. *C/C++ Users J.*, 18(6):38–43, 2000. ISSN 1075-2838.

[234] C. M. Phillips, A. Parmar, H. Guo, D. Schwartz, W. Isaranuwatchai, J. Beca, W. Dai, J. Arias, S. Gavura, and K. K. Chan. Assessing the efficacy-effectiveness gap for cancer therapies: A comparison of overall survival and toxicity between clinical trial and population-based, real-world data for contemporary parenteral cancer therapeutics. *Cancer*, 126(8):1717–1726, 2020. ISSN 1097-0142. doi: 10.1002/CNCR.32697.

[235] P. Pimpalkhute, A. Patki, A. Nikfarjam, and G. Gonzalez. Phonetic spelling filter for keyword selection in drug mention mining from social media. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2014:90–5, 2014. ISSN 2153-4063.

[236] K. Plueschke, P. McGettigan, A. Pacurariu, X. Kurz, and A. Cave. EU-funded initiatives for real world evidence: Descriptive analysis of their characteristics and relevance for regulatory decision-making. *BMJ Open*, 8(6):21864, 2018. ISSN 20446055. doi: 10.1136/bmjopen-2018-021864.

[237] G. R. Polich. Rare disease patient groups as clinical researchers. *Drug Discovery Today*, 17(3-4):167–172, 2012. ISSN 13596446. doi: 10.1016/j.drudis.2011.09.020.

[238] J. Pols. Knowing Patients: Turning Patient Knowledge into Science. *Science, Technology, & Human Values*, 39(1):73–97, 2014. doi: 10.1177/0162243913504306.

[239] H. Poort, W. T. van der Graaf, R. Tielen, M. Vlenterie, J. A. Custers, J. B. Prins, C. A. Verhagen, M. F. Gielissen, and H. Knoop. Prevalence, Impact, and Correlates of Severe Fatigue in Patients With Gastrointestinal Stromal Tumors. *Journal of pain and symptom management*, 52(2):265–271, 2016. ISSN 1873-6513. doi: 10.1016/J.JPAINSYMMAN.2016.02.019.

[240] J. Popay and G. Williams. Public health research and lay knowledge. *Social Science and Medicine*, 42(5):759–768, 1996. ISSN 02779536. doi: 10.1016/0277-9536(95)00341-X.

[241] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, and G. Savova. SemEval-2014 Task 7: Analysis of Clinical Text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, 2014. doi: 10.3115/v1/s14-2007.

[242] J. Price. What Can Big Data Offer the Pharmacovigilance of Orphan Drugs? *Clinical Therapeutics*, 38(12):2533–2545, 2016. ISSN 1879114X. doi: 10.1016/j.clinthera.2016.11.009.

[243] L. Prior. Belief, knowledge and expertise: the emergence of the lay expert in medical sociology. *Sociology of Health & Illness*, 25(3):41–57, 2003. ISSN 01419889. doi: 10.1111/1467-9566.00339.

[244] C. A. Radawski, T. A. Hammad, S. Colilla, P. Coplan, K. Hornbuckle, E. Freeman, M. Y. Smith, R. E. Sobel, P. Bahri, A. E. Arias, and D. Bennett. The utility of real-world evidence for benefit-risk assessment, communication, and evaluation of pharmaceuticals: Case studies. *Pharmacoepidemiology and Drug Safety*, 29(12):1532–1539, 2020. ISSN 1053-8569. doi: 10.1002/pds.5167.

[245] L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 157–176, Dordrecht, 1999. Springer Netherlands. ISBN 978-94-017-2390-9. doi: 10.1007/978-94-017-2390-9_10.

[246] P. Reichardt. The Story of Imatinib in GIST-a Journey through the Development of a Targeted Therapy. *Oncol Res Treat*, 41:472–477, 2018. doi: 10.1159/000487511. URL www.karger.com/ort.

[247] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992, 2019.

[248] M. Reynaert. *Text-Induced Spelling Correction*. PhD thesis, Tilburg University, 2005.

[249] T. Richter, S. Nestler-Parr, R. Babela, Z. M. Khan, T. Tesoro, E. Molsen, and D. A. Hughes. Rare Disease Terminology and Definitions-A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. *Value in Health*, 18(6): 906–914, 2015. ISSN 15244733. doi: 10.1016/j.jval.2015.05.008.

[250] A. Rios and R. Kavuluru. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3132–3142, 2018. doi: 10.18653/V1/D18-1352.

[251] K. M. Robinson. Unsolicited narratives from the internet: A rich source of qualitative data. *Qualitative Health Research*, 11(5):706–714, 2001. ISSN 10497323. doi: 10.1177/104973201129119398.

[252] L. Rolfes, F. van Hunsel, K. Taxis, and E. van Puijenbroek. The Impact of Experiencing Adverse Drug Reactions on the Patient's Quality of Life: A Retrospective Cross-Sectional Study in the Netherlands. *Drug Safety*, 39(8):769–776, 2016. ISSN 11791942. doi: 10.1007/s40264-016-0422-0.

[253] L. Rolfes, F. van Hunsel, L. van der Linden, K. Taxis, and E. van Puijenbroek. The Quality of Clinical Information in Adverse Drug Reaction Reports by Patients and Healthcare Professionals: A Retrospective Comparative Analysis. *Drug Safety*, 40(7): 607–614, 2017. ISSN 11791942. doi: 10.1007/s40264-017-0530-5.

[254] S. Rosenthal and K. McKeown. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4625.

[255] P. Ruch, R. Baud, and A. Geissbühler. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial Intelligence in Medicine*, 29(1):169 – 184, 2003. ISSN 0933-3657. doi: https://doi.org/10.1016/S0933-3657(03)00052-6.

[256] M. Ryu and K. Lee. Knowledge Distillation for BERT Unsupervised Domain Adaptation. *arXiv*, 2020. URL https://arxiv.org/abs/2010.11478v2.

[257] A. Sakhovskiy, Z. Miftahutdinov, and E. Tutubalina. KFU NLP Team at SMM4H 2021 Tasks: Cross-lingual and Cross-modal BERT-based Models for Adverse Drug Effects. In *Proceedings of the Sixth Social Media Mining for Health Workshop 2021*, pages 39–43, 2021. doi: 10.18653/v1/2021.smm4h-1.6.

[258] H. Sampathkumar, X.-W. Chen, and B. Luo. Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model. *BMC Medical Informatics and Decision Making*, 14, 2014.

[259] O. Sangha, G. Stucki, M. H. Liang, A. H. Fossel, and J. N. Katz. The Self-Administered Comorbidity Questionnaire: a new method to assess comorbidity for clinical and health services research. *Arthritis and rheumatism*, 49(2):156–163, 2003. ISSN 0004-3591. doi: 10.1002/ART.10993.

[260] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*, 2019.

[261] A. Sarker. A customizable pipeline for social media text normalization. *Social Network Analysis and Mining*, 7(1):45, 2017. ISSN 1869-5450. doi: 10.1007/s13278-017-0464-z.

[262] A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207, 2015. ISSN 1532-0464. doi: 10.1016/J.JBI.2014.11.002.

[263] A. Sarker and G. Gonzalez. A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. *Data in Brief*, 10:122–131, 2017. ISSN 2352-3409. doi: 10.1016/J.DIB.2016.11.056.

[264] A. Sarker and G. Gonzalez. HLP@UPenn at SemEval-2017 Task 4A: A simple, self-optimizing text classification system combining dense and sparse vectors. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 640–643, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2105.

[265] A. Sarker and G. Gonzalez-Hernandez. Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017. In *Proceedings of the Second Workshop on Social Media Mining for Health Applications (SMM4H)*, 2017.

[266] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212, 2015. ISSN 1532-0464. doi: 10.1016/J.JBI.2015.02.004.

[267] A. Sarker, K. O'Connor, R. Ginn, M. Scotch, K. Smith, D. Malone, and G. Gonzalez. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Safety*, 39(3):231–240, 2016. ISSN 0114-5916. doi: 10.1007/s40264-015-0379-4.

[268] A. Sarker, M. Belousov, J. Friedrichs, K. Hakala, S. Kiritchenko, F. Mehryary, S. Han, T. Tran, A. Rios, R. Kavuluru, B. de Bruijn, F. Ginter, D. Mahata, S. M. Mohammad, G. Nenadic, and G. Gonzalez-Hernandez. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283, 2018. ISSN 1067-5027. doi: 10.1093/jamia/ocy114.

[269] A. Sarker, S. Lakamana, W. Hogg-Bremer, A. Xie, M. A. Al-Garadi, and Y.-C. Yang. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. *Journal of the American Medical Informatics Association*, 27(8):1310–1315, 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa116.

[270] C. Seale, J. Charteris-Black, A. MacFarlane, and A. McPherson. Interviews and internet forums: a comparison of two sources of qualitative data. *Qualitative health research*, 20(5):595–606, 2010. ISSN 1049-7323. doi: 10.1177/1049732309354094.

[271] I. Segura-Bedmar, P. Martínez, R. Revert, and J. Moreno-Schneider. Exploring Spanish health social media for detecting drug effects. *BMC Medical Informatics and Decision Making*, 15(2):1–9, 2015. ISSN 14726947. doi: 10.1186/1472-6947-15-S2-S6.

[272] J. Shang, L. Liu, X. Gu, X. Ren, T. Ren, and J. Han. Learning named entity tagger using domain-specific dictionary. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2054–2064, 2020. doi: 10.18653/v1/d18-1230.

[273] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[274] P. Shenoy and A. Harugeri. Elderly patients' participation in clinical trials. *Perspectives in Clinical Research*, 6(4):184, 2015. ISSN 2229-3485. doi: 10.4103/2229-3485.167099.

[275] Y. Si, J. Wang, H. Xu, and K. Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 2019. doi: https://doi.org/10.1093/jamia/ocz096.

[276] R. Sloane, O. Osanlou, D. Lewis, D. Bollegala, S. Maskell, and M. Pirmohamed. Social media and pharmacovigilance: A review of the opportunities and challenges. *British journal of clinical pharmacology*, 80(4):910–20, 2015. ISSN 1365-2125. doi: 10.1111/bcp.12717.

[277] E. Smailhodzic, W. Hooijsma, A. Boonstra, and D. J. Langley. Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Services Research*, 16(1), 2016. ISSN 1472-6963. doi: 10.1186/s12913-016-1691-0.

[278] M. V. Smeden, C. Moons, I. Kant, and H. V. Os. Leidraad voor kwalitatieve diagnostische en prognostische toepassingen van AI in de zorg. Technical report, Ministerie van Volksgezondheid, Welzijn en Sport, 2021. URL https://www.datavoorgezondheid.nl/documenten/publicaties/2021/12/17/leidraad-kwaliteit-ai-in-de-zorg.

[279] D. Smedt, R. H. . Jaarsma, T. . Ranchor, A. V. . Van Der Meer, K. . Groenier, K. H. . Haaijer-Ruskamp, and F. M. . Denig. Coping with adverse drug events in patients with heart failure: Exploring the role of medication beliefs and perceptions. *Psychology & Health*, 27(5):570–587, 2012. doi: 10.1080/08870446.2011.605886.

[280] K. Smith, S. Golder, A. Sarker, Y. Loke, K. O'Connor, and G. Gonzalez-Hernandez. Methods to Compare Adverse Events in Twitter to FAERS, Drug Information Databases, and Systematic Reviews: Proof of Concept with Adalimumab. *Drug Safety*, 41(12):1397–1410, 2018. ISSN 11791942. doi: 10.1007/s40264-018-0707-6.

[281] L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *ArXiv*, 2018. URL https://arxiv.org/pdf/1803.09820.pdf.

[282] S. C. Sodergren, A. White, F. Efficace, M. Sprangers, D. Fitzsimmons, A. Bottomley, and C. D. Johnson. Systematic review of the side effects associated with tyrosine kinase inhibitors used in the treatment of gastrointestinal stromal tumours on behalf of the EORTC Quality of Life Group. *Critical reviews in oncology/hematology*, 91(1):35–46, 2014. ISSN 1879-0461. doi: 10.1016/J.CRITREVONC.2014.01.002.

[283] S. C. Sodergren, S. J. Wheelwright, D. Fitzsimmons, F. Efficace, M. Sprangers, P. Fayers, A. Harle, H. Schmidt, A. Bottomley, A. S. Darlington, C. Benson, A. Bredart, L. Hentschel, J. I. Arraras, G. Ioannidis, M. Leahy, I. Lugowska, O. Nicolatou-Galitis, D. Petranovic, G. E. Rohde, V. Vassiliou, and C. D. Johnson. Developing Symptom Lists for People with Cancer Treated with Targeted Therapies. *Targeted oncology*, 16 (1):95–107, 2021. ISSN 1776-260X. doi: 10.1007/S11523-020-00769-Z.

[284] C. Song, S. Zhang, N. Sadoughi, P. Xie, and E. Xing. Generalized Zero-Shot Text Classification for ICD Coding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Code*, pages 4018–4024, 2020.

[285] K. Søreide, O. M. Sandvik, J. A. Søreide, V. Giljaca, A. Jureckova, and V. R. Bulusu. Global epidemiology of gastrointestinal stromal tumours (GIST): A systematic review of population-based cohort studies. *Cancer Epidemiology*, 40:39–46, 2016. ISSN 1877783X. doi: 10.1016/j.canep.2015.10.031.

[286] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh. A systematic literature review of automated clinical coding and classification systems. *JAMIA*, 17: 646–651, 2010. doi: 10.1136/jamia.2009.001024.

[287] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger. Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International*

Journal of Information Management, 39:156–168, 2018. ISSN 02684012. doi: 10.1016/j.ijinfomgt.2017.12.002.

[288] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374, 2000.

[289] B. H. Stricker and B. M. Psaty. Detection, verification, and quantification of adverse drug reactions. *BMJ*, 329:44–47, 2004. doi: 10.1136/bmj.329.7456.44.

[290] L. Sun, K. Hashimoto, W. Yin, A. Asai, J. Li, P. Yu, and C. Xiong. Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT. *arXiv*, 2020. ISSN 23318422. URL `https://arxiv.org/abs/2003.04985`.

[291] M. Sung, H. Jeon, J. Lee, and J. Kang. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, 2020. doi: 10.18653/v1/2020.acl-main.335.

[292] D. Supranovich and V. Patsepnia. Ihs_rd: Lexical normalization for english tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 78–81, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4311.

[293] D. Sánchez, M. Batet, and A. Viejo. Utility-preserving privacy protection of textual healthcare documents. *Journal of Biomedical Informatics*, 52:189 – 198, 2014. doi: https://doi.org/10.1016/j.jbi.2014.06.008.

[294] H. Taipale, J. Schneider-Thoma, J. Pinzón-Espinosa, J. Radua, O. Efthimiou, C. H. Vinkers, E. Mittendorfer-Rutz, N. Cardoner, L. Pintor, A. Tanskanen, A. Tomlinson, P. Fusar-Poli, A. Cipriani, E. Vieta, S. Leucht, J. Tiihonen, and J. J. Luykx. Representation and Outcomes of Individuals With Schizophrenia Seen in Everyday Practice Who Are Ineligible for Randomized Clinical Trials. *JAMA Psychiatry*, 2022. ISSN 2168-622X. doi: 10.1001/JAMAPSYCHIATRY.2021.3990.

[295] B. Tang, Q. Chen, X. Wang, Y. Wu, Y. Zhang, M. Jiang, J. W. Wang, and H. Xu. Recognizing Disjoint Clinical Concepts in Clinical Text Using Machine Learning-based Methods. In *AMIA Annu Symp Proc.*, pages 1184–1193, 2015.

[296] B. Tang, J. Hu, X. Wang, and Q. Chen. Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF. *Wireless Communications and Mobile Computing*, 2018. doi: 10.1155/2018/2379208.

[297] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29 (1):24–54, 2010. ISSN 0261927X. doi: 10.1177/0261927X09351676.

[298] E. Tjong Kim Sang and F. de Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[299] G. Tortoreto, E. A. Stepanov, A. Cervone, M. Dubiel, and G. Riccardi. Affective Behaviour Analysis of On-line User Interactions: Are On-line Support Groups more Therapeutic than Twitter? In *Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 79–88, 2019. doi: 10.18653/v1/W19-3211.

[300] H. Tu, Z. Ma, A. Sun, and X. Wang. When MetaMap Meets Social Media in Healthcare: Are the Word Labels Correct? In *Asia Information Retrieval Symposium*, pages 356–362. Springer, 2016. doi: 10.1007/978-3-319-48051-0_31.

[301] Z. Tufekci. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *Proceedings of the eighth international AAAI conference on weblogs and social media.*, pages 505–514, 2014. doi: 10.1016/0022-5193(78)90170-4.

[302] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, and V. Malykh. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84:93–102, 2018. ISSN 1532-0464. doi: 10.1016/J.JBI.2018.06.006.

[303] E. Tutubalina, I. Alimova, Z. Miftahutdinov, A. Sakhovskiy, V. Malykh, and S. Nikolenko. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews. *Bioinformatics*, 37(2), 2020. doi: 10.1093/bioinformatics/btaa675.

[304] E. Tutubalina, A. Kadurin, and Z. Miftahutdinov. Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models. In *COLING 2020*, 2020. doi: 10.18653/v1/2020.coling-main.588.

[305] U.S Congress Office of Technology Assessment. *Pharmaceutical R&D: Costs, Risks, and Rewards*. U.S. Government Printing Office, Washington, DC, 1993. ISBN 0-16-041658-2. URL `https://ota.fas.org/reports/9336.pdf`.

[306] US Department of Health & Human Services. FAQs About Rare Diseases, 2021. URL `https://rarediseases.info.nih.gov/diseases/pages/31/faqs-about-rare-diseases`.

[307] US Department of Health and Human Services. Common Terminology Crtieria for Adverse Drug Events (CTCAE) version 5. Technical report, US Department of Health and Human Services, 2017. URL `https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcae_v5_quick_reference_5x7.pdf`.

[308] U.S. Food and Drug Administration (FDA). GLEEVEC (imatinib mesylate) tablets label. Technical report, U.S. Food and Drug Administration (FDA),

2008. URL https://www.accessdata.fda.gov/drugsatfda_docs/label/2008/021588s024lbl.pdf.

[309] U.S. Food and Drug Administration (FDA). SUTENT (sunitinib malate) capsules label. Technical report, U.S. Food and Drug Administration (FDA), 2011. URL https://www.accessdata.fda.gov/drugsatfda_docs/label/2011/021938s13s17s18lbl.pdf.

[310] U.S. Food and Drug Administration (FDA). STIVARGA (regorafenib) tablets label. Technical report, U.S. Food and Drug Administration (FDA), 2017. URL https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/203085s007lbl.pdf.

[311] U.S. Food and Drug Administration (FDA). FDA approves ripretinib for advanced gastrointestinal stromal tumor, 2020. URL https://www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-ripretinib-advanced-gastrointestinal-stromal-tumor.

[312] U.S. Food and Drug Administration (FDA). Grapefruit juice can affect how well some medicines work, 2021. URL https://www.fda.gov/consumers/consumer-updates/grapefruit-juice-and-some-drugs-dont-mix.

[313] U.S. National Library of Medicine. Unified Medical Language System (UMLS). URL https://www.nlm.nih.gov/research/umls/.

[314] U.S. National Library of Medicine. RxNorm, 2020. URL https://www.nlm.nih.gov/research/umls/rxnorm/.

[315] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011. ISSN 10675027. doi: 10.1136/amiajnl-2011-000203.

[316] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. ISSN 1548-7660. doi: 10.18637/JSS.V045.I03.

[317] L. V. Van De Poll-Franse, N. Horevoorts, M. V. Eenbergen, J. Denollet, J. A. Roukema, N. K. Aaronson, A. Vingerhoets, J. W. Coebergh, J. De Vries, M. L. Essink-Bot, and F. Mols. The Patient Reported Outcomes Following Initial treatment and Long term Evaluation of Survivorship registry: Scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. *European Journal of Cancer*, 47(14):2188–2194, 2011. ISSN 09598049. doi: 10.1016/j.ejca.2011.04.034.

[318] R. van der Goot and G. van Noord. Monoise: Modeling noise using a modular normalization system. *Computational Linguistics in the Netherlands Journal*, 7:129–144, 2017.

[319] M. C. van Eenbergen, L. V. van de Poll-Franse, P. Heine, and F. Mols. The Impact of Participation in Online Cancer Communities on Patient Reported Outcomes: Systematic Review. *JMIR Cancer*, 3(2), 2017. doi: 10.2196/cancer.7312.

[320] F. Van Hunsel, Susan De Waal, and L. Härmark. The contribution of direct patient reported ADRs to drug safety signals in the Netherlands from 2010 to 2015. *Pharmacoepidemiol Drug Saf*, 26:977–983, 2017. doi: 10.1002/pds.4236.

[321] J. van Stekelenborg, J. Ellenius, S. Maskell, T. Bergvall, O. Caster, N. Dasgupta, J. Dietrich, S. Gama, D. Lewis, V. Newbould, S. Brosch, C. E. Pierce, G. Powell, A. Ptaszyńska-Neophytou, A. F. Wiśniewski, P. Tregunno, G. N. Norén, and M. Pirmohamed. Recommendations for the Use of Social Media in Pharmacovigilance: Lessons from IMI WEB-RADR. *Drug Safety*, 42(12):1393–1407, 2019. ISSN 11791942. doi: 10.1007/s40264-019-00858-7.

[322] C. F. van Uden-Kraan, C. H. Drossaert, E. Taal, C. E. Lebrun, K. W. Drossaers-Bakker, W. M. Smit, E. R. Seydel, and M. A. van de Laar. Coping with somatic illnesses in online support groups: Do the feared disadvantages actually occur? *Computers in Human Behavior*, 24(2):309–324, 2008. ISSN 07475632. doi: 10.1016/j.chb.2007.01.014.

[323] C. F. van Uden-Kraan, C. H. Drossaert, E. Taal, E. R. Seydel, and M. A. van de Laar. Self-reported differences in empowerment between lurkers and posters in online patient support groups. *Journal of medical Internet research*, 10(2):1–9, 2008. ISSN 14388871. doi: 10.2196/jmir.992.

[324] C. F. van Uden-Kraan, C. H. Drossaert, E. Taal, B. R. Shaw, E. R. Seydel, and M. A. F. J. van de Laar. Empowering processes and outcomes of participation in online support groups for patients with breast cancer, arthritis, or fibromyalgia. *Qualitative Health Research*, 18(3):405–417, 2008. ISSN 1049-7323. doi: 10.1177/1049732307313429.

[325] C. F. van Uden-Kraan, C. H. Drossaert, E. Taal, E. R. Seydel, and M. A. van de Laar. Participation in online patient support groups endorses patients' empowerment. *Patient Education and Counseling*, 74(1):61–69, 2009. ISSN 07383991. doi: 10.1016/j.pec.2008.07.044.

[326] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008, 2017.

[327] S. Verberne. Context-sensitive spell checking based on word trigram probabilities. Master's thesis, Radboud University, 2002.

[328] S. Verberne, A. Batenburg, R. Sanders, and M. V. Eenbergen. Social processes of online empowerment on a cancer patient discussion form: using text mining to analyze linguistic patterns of empowerment processes. *JMIR Cancer*, 1(5), 2018. doi: 10.2196/cancer.9887.

[329] S. Verberne, A. Batenburg, R. Sanders, M. van Eenbergen, E. Das, and M. S. Lambooij. Analyzing empowerment processes among cancer patients in an online community: A text mining approach. *JMIR Cancer*, 5(1):e9887, 2019. ISSN 2369-1999. doi: 10.2196/cancer.9887.

[330] K. Verspoor, J. Cohn, S. Mniszewski, and C. Joslyn. A categorization approach to automated ontological function annotation. *Protein Science*, 15(6):1544–1549, 2006. ISSN 09618368. doi: 10.1110/ps.062184006.

[331] J. Verweij, P. G. Casali, J. Zalcberg, A. LeCesne, P. Reichardt, J.-Y. Blay, R. Issels, A. van Oosterom, P. C. Hogendoorn, M. Van Glabbeke, R. Bertulli, and I. Judson. Progression-free survival in gastrointestinal stromal tumours with high-dose imatinib: randomised trial. *The Lancet*, 364(9440):1127–1134, 9 2004. ISSN 0140-6736. doi: 10.1016/S0140-6736(04)17098-0. URL https://www.sciencedirect.com/science/article/pii/S0140673604170980?via%3Dihub.

[332] P. Vijayaraghavan and D. Roy. Modeling human motives and emotions from personal narratives using external knowledge and entity tracking. In *Proceedings of the Web Conference 2021 (WWW '21)*, pages 529–540. ACM, 2021. ISBN 978-1-4503-8312-7. doi: 10.1145/3442381.3449997.

[333] J. E. Ware and C. D. Sherbourne. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical care*, 30(6):473–483, 1992. ISSN 0025-7079. doi: 10.1097/00005650-199206000-00002.

[334] WEB-RADR 2 project, 2021. URL https://www.snomed.org/news-and-events/articles/new-collaboration-SNOMED-ICH-MedDRA.

[335] D. Weissenbacher, A. Sarker, M. Paul, and G. Gonzalez-Hernandez. Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018. In *Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task*, pages 13–16, 2018. doi: 10.18653/v1/W18-5904.

[336] D. Weissenbacher, A. Sarker, A. Klein, K. O'connor, A. Magge, and G. Gonzalez-Hernandez. Deep neural networks ensemble for detecting medication mentions in tweets. *Journal of the American Medical Informatics Association*, 26(12):1618–1626, 2019. ISSN 1527974X. doi: 10.1093/jamia/ocz156.

[337] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. Paul, and G. Gonzalez-Hernandez. Overview of the Fourth Social Media Mining for Health (#SMM4H) Shared Task at ACL 2019. In *Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, 2019. doi: 10.18653/v1/W19-3203.

[338] D. William and D. Suhartono. Text-based Depression Detection on Social Media Posts: A Systematic Literature Review. *Procedia Computer Science*, 179:582–589, jan 2021. ISSN 1877-0509. doi: 10.1016/J.PROCS.2021.01.043.

[339] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, 2019. URL https://arxiv.org/abs/1910.03771.

[340] World Health Organisation. The Safety of Medicines in Public Health Programmes: Pharmacovigilance, an essential tool. Technical report, World Health Organisation, 2006. URL https://apps.who.int/iris/handle/10665/43384.

[341] World Health Organization. The Importance of Pharmacovigilance. Technical report, World Health Organization, Geneva, 2002. URL https://apps.who.int/iris/handle/10665/42493.

[342] World Health Organization. Safety of Medicines: A guide to detecting and reporting adverse drug reactions, why health professionals need to take action. Technical report, World Health Organization, Geneva, 2002. URL https://apps.who.int/iris/handle/10665/67378.

[343] Y. Wu, B. Tang, M. Jiang, S. Moon, J. C. Denny, and H. Xu. Clinical acronym/abbreviation normalization using a hybrid approach. In *CLEF (Working Notes)*, 2013.

[344] C. C. Yang, H. Yang, L. Jiang, and M. Zhang. Social media mining for drug safety signal detection. In *Proceedings of the 2012 international workshop on Smart health and wellbeing - SHB '12*, pages 33–40, 2012. ISBN 9781450317122. doi: 10.1145/2389707.2389714.

[345] A. Yates and N. Goharian. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. In *ECIR 2013: Advances in Information Retrieval*, pages 816–819, 2013. doi: 10.1007/978-3-642-36973-5_92.

[346] S. Yeleswarapu, A. Rao, T. Joseph, V. Govindakrishnan Saipradeep, and R. Srinivasan. A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC medical informatics and decision making*, 14(13), 2014. doi: 10.1186/1472-6947-14-13.

[347] Y. Zang, B. Hou, F. Qi, Z. Liu, X. Meng, and M. Sun. Learning to attack: Towards textual adversarial attacking in real-world situations. *arXiv*, 2020. ISSN 23318422. URL https://arxiv.org/abs/2009.09192.

[348] Q. Zeng and T. Tse. Exploring and developing consuming health vocabulary. *J Am Med Inform Assoc*, 13(1):24–29, 2006. doi: 10.1197/jamia.M1761.A.

[349] B. Zhang, X. Zhang, Y. Liu, L. Cheng, and Z. Li. Matching Distributions between Model and Data: Cross-domain Knowledge Distillation for Unsupervised Domain Adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5423–5433, 2021.

[350] M. Zhang, M. Zhang, C. Ge, Q. Liu, J. Wang, J. Wei, and K. Q. Zhu. Automatic discovery of adverse reactions through Chinese social media. *Data Mining and Knowledge Discovery*, 33(4):848–870, 2019. ISSN 1573756X. doi: 10.1007/s10618-018-00610-2.

[351] Y. Zhang, J. Wang, B. Tang, Y. Wu, M. Jiang, Y. Chen, and H. Xu. UTH_CCB: A report for SemEval 2014 – Task 7 Analysis of Clinical Text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 802–806, 2014. doi: 10.3115/v1/s14-2142.

[352] X. Zhou, A. Zheng, J. Yin, R. Chen, X. Zhao, W. Xu, W. Cheng, T. Xia, and S. Lin. Context-Sensitive Spelling Correction of Consumer-Generated Content on Health Care. *JMIR Medical Informatics*, 3(3):e27, 2015. ISSN 2291-9694. doi: 10.2196/medinform.4211.

[353] M. Zolnoori, K. W. Fung, T. B. Patrick, P. Fontelo, H. Kharrazi, A. Faiola, N. D. Shah, Y. S. Shirley Wu, C. E. Eldredge, J. Luo, M. Conway, J. Zhu, S. K. Park, K. Xu, and H. Moayyed. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in brief*, 24, 2019. ISSN 2352-3409. doi: 10.1016/j.dib.2019.103838.

[354] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, K. Bontcheva, T. Cohn, and I. Augenstein. Discourse-Aware Rumour Stance Classification in Social Media Using Sequential Classifiers. *Information Processing & Management*, 54(2):273–390, 2018. doi: 10.1016/j.ipm.2017.11.009.

SUMMARY

Patients share valuable advice and experiences with their peers in online patient discussion groups. These uncensored experiences can provide a complementary perspective to that of the health professional and thereby yield novel hypotheses which could be tested in further rigorous medical research. This thesis focuses on the development of automatic extraction methods to harvest these patient experiences from online patient forums using text mining techniques. We also examine the complementary value of these patient-reported outcomes to traditional sources of medical knowledge for scientific hypothesis generation. Specifically, we focus on the extraction of adverse drug events (i.e. side effects) and coping strategies for dealing with adverse drug events.

In the first part, we investigated how spelling mistakes in medical social media messages can be reduced. We studied which unsupervised spelling correction method is most suitable for correcting spelling mistakes in medical social media data without losing valuable information due to false positives (domain-specific terms or layman's terms that are corrected because they are not in the dictionary). We also examined how posts containing patient experiences can best be distinguished from those that do not to weed out irrelevant posts. This helped to select the discussion threads that are most likely to contain adverse drug events (ADEs) and provided insight into the different types of patient experiences shared on the forum. In a third study, we showed that despite the fact that relevant posts cluster together, incorporating the structure of the conversation into state-of-the-art text classification models did not help to identify relevant posts.

In the second part, we addressed challenges presented by the extraction of text snippets containing patient-reported ADEs. We tested the efficacy of default Transformer models, including the popular BERT model, for this task and evaluated the vulnerability of BERT models to being fooled by variation in the input data. We also tackled the challenge of discontinuous entities, which can be either composite (e.g. "*hand* and *foot pain*") or disjoint (e.g. "*eyes* are feeling *dry*"). We presented a more flat, continuous representation of these entities that can benefit end-to-end extraction of ADEs.

In the third part, we showcased a novel task: the extraction of coping strategies for adverse drug responses. We presented the first ontology for coping strategies, compared the success of different conceptualizations of this task, and showed that automatically derived coping strategies from an online patient forum could be used for hypothesis generation.

In the fourth part, we described a case study on a specific patient forum for Gastro-Intestinal Stromal Tumor (GIST) patients and demonstrated the value of extracting ADE from patient forum posts for post-market drug monitoring. We showed that adverse drug events can be extracted from patient forum messages with sufficient success to enable the discovery of novel ADEs, long-term ADEs, and an indication of which ADEs are most important to patients. A comparison of these results with ADEs reported by GIST patients in a survey revealed that automatically extracted ADEs from patient forum data

can be used to select the most appropriate questionnaire for the patient population and to keep questionnaires up to date. To better understand the limitations of knowledge discovery from patient forum data, we also investigated how representative the online patient population of GIST patients is. We found that patients in relatively better condition are generally under-represented on the patient forum.

Our work offers a starting point for knowledge discovery from online patient forums and its use as a complementary data source for hypothesis generation. Future work will need to elucidate to what extent the complementary value of patient knowledge may differ between different types of disorders, such as between rare and more common disorders and between chronic and more acute disorders.

SAMENVATTING

Patiënten delen waardevolle adviezen en ervaringen met hun lotgenoten op online patiëntenfora. Deze ongecensureerde ervaringen kunnen een aanvullend perspectief bieden op dat van de zorgprofessional en nieuwe hypotheses opleveren die vervolgens gevalideerd kunnen worden met medisch onderzoek. Deze dissertatie richt zich op het ontwikkelen van automatische methoden om patiëntenervaringen uit online patiëntenfora te extraheren met behulp van text mining technieken. We onderzoeken de complementaire waarde van deze kennis voor het genereren van wetenschappelijke hypotheses ten opzichte van traditionele bronnen van medische kennis. In het bijzonder richten wij ons op de extractie van bijwerkingen en copingstrategieën die patienten gebruiken om om te gaan met hun bijwerkingen.

In het eerste deel van deze dissertatie onderzochten we hoe spelfouten in berichten op het patiëntenforum verminderd kunnen worden. We onderzochten welke unsupervised spellingcorrectiemethode het meest geschikt is voor het corrigeren van spelfouten, zonder waardevolle informatie te verliezen door valse positieven (domein-specifieke termen of termen van leken die onterecht gecorrigeerd worden omdat ze niet in het woordenboek staan). We onderzochten ook hoe we berichten met patiëntenervaringen het best kunnen onderscheiden van berichten zonder om zo irrelevante berichten uit de data te kunnen filteren. Bovendien konden we hiermee inzicht verkrijgen in de verschillende soorten ervaringen die patiënten op het forum delen. In een andere studie lieten we zien dat het meenemen van de structuur van discussiedraad tekstclassificatiemodellen niet helpt om relevante berichten te identificeren, ondanks dat relevante berichten vaak gegroepeerd zijn.

In het tweede deel gingen we in op de uitdagingen die ontstaan bij het extraheren van tekstfragmenten waar bijwerkingen in staan. We toetsten hoe goed Transformer-modellen, waaronder het populaire BERT model, werken voor deze taak. Daarnaast toetsten we de kwetsbaarheid van deze modellen voor variatie in de data: hoe makkelijk kunnen ze voor de gek gehouden worden? We presenteerden daarnaast een eenvoudige representatie voor discontinue of onderbroken concepten (bv.“pijn in handen en voeten” wat bestaat uit zowel “pijn in handen” als “pijn in voeten”) die de extractie van bijwerkingen ten goede komt.

In het derde deel introduceerden we een nieuwe taak: de extractie van coping mechanismen voor bijwerkingen van geneesmiddelen. Hoe gaan patiënten om met hun bijwerkingen en welk advies geven ze aan elkaar? We presenteerden de eerste ontologie voor coping mechanismen, vergeleken verschillende mogelijke conceptualisaties van deze taak en lieten zien dat automatisch geëxtraheerde coping strategieën uit berichten van een patiëntenforum gebruikt kunnen worden om nieuwe hypotheses te genereren.

In het vierde en laatste deel beschreven we een casus: een specifiek patiëntenforum voor Gastro-Intestinal Stromal Tumor (GIST) patiënten. We onderzochten de waarde van automatisch geëxtraheerde bijwerkingen uit dit patiëntenforum voor het monitoren van

geneesmiddelen die al op de markt zijn (ook wel post-market surveillance genoemd). We lieten zien dat onze methode niet eerder gevonden bijwerkingen en langetermijnbijwerkingen kan ontdekken. Ook kan onze methode een indicatie geven van welke bijwerkingen het meest belangrijk zijn voor de patiënt zelf. Een vergelijking van deze resultaten met de bijwerkingen die door GIST-patiënten in een enquête worden gemeld liet zien dat automatisch geëxtraheerde bijwerkingen uit patiëntenfora gebruikt kunnen worden om de meest geschikte vragenlijst te selecteren en om vragenlijsten up-to-date te houden. Tot slot hebben we onderzocht hoe representatief de online patiëntenpopulatie van GIST-patiënten is om de beperkingen van kennis uit patiëntenfora beter te overzien. Onze conclusie was dat patiënten die relatief gezonder zijn minder goed vertegenwoordigd zijn op het patiëntenforum.

Ons werk biedt aanknopingspunten voor het ontsluiten van de waardevolle kennis die gedeeld wordt op online patiëntenfora en het gebruik ervan als aanvullende informatiebron voor het genereren van hypothesen. Toekomstig werk zal moeten uitwijzen in hoeverre de complementaire waarde van patiëntenkennis verschilt voor verschillende aandoeningen, waaronder tussen zeldzame en veelvoorkomende aandoeningen en tussen acute en chronische aandoeningen.

ACKNOWLEDGEMENTS

There are many people who helped shape this thesis and who made my Ph.D. easier and more rewarding. One person without whom this thesis would not exist is my supervisor, Suzan. Thank you for helping me tame the chaotic tsunami of ideas in my head into concrete plans and for motivating me in moments of uncertainty. Your advice and steady pragmatism gave me renewed energy and purpose.

I would also like to give a big thank you to my other supervisors, Wessel, Hans, and Gerard, for their guidance and unwavering enthusiasm. I would also like to thank Abeed Sarker who helped me get to grips with the niche community of medical social media mining and Dide den Hollander and Olga Husson for giving me a valuable medical perspective on my work.

I was fortunate to be able to share my PhD struggles and victories with other PhD students around me. Prajit, thanks for transforming my initially dreary lonely office into a den of gossip and fun in my first year. Alex, thanks for dropping your own work to help me with any text mining or web server questions I might have, or just to hang out and chat. Daniela, thanks for making me feel less out of place at LIACS, going on writing retreats with me, and helping me with motivational fairies whenever I was stuck. Also, a special thank you to both of you for being amazing colleagues and friends during the pandemic. Our daily updates and shared suffering on slack kept me going. Hugo, thanks for the fun conversations on the train home when we still worked at the office, the food tips for the Hague, and the coffees in the park when we no longer worked at the office. Gineke, thank you for the cappuccinos (both inside and outside the office), going to the chocolate museum in Cologne with me, and for being the only other non-Greek on the Greek island. Iris, thanks for sharing your post-PhD wisdom during our writing retreats and all our other conversations about life, academia, and burnout.

I also would like to thank all my colleagues in the Data Science Program with a special thank you to both Wouters, Gineke, Daniela, Hugo, Alex, Manon, Marieke, Annelieke, and Shannon. Thank you for the table football competitions, pubquiz on Tuesdays, broodje kroket on Fridays, being happy for me when I got engaged although you didn't know me at all, the eternal talk about how we were going to make a talking fish, joining my murder mystery parties, and coffee chats at our illegal coffee machine. Thank you to the Greek Island crew, Antonis and Manolis, for including me even if I was not Greek.

I also want to thank the other PhD students at LIACS for the geeky and fun conversations during lunch. Sander, I know it was your job, but you really went out of your way to help me with the Data Science cluster, thank you. Another thank you goes to the other members of the "best office" at LIACS: Antonio, Gerrit-Jan, Xue, and Yuchen. I would like to thank Gerrit-Jan specifically for getting me hooked up with the supercomputers, setting up my LIACS website, and helping me with all sorts of other technical struggles at the start.

Of course, I would not have felt half as supported during this PhD without my friends and family. Thank you Corine, Dominic, Bas, Tessa, Wing, Joyce, Hanna, Dirk, and Alex for dragging me out into normal life, celebrating my victories with me, helping me with my imposter syndrome, and just being great friends who reminded me that there is life outside of and after a PhD. Thank you for my loving family who may not have always understood what I was up to, but supported me nonetheless. A special thank you to my grandma who was always asking about what I was up to even though she grew up in an era without computers, let alone machine learning. A second special thank you goes to my sister-in-law Gaby whose great design skills went into making the cover of this book.

Finally, I want to thank my husband, Ralph. Thank you for being there when things got tough, for making me feel accomplished and capable, for being my colleague during the last two years working at home, but most of all thank you for making me happy. Without you, I could not have done this.

APPENDICES

A

TECHNICAL DETAILS OF ADE EXTRACTION

In Appendix A, we will elaborate on the technical details of how we preprocessed our data (Section A.0.1), how we trained and evaluated models to extract adverse drug events (ADE) (Section A.0.2), how we trained and evaluated machine learning models to map ADEs to the medical ontology SNOMED-CT (Section A.0.3) and how we linked reported ADEs to the medication for which the patient reports them (Section A.0.4). The Python code is publicly available at `https://github.com/AnneDirkson/CHyMer`.

A.0.1. DATA PREPROCESSING

We preprocess the data with the pipeline described in Chapter 2 that includes replacing URLs and email addresses with the strings -URL- and -EMAIL- with regular expressions, lower-casing and tokenizing the text using NLTK, converting British to American English, expanding abbreviations to their full form (e.g., lol to laughing out loud) and expanding contractions (e.g., I"m to I am). Spelling mistakes were corrected using a combination of relative Levenshtein edit distance (i.e., how many insertions, deletions, and replacements are necessary to change one word into another word relative to the length of the word) and cosine similarity based on a static (or context independent) word2vec language model. A word2vec language model represents words based on how they are used, meaning that words used in similar contexts are closer together in the model and therefore have a lower cosine similarity. We excluded drug names in the FDA database of drugs[1] from spelling correction to prevent common drug names from replacing uncommon, similar drug names. Removing empty messages (567) and messages in a language other than English (1,493) left 121,516 messages. We also normalized drug names to their generic forms using the FDA database. We manually added experimental names before FDA approval for novel GIST drugs (BLU-285 for Avapritinib and DCC-2618 for Ripretinib).

[1] Downloaded from: `https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files`

	F_1	Precision	Specificity	Sensitivity/Recall
Token level performance	0.626	0.723	0.995	0.553
Entity level performance	0.716	0.739	0.998	0.695

Table A.1: The performance of the ADE extraction model on the held-out test set. Here the entity-level performance is lenient: an entity is regarded as true positive if at least one token has been retrieved correctly.

A.0.2. EXTRACTING ADES FROM TEXT

The task of extracting words from a text that contain a certain concept (like an Adverse Drug Event) is called Named Entity Recognition. For Named Entity Recognition, entities are represented using the BIO scheme (B-Beginning, I-Inside and O-Outside). By default, this representation is not able to represent entities that overlap with other entities (e.g., hand and foot pain) or are split (eyes are feeling dry). These entities are termed discontinuous entities. We converted annotated data labels to the FuzzyBIO annotation scheme (described in Chapter 7) to deal with these entity types. Discontinuous entities are transformed into continuous sequences that the BIO scheme can handle by annotating the non-entity words in between.

We make use of a state-of-the-art machine learning model for named entity recognition (BERT [84]) that has been trained on a large data set of English medical social media (EnDR BERT [303]). BERT models are a type of transfer learning model. Transfer learning models reuse a model trained on one (usually larger) data set as a starting point for training a model on another (usually smaller) data set to perform a task such as named entity recognition. For our model, we experimented with BERT models trained on biomedical text (i.e., PubmedBERT [119], BioBERT [174], and SciBERT[28]), but they did not perform as well as EnDR BERT.

The initialization of such models is stochastic (i.e., has a degree of randomness). This can result in suboptimal models [336]. To reduce this effect and create a more robust model, we use an ensemble of 10 models trained with different initialization seeds (1, 2, 4, 8, 16, 32, 64, 128, 256 and 512) following Weissenbacher et al. [336] and Miftahutdinov and Tutubalina [206]. We split our labeled data into training (80%), a validation (penultimate most recent 10% of the data), and a test set (most recent 10% of the data). We added a second publicly available data set of patient forum texts labeled for ADEs (CADEC [151]) to the training set. We also tried using PsyTAR [353] to increase the amount of data, but this was not beneficial. For each of the 10 models, we train for 3 or 4 epochs based on the results of the model on the first validation data. We use a one-cycle learning rate (LR) policy (max LR of 5^{-5}) to train the models. We average the output of the 10 models using majority voting.

Table A.1 reports the performance of the extraction of ADEs from text. The metrics used to calculate performance are recall, precision, and the F_1 score. Recall is the percentage of true positives that have been found. Precision is the percentage of true positives among the retrieved instances. The F_1 score is a measure of the overall performance: it is the harmonic mean of precision and recall.

Here, tokens are relevant words that are part of an ADE. Our algorithm was able to retrieve 55.3% of all true positive tokens (“Recall”) in a held-out test set and 72.3% of

the retrieved tokens are true positives ("Precision"). An entity is another term for the full concept e.g., "pain in chest" is an entity while "pain", "in" and "chest" are the tokens belonging to the entity. Our algorithm was able to retrieve 52.3% of all entities fully and 16.6% partially. On average, 69.5% of all retrieved entities were true positives. With this performance, our model performs better than state-of-the-art models on this task [194, 337]. It still falls below human performance (average mutual $F_1 = 0.80$).

In addition to good performance, a model needs to be able to find entities that it has not seen in the training data [304]. We find that 40.2% of the true positive entities that our model finds are not present in the training data, indicating that our model is able to find novel entities in unlabeled data.

A.0.3. ADE NORMALIZATION

Normalization of adverse drug responses is the mapping of the text containing the ADE to concepts in an ontology (e.g., "cannot sleep" to Insomnia in SNOMED-CT). We use the current state-of-the-art method BioSyn [291] for normalizing the entities. We used the default parameters of BioSyn. BioSyn uses BioBERT [174] (a BERT model trained on biomedical text) to rank all possible concept labels for an extracted ADE. The highest ranking label is selected. As was done in Sung et al. [291], we split composite mentions to separate mentions using heuristic rules by D' Souza and Ng [70] (e.g., non-familial breast and ovarian cancers into non-familial breast cancer and ovarian cancers).

Our data does not contain annotations for normalization (i.e., the relevant concept IDs for each ADE mention). We rely on three publicly available data sets for training our normalization model: CADEC [151], PsyTAR [353] and the Clinical Findings subset of the COMETA corpus [20].

We choose a curated subset of SNOMED, the CORE Problem List Subset as our target ontology. We try to map the concepts in the data sets to synonymous concepts in the CORE subset if possible. We do so by checking for a direct mapping in the community-based mappings in BioPortal [220] between the original concept. We also map the concept to its parent if the parent is in the CORE (e.g., "moderate anxiety" to "anxiety"). As target concepts, we include all concepts of the CORE subset. SNOMED concepts present in the training data that could not be mapped to a CORE concept and SNOMED concepts present in the registration trial data that could not be mapped to a CORE concept as candidates. We also removed all concepts that are not in the Clinical Findings of SNOMED CT (e.g., procedures like knee replacement). This results in a total of 5813 concepts. To employ the BioSyn method, we need to collect all synonyms of the target SNOMED concepts. Synonyms for each concept are retrieved from the community based mappings in BioPortal [220] using the REST API and from the UMLS using pymedtermino [170].

The performance of the normalization model is shown in Table A.2. On average, the model could accurately label 64.5% of the ADEs when tested on a different data set than those on which the model was trained. For an additional 14.6% of the cases, the correct label was included in the top 5. We manually inspected these cases and found that 53 of 100 randomly selected cases, we would consider the first label to be correct or even better than the gold label. For example, the mention "severe abdominal pain" has the gold label "painful" (i.e., the label given by humans) and the predicted label "abdominal pain". Moreover, we inspected 100 random cases in which the correct label was not included in

Trained on	Tested on	Acc @1	Acc @5
CADEC & COMETA	PsyTAR	0.586	0.771
COMETA & PsyTAR	CADEC	0.663	0.807
CADEC & PsyTAR	COMETA	0.688	0.795
		0.645	**0.791**

Table A.2: The performance of our normalization model on a held-out data set. As the normalization model provides a ranking of candidate labels, Acc @1 and Acc @5 indicate the percentage of cases with the correct label in the top 1 and top 5, respectively. The bold numbers indicate mean values

Category	Frequency	Example
Correct concept	67	-
Extraction errors	22	“feet”, “nose”, “losing”
Predicted concept is related	6	“kidney issues” instead of “nephrosis”
No SNOMED equivalent	2	“comfy eyes, woozy face”
Wrong but no clear reason	2	

Table A.3: Manual analysis of 100 randomly selected found ADEs in the GIST data

the top 5 and found that for 36 of those we would consider the predicted label as correct. Thus, the performance in Table A.2 is likely an underestimation.

One concern is the propagation of errors in the pipeline (i.e., errors from extraction will influence normalization). Previous work has shown that ADE normalization is mainly hampered by errors made during extraction [337]. To evaluate the pipeline end-to-end, we manually inspect 100 of the ADEs found in the GIST data. As can be seen in Table A.3, we find that 67 of the 100 cases are correct, while 22 of the 100 are incorrect due to extraction errors prior to normalization. Thus, extraction appears to still be the major bottleneck.

Another concern is that the normalization model should be able to predict new types of ADEs that are absent in the training data. The BioSyn model should theoretically be able to do so because all the concepts of the SNOMED CT are considered as possible targets for mapping. Our normalization model is indeed able to predict classes that were not part of the training data at an Accuracy@1 of 0.417 on average and an Accuracy@5 of 0.612 on average for the external data sets. On our own GIST data, we also see empirically that 15.0% of the predicted concepts are not part of the training data.

Manual analysis of the predicted concepts in the GIST data revealed that some ADEs for tyrosine kinase inhibitors (TKIs) (e.g. split nails, hair color change, and hand-foot syndrome) were not included in the target concepts. We manually added 5 concepts and 2 synonyms to existing concepts to improve normalization.

A.0.4. LINKING ADES TO MEDICATION

First, we identify all drugs mentioned in each message using a dictionary based on the RxNORM [313]. During preprocessing, we already converted all brand names to their generic equivalents (e.g., Gleevec to Imatinib).

We use heuristic rules to determine which drug is linked to each ADE. Whenever possible, we select the drug mentioned prior to the ADE in the message. If there is

none, we select the drug mentioned after the ADE in the message. If there are no drugs mentioned in the message, we select the first drug mentioned in the conversational thread prior to the message. These rules were determined based on further manual annotation of our annotated subset by the first author. In some cases, it was not clear which drug the patient believed was causing the ADE and these cases were excluded. In this data set, our rules were 93% accurate if we restricted the possible drugs for linking to a predetermined list (Imatinib, Sunitinib, Regorafenib, Avapritinib, Ripretinib, Nilotinib, Pazopanib, Ponatinib, Sorafenib).

thread_id	post_id	word	tag
7581	155	I	O
7581	155	really	O
7581	155	do	O
7581	155	not	O
7581	155	know	O
7581	155	what	O
7581	155	I	O
7581	155	would	O
7581	155	do	O
7581	155	without	O
7581	155	you	O
7581	155	guys	O
7581	155	!!	O
7581	155	I	O
7581	155	started	O
7581	155	sunitinib	O
7581	155	12	O
7581	155	days	O
7581	155	ago	O
7581	155	and	O
7581	155	now	O
7581	155	have	O
7581	155	some	O
7581	155	crazy	O
7581	155	rash	B-ADR
7581	155	all	I-ADR
7581	155	over	I-ADR
7581	155	my	I-ADR
7581	155	chest	I-ADR
7581	155	and	I-ADR
7581	155	back	I-ADR
7581	155	.	O
7581	155	They	O
7581	155	are	O
7581	155	little	O
7581	155	red	B-ADR
7581	155	elevated	I-ADR
7581	155	bumps	I-ADR
7581	155	all	O
7581	155	over	O
7581	155	that	O
7581	155	itch	O
7581	155	.	O
7581	155	Has	O
7581	155	anyone	O
7581	155	else	O
7581	155	had	O

7581	155	this	O
7581	155	issue	O
7581	155	?	O
7581	156	My	O
7581	156	husband	O
7581	156	has	O
7581	156	it	O
7581	156	but	O
7581	156	not	O
7581	156	itchy	O
7581	156	and	O
7581	156	his	O
7581	156	is	O
7581	156	from	O
7581	156	imatinib	O
7581	157	"	O
7581	157	I	O
7581	157	did	O
7581	157	,	O
7581	157	it	O
7581	157	only	O
7581	157	showed	O
7581	157	up	O
7581	157	on	O
7581	157	my	O
7581	157	legs	O
7581	157	and	O
7581	157	but	O
7581	157	it	O
7581	157	did	O
7581	157	not	O
7581	157	itch	O
7581	157	.	O
7581	157	I	O
7581	157	went	O
7581	157	to	O
7581	157	the	O
7581	157	emergency	O
7581	157	room	O
7581	157	,	O
7581	157	just	O
7581	157	in	O
7581	157	case	O
7581	157	it	O
7581	157	was	O
7581	157	an	O
7581	157	allergic	O
7581	157	reaction	O
7581	157	,	O

7581	157	but	O
7581	157	it	O
7581	157	ended	O
7581	157	up	O
7581	157	being	O
7581	157	“	O
7581	157	nothing	O
7581	157	”.	O
7581	158	Yes	O
7581	158	I	O
7581	158	had	O
7581	158	some	O
7581	158	of	O
7581	158	that	O
7581	158	plus	O
7581	158	it	O
7581	158	dried	B-ADR
7581	158	out	I-ADR
7581	158	my	I-ADR
7581	158	hands	I-ADR
7581	158	and	I-ADR
7581	158	feet	I-ADR
7581	158	I	O
7581	158	fought	O
7581	158	that	O
7581	158	with	O
7581	158	immunotherapy	O
7581	158	by	O
7581	158	putting	O
7581	158	it	O
7581	158	on	O
7581	158	my	O
7581	158	feet	O
7581	158	then	O
7581	158	socks	O
7581	158	and	O
7581	158	put	O
7581	158	it	O
7581	158	on	O
7581	158	my	O
7581	158	hands	O
7581	158	with	O
7581	158	cotton	O
7581	158	gloves	O
7581	158	with	O
7581	158	the	O
7581	158	blue	O
7581	158	plastic	O
7581	158	gloves	O

7581	158	on	O
7581	158	top	O
7581	159	My	O
7581	159	mouth	B-ADR
7581	159	is	I-ADR
7581	159	constantly	I-ADR
7581	159	on	I-ADR
7581	159	fire	I-ADR
7581	159	too	O
7581	159	.	O
7581	159	Stuff	O
7581	159	that	O
7581	159	is	O
7581	159	not	O
7581	159	even	O
7581	159	spicy	O
7581	159	burns	O
7581	159	!!!	O
7581	159	I	O
7581	159	was	O
7581	159	just	O
7581	159	trying	O
7581	159	to	O
7581	159	eat	O
7581	159	some	O
7581	159	cheese	O
7581	159	Doritos	O
7581	159	with	O
7581	159	melted	O
7581	159	colby	O
7581	159	jack	O
7581	159	cheese	O
7581	159	on	O
7581	159	them	O
7581	159	and	O
7581	159	my	O
7581	159	mouth	O
7581	159	is	O
7581	159	on	O
7581	159	fire	O
7581	159	.	O
7581	159	My	O
7581	159	first	O
7581	159	9	O
7581	159	days	O
7581	159	on	O
7581	159	it	O
7581	159	were	O
7581	159	really	O

7581	159	good	O
7581	159	but	O
7581	159	I	O
7581	159	think	O
7581	159	the	O
7581	159	sunitinib	O
7581	159	side	O
7581	159	effects	O
7581	159	are	O
7581	159	way	O
7581	159	worse	O
7581	159	than	O
7581	159	imatinib	O
7581	159	!!	O
7581	160	"	O
7581	160	Yes	O
7581	160	I	O
7581	160	get	O
7581	160	rash	B-ADR
7581	160	on	I-ADR
7581	160	my	I-ADR
7581	160	neck	I-ADR
7581	160	and	I-ADR
7581	160	chest	I-ADR
7581	160	,	O
7581	160	some	O
7581	160	mornings	O
7581	160	I	O
7581	160	wake	O
7581	160	with	O
7581	160	eyes	B-ADR
7581	160	so	I-ADR
7581	160	swollen	I-ADR
7581	160	I	O
7581	160	can	O
7581	160	hardly	O
7581	160	see	O
7581	160	"	O
7581	161	Had	O
7581	161	rash	B-ADR
7581	161	and	O
7581	161	it	O
7581	161	took	O
7581	161	a	O
7581	161	bit	O
7581	161	for	O
7581	161	me	O
7581	161	to	O
7581	161	adjust	O

7581	161	to	O
7581	161	sunitinib	O
7581	161	-	O
7581	161	rash	O
7581	161	finally	O
7581	161	gone	O
7581	162	"	O
7581	162	My	O
7581	162	wife	O
7581	162	had	O
7581	162	similar	O
7581	162	issues	O
7581	162	when	O
7581	162	she	O
7581	162	initially	O
7581	162	started	O
7581	162	on	O
7581	162	imatinib	O
7581	162	,	O
7581	162	but	O
7581	162	went	O
7581	162	away	O
7581	162	after	O
7581	162	1	O
7581	162	.	O
7581	162	5	O
7581	162	weeks	O
7581	162	.	O
7581	162	What	O
7581	162	helped	O
7581	162	was	O
7581	162	that	O
7581	162	her	O
7581	162	oncologist	O
7581	162	prescribed	O
7581	162	some	O
7581	162	steroid	O
7581	162	,	O
7581	162	and	O
7581	162	also	O
7581	162	I	O
7581	162	had	O
7581	162	her	O
7581	162	taking	O
7581	162	Epsom	O
7581	162	Salt	O
7581	162	with	O
7581	162	Baking	O
7581	162	Powder	O

7581	162	baths	O
7581	162	.	O
7581	162	It	O
7581	162	helped	O
7581	162	pull	O
7581	162	toxins	O
7581	162	out	O
7581	162	of	O
7581	162	the	O
7581	162	skin	O
7581	162	and	O
7581	162	relieve	O
7581	162	itching	O
7581	162	and	O
7581	162	discomfort	O
7581	162	almost	O
7581	162	immediately	O
7581	162	.	O
7581	162	She	O
7581	162	took	O
7581	162	these	O
7581	162	baths	O
7581	162	x2	O
7581	162	per	O
7581	162	day	O
7581	162	."	O

Table A.4: Example annotation of NER data for ADE extraction

B

SUPPLEMENTARY FILES FOR CHAPTER 8

thread_id	post_id	word	tag	concept
5065	1873	"	O	-
5065	1873	On	O	-
5065	1873	day	O	-
5065	1873	4	O	-
5065	1873	of	O	-
5065	1873	imatinib	O	-
5065	1873	,	O	-
5065	1873	was	O	-
5065	1873	very	O	-
5065	1873	nauseous	B-ADR	-
5065	1873	all	O	-
5065	1873	day	O	-
5065	1873	.	O	-
5065	1873	Previous	O	-
5065	1873	days	O	-
5065	1873	I	O	-
5065	1873	had	O	-
5065	1873	taken	O	-
5065	1873	the	O	-
5065	1873	imatinib	O	-
5065	1873	with	O	-
5065	1873	yogurt	O	-
5065	1873	.	O	-
5065	1873	But	O	-
5065	1873	was	O	-
5065	1873	out	O	-
5065	1873	last	O	-
5065	1873	night	O	-
5065	1873	so	O	-
5065	1873	I	O	-
5065	1873	took	O	-
5065	1873	it	O	-
5065	1873	with	O	-
5065	1873	something	O	-
5065	1873	else	O	-
5065	1873	.	O	-
5065	1873	Going	O	-
5065	1873	to	O	-
5065	1873	try	O	-
5065	1873	yogurt	O	-
5065	1873	again	O	-
5065	1873	tonight	O	-
5065	1873	and	O	-
5065	1873	see	O	-
5065	1873	if	O	-
5065	1873	it	O	-
5065	1873	makes	O	-

5065	1873	a	O	-
5065	1873	difference	O	-
5065	1873	.	O	-
5065	1873	Only	O	-
5065	1873	other	O	-
5065	1873	side	O	-
5065	1873	effect	O	-
5065	1873	is	O	-
5065	1873	cold	B-ADR	-
5065	1873	hands	I-ADR	-
5065	1873	and	I-ADR	-
5065	1873	feet	I-ADR	-
5065	1873	while	O	-
5065	1873	the	O	-
5065	1873	rest	O	-
5065	1873	of	O	-
5065	1873	the	O	-
5065	1873	body	O	-
5065	1873	is	O	-
5065	1873	warm	O	-
5065	1873	.	O	-
5065	1873	These	O	-
5065	1873	side	O	-
5065	1873	effects	O	-
5065	1873	are	O	-
5065	1873	workable	O	-
5065	1873	!	O	-
5065	1873	Do	O	-
5065	1873	you	O	-
5065	1873	take	O	-
5065	1873	it	O	-
5065	1873	with	O	-
5065	1873	a	O	-
5065	1873	certain	O	-
5065	1873	food	O	-
5065	1873	every	O	-
5065	1873	night	O	-
5065	1873	?"	O	-
5065	1874	I	O	-
5065	1874	have	O	-
5065	1874	been	O	-
5065	1874	taking	O	-
5065	1874	mine	O	-
5065	1874	at	O	-
5065	1874	night	O	-
5065	1874	around	O	-
5065	1874	11pm	O	-
5065	1874	with	O	-
5065	1874	a	O	-

5065	1874	couple	O	-
5065	1874	cookies	O	-
5065	1874	and	O	-
5065	1874	a	O	-
5065	1874	large	O	-
5065	1874	glass	O	-
5065	1874	of	O	-
5065	1874	water	O	-
5065	1874	and	O	-
5065	1874	then	O	-
5065	1874	I	O	-
5065	1874	drink	O	-
5065	1874	a	O	-
5065	1874	bottle	O	-
5065	1874	of	O	-
5065	1874	ensure	O	-
5065	1874	after	O	-
5065	1874	it	O	-
5065	1874	.	O	-
5065	1874	It	O	-
5065	1874	seems	O	-
5065	1874	to	O	-
5065	1874	work	O	-
5065	1874	I	O	-
5065	1874	have	O	-
5065	1874	not	O	-
5065	1874	felt	O	-
5065	1874	nauseous	O	-
5065	1874	just	O	-
5065	1874	a	O	-
5065	1874	lot	O	-
5065	1874	of	O	-
5065	1874	tummy	O	-
5065	1874	rumbling	O	-
5065	1874	.	O	-
5065	1874	it	O	-
5065	1874	will	O	-
5065	1874	be	O	-
5065	1874	my	O	-
5065	1874	6th	O	-
5065	1874	day	O	-
5065	1874	tonight	O	-
5065	1875	Hey	O	-
5065	1875	NAME	O	-
5065	1875	...	O	-
5065	1875	I	O	-
5065	1875	take	O	-
5065	1875	mine	O	-
5065	1875	about	O	-

5065	1875	an	O	-
5065	1875	hour	O	-
5065	1875	before	O	-
5065	1875	bedtime	O	-
5065	1875	.	O	-
5065	1875	I	O	-
5065	1875	have	O	-
5065	1875	tried	O	-
5065	1875	numerous	O	-
5065	1875	things	O	-
5065	1875	to	O	-
5065	1875	see	O	-
5065	1875	which	O	-
5065	1875	works	O	-
5065	1875	best	O	-
5065	1875	.	O	-
5065	1875	I	O	-
5065	1875	am	O	-
5065	1875	now	O	-
5065	1875	taking	O	-
5065	1875	my	O	-
5065	1875	imatinib	O	-
5065	1875	with	O	-
5065	1875	a	O	-
5065	1875	glass	O	-
5065	1875	of	O	-
5065	1875	dark	B-STR	CS06033
5065	1875	chocolate	I-STR	CS06033
5065	1875	almond	I-STR	CS06033
5065	1875	milk	I-STR	CS06033
5065	1875	with	O	-
5065	1875	much	O	-
5065	1875	success	O	-
5065	1875	.	O	-
5065	1875	Dark	B-STR	CS05345
5065	1875	chocolate	I-STR	CS05345
5065	1875	also	O	-
5065	1875	helps	O	-
5065	1875	with	O	-
5065	1875	nausea	B-ADR	-
5065	1875	and	O	-
5065	1875	the	O	-
5065	1875	almond	O	-
5065	1875	milk	O	-
5065	1875	has	O	-
5065	1875	lots	O	-
5065	1875	do	O	-
5065	1875	health	O	-
5065	1875	benefits	O	-

5065	1875	.	O	-
5065	1875	I	O	-
5065	1875	try	O	-
5065	1875	to	O	-
5065	1875	stay	B-STR	CS03264
5065	1875	away	I-STR	CS03264
5065	1875	from	I-STR	CS03264
5065	1875	dairy	I-STR	CS03264
5065	1875	as	O	-
5065	1875	much	O	-
5065	1875	as	O	-
5065	1875	possible	O	-

Table B.1: Example of real annotated data for NER and entity linking of coping strategies.

thread_id	Text	Label*
5065	'" On day 4 of imatinib , was very **nauseous** all day . Previous days I had taken the imatinib with yogurt . But was out last night so I took it with something else . Going to try yogurt again tonight and see if it makes a difference . Only other side effect is cold hands and feet while the rest of the body is warm . These side effects are workable ! Do you take it with a certain food every night ?" ', 'I have been taking mine at night around 11pm with a couple cookies and a large glass of water and then I drink a bottle of ensure after it . It seems to work I have not felt nauseous just a lot of tummy rumbling . it will be my 6th day tonight ', 'Hey NAME ... I take mine about an hour before bedtime . I have tried numerous things to see which works best . I am now taking my imatinib with a glass of dark chocolate almond milk with much success . Dark chocolate also helps with **nausea** and the almond milk has lots do health benefits . I try to ***stay away from dairy*** as much as possible'	1
5065	'" On day 4 of imatinib , was very nauseous all day . Previous days I had taken the imatinib with yogurt . But was out last night so I took it with something else . Going to try yogurt again tonight and see if it makes a difference . Only other side effect is **cold hands and feet** while the rest of the body is warm . These side effects are workable ! Do you take it with a certain food every night ?" ', 'I have been taking mine at night around 11pm with a couple cookies and a large glass of water and then I drink a bottle of ensure after it . It seems to work I have not felt nauseous just a lot of tummy rumbling . it will be my 6th day tonight ', 'Hey NAME ... I take mine about an hour before bedtime . I have tried numerous things to see which works best . I am now taking my imatinib with a glass of dark chocolate almond milk with much success . Dark chocolate also helps with nausea and the almond milk has lots do health benefits . I try to ***stay away from dairy*** as much as possible'	0

5065	'" On day 4 of imatinib , was very **nauseous** all day . Previous days I had taken the imatinib with yogurt . But was out last night so I took it with something else . Going to try yogurt again tonight and see if it makes a difference . Only other side effect is cold hands and feet while the rest of the body is warm . These side effects are workable ! Do you take it with a certain food every night ?" ', 'I have been taking mine at night around 11pm with a couple cookies and a large glass of water and then I drink a bottle of ensure after it . It seems to work I have not felt nauseous just a lot of tummy rumbling . it will be my 6th day tonight ', 'Hey NAME ... I take mine about an hour before bedtime . I have tried numerous things to see which works best . I am now taking my imatinib with a glass of dark chocolate almond milk with much success . Dark chocolate also helps with **nausea** and the almond milk has lots do health benefits . I try to stay away from dairy as much as possible ', 'I take mine nightly after a full meal normally diner but if I take later I normally have a peanut butter sandwich and then my pill . Sometimes I will have a couple Heresy kisses after taking my pill have had little or no nausea since I started my imatinib almost 9 months ago . ', 'My husband has been taking 400 mg for 6 years . He started breaking it in 1 / 2 and ***taking one half in the morning after breakfast and the other half at night with dinner*** . That seems to have helped his **nausea**.'	1
5065	'" On day 4 of imatinib , was very nauseous all day . Previous days I had taken the imatinib with yogurt . But was out last night so I took it with something else . Going to try yogurt again tonight and see if it makes a difference . Only other side effect is **cold hands and feet** while the rest of the body is warm . These side effects are workable ! Do you take it with a certain food every night ?" ', 'I have been taking mine at night around 11pm with a couple cookies and a large glass of water and then I drink a bottle of ensure after it . It seems to work I have not felt nauseous just a lot of tummy rumbling . it will be my 6th day tonight ', 'Hey NAME ... I take mine about an hour before bedtime . I have tried numerous things to see which works best . I am now taking my imatinib with a glass of dark chocolate almond milk with much success . Dark chocolate also helps with nausea and the almond milk has lots do health benefits . I try to stay away from dairy as much as possible ', 'I take mine nightly after a full meal normally diner but if I take later I normally have a peanut butter sandwich and then my pill . Sometimes I will have a couple Heresy kisses after taking my pill have had little or no nausea since I started my imatinib almost 9 months ago . ', 'My husband has been taking 400 mg for 6 years . He started breaking it in 1 / 2 and ***taking one half in the morning after breakfast and the other half at night with dinner***. That seems to have helped his nausea .'	0

Table B.2: Example of real annotated data for CS-ADE relation extraction. *1 indicates true CS-ADE relation. **Bold** indicates the ADE mentions and ***bold italic*** indicates the coping strategy (CS).

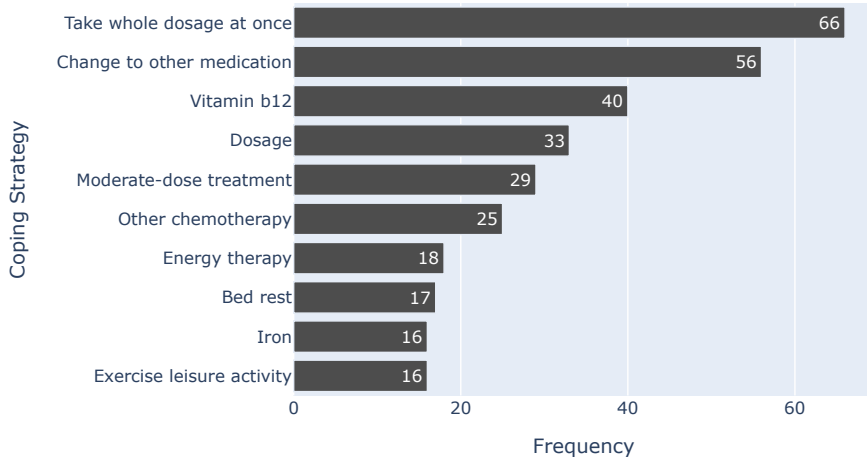


Figure B.1: **Top 10 coping strategies for Fatigue.** Manual examination of messages shows that patients recommend taking B12 or iron pills as a supplement. They also recommend naps ("bed rest"). Energy therapy appears to be a false positive and concerns messages about energy levels. Strategies regarding dosage ("Dosage", "Change to other medication", "Take whole dosage at once", "Moderate-dose treatment", "Other chemotherapy") do not appear to be about fatigue although they do relate to dosage.

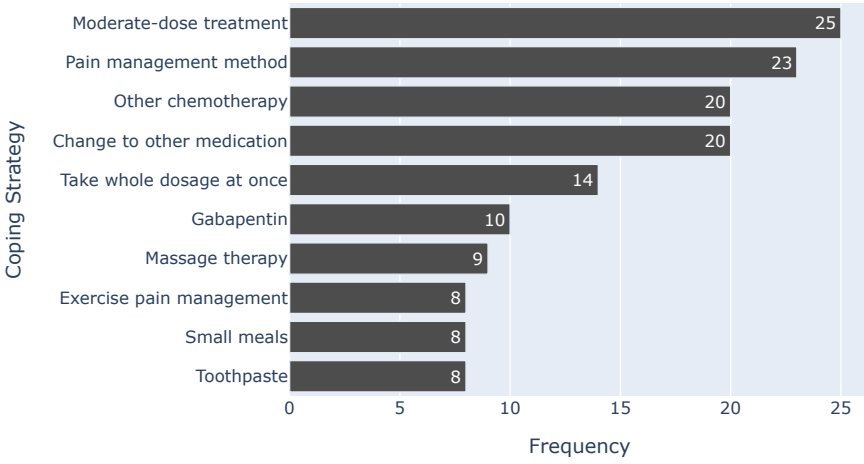


Figure B.2: **Top 10 coping strategies for Pain.** Manual examination of messages shows that patients recommend reducing the dosage of the primary medication ("moderate-dose treatment"), using gabapentin ("Gabapentin") or other pain medication ("Pain management method"), getting a massage ("massage therapy") or exercising ("Exercise pain management"). Small meals and toothpaste do concern messages around these topics but do not appear to be about pain management. Other categories related to dosage ("Other chemotherapy", "Take whole dosage at once") are not insightful.

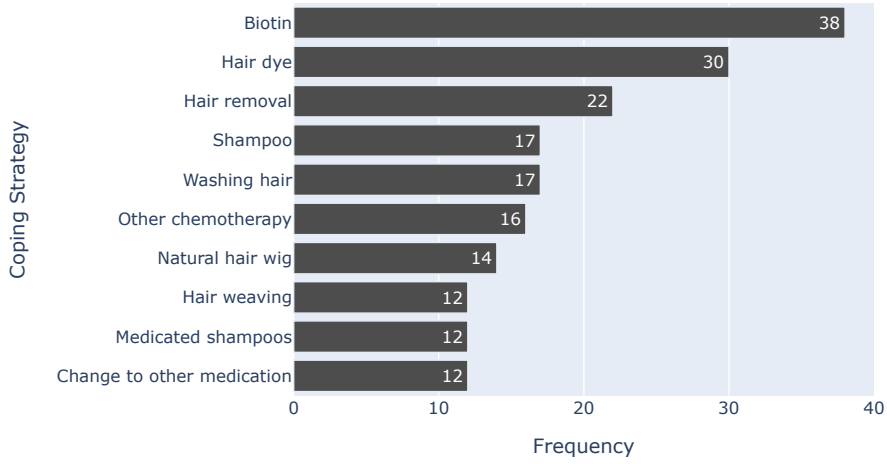


Figure B.3: **Top 10 coping strategies for Alopecia (hair loss).** Manual examination of messages reveals several true positives: Patients recommend using Biotin, using special shampoos, washing one's hair less, and wearing a wig. Other categories (e.g. "Hair dye", "Hair removal" and "Hair weaving") are false positives. The category "Hair dye" specifically mainly concerns messages where patients relay that their hair color has changed due to medication.

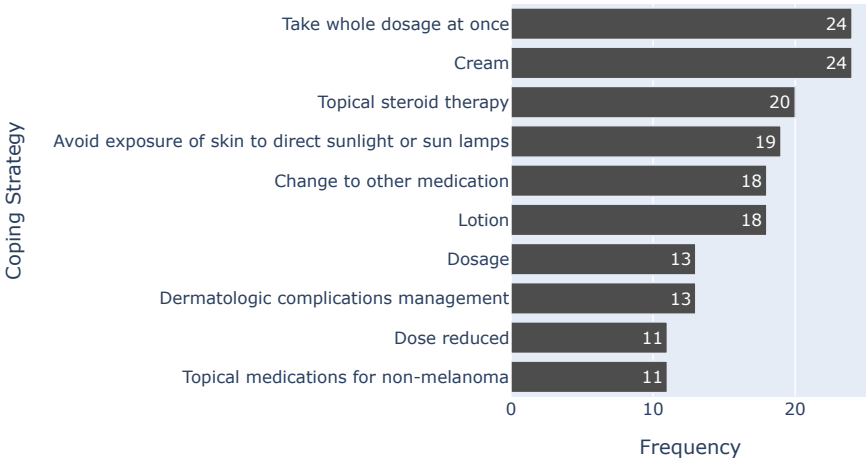


Figure B.4: **Top 10 coping strategies for Eruption (Rash).** Manual examination of messages shows that patients recommend cream, lotion, steroid creams, seeing a dermatologist, staying out of the sun and using sunscreen. The category "taking whole dosage at once" mainly contains the advice to the contrary i.e. split the dosage

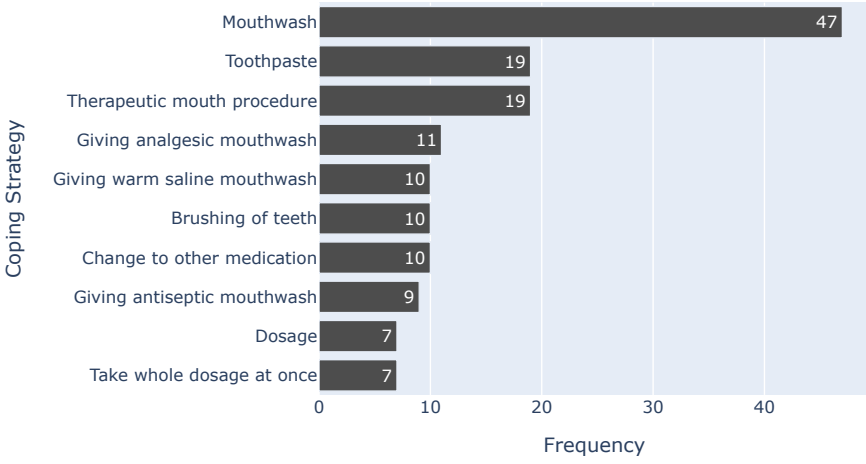


Figure B.5: **Top 10 coping strategies for a painful mouth.** Manual analysis of the messages relating to these coping strategies showed that patients recommend certain mouthwashes (e.g. magic mouthwash), using or avoiding certain toothpastes, and rinsing with saline water.

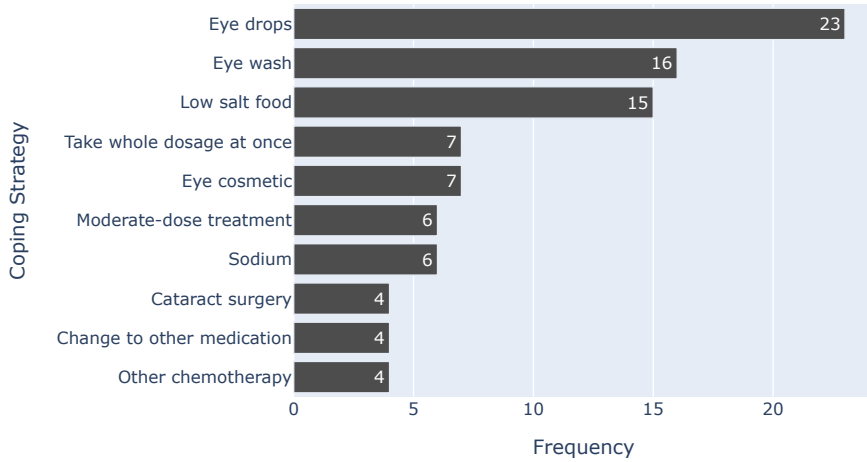


Figure B.6: **Top 10 coping strategies for Periorbital Edema.** Manual examination of messages shows patients recommend certain eye products including eye drops, or eye washes. They also discuss home remedies like cucumber and cotton pads with chamomile for on the eyes. Low salt food is also recommended.

C

SUPPLEMENTARY TABLES FOR CHAPTER 10

	Sunitinib (n=6)	Regorafenib (n=6)	Ripretinib (n=3)
Age (mean ± SD (range))	74.4 ± 8.0 (64-86)	65.5 ± 4.3 (60-71)	64.9 ± 4.6 (60-69)
Time since diagnosis in years (mean ± SD (range))	6.0 ±2.1 (3.8-9.4)	5.8 ± 1.8 (3.8 – 8.2)	3.4 ± 1.2 (2.1-4.6)
Sex			
– Male	2 (33.3%)	5 (83.3%)	2 (66.7%)
– Female	4 (66.7%)	1 (16.7%)	1 (33.3%)
Highest formal education			
– Primary school only	0	1 (16.7%)	0
– High school	1 (16.7%)	2 (33.3%)	0
– College or university	5 (83.3%)	3 (50%)	3 (100%)
Relationship status			
– Single	0	0	0
– Married/relationship	4 (66.7%)	5 (83.3%)	3 (100%)
– Separated/divorced	0	0	0
– Widowed	2 (33.3%)	1 (16.7%)	0
Comorbidities			
– None	2 (33.3%)	3 (50%)	1 (33.3%)
– One	1 (16.7%)	2 (33.3%)	1 (33.3%)
– Two or more	3 (50%)	1 (16.7%)	1 (33.3%)

Table C.1: Patient characteristics from the survey study for different tyrosine kinase inhibitors than imatinib

	Sunitinib (n=6)	Regorafenib (n=6)	Ripretinib (n=3)
Symptoms	Prevalence (%)	Prevalence (%)	Prevalence (%)
Swelling of the face or around the eyes	3 (50)	2 (33)	0 (0)
Swelling in any part of the body	1 (18)	1 (18)	0 (0)
Muscle aches, pains, or cramps	4 (67)	4 (67)	3 (100)
Aches or pains in joints	4 (67)	2 (33)	1 (33)
Food and drink tasting different from usual	5 (83)	4 (67)	0 (0)
Pain or soreness in mouth	5 (83)	2 (33)	0 (0)
Indigestion or heartburn	5 (83)	1 (18)	1 (33)
Skin problems (e.g. itchy skin, dry skin, skin discoloration)	5 (83)	4 (67)	2 (67)
Hand-foot syndrome	3 (50)	3 (50)	0 (0)
Problems because of changed appearance	2 (33)	0 (0)	3 (100)
Feeling confused	1 (18)	1 (18)	0 (0)
Trouble speaking	2 (33)	1 (18)	0 (0)
Auditory hallucinations	0 (0)	0 (0)	0 (0)
Visual hallucinations	1 (18)	0 (0)	0 (0)
Shortness of breath	1 (18)	4 (67)	1 (33)
Pain	4 (67)	3 (50)	2 (67)
Feeling weak	5 (83)	4 (67)	1 (33)
Appetite loss	4 (67)	2 (33)	1 (33)
Nausea	5 (83)	1 (18)	0 (0)
Vomiting	2 (33)	1 (18)	0 (0)
Constipation	3 (50)	1 (18)	2 (67)
Diarrhea	4 (67)	2 (33)	1 (33)
Fatigue	6 (100)	5 (83)	3 (100)
Problems with concentrating	3 (50)	1 (18)	1 (33)
Problems with remembering things	3 (50)	2 (33)	1 (33)

Table C.2: Prevalence scores for symptoms in the survey study for different tyrosine kinase inhibitors (TKI) than imatinib. Prevalence is based on percentage of patients with this symptom out of the total number of patients taking each TKI.

	Sunitinib (n=6)	Regorafenib (n=6)	Ripretinib (n=3)
Symptoms	Prevalence (%)	Prevalence (%)	Prevalence (%)
Fatigue	184 (8.0)	117 (9.5)	40 (12.6)
Nausea	111 (4.8)	35 (2.8)	14 (4.4)
Cramp	32 (1.4)	30 (2.4)	14 (4.4)
Disorder of skin	59 (2.6)	36 (2.9)	12 (3.8)
Edema	-	-	-
Pain[a]	92 (4.0)	80 (6.5)	13 (4.1)
Alopecia	90 (3.9)	72 (5.8)	42 (13.4)
Altered bowel function[b]	121 (5.2)	42 (3.4)	5 (1.6)
Pain in limb[c]	137 (5.9)	87 (7.1)	13 (4.1)
Facial swelling	-	-	-
Painful mouth	142 (6.1%)	27 (2.2)	-
Weight loss	20 (0.9)	38 (3.1)	6 (1.9)
Hand-foot syndrome	27 (1.2)	58 (4.7)	10 (3.1)
Hypertensive disorder	86 (3.8)	-	26 (2.1)
Taste sense altered	77 (3.3)	-	-

Table C.3: Prevalence scores for symptoms in the forum study for different tyrosine kinase inhibitors (TKI) than imatinib. Forum data was adapted from `https://dashboard-gist-adr.herokuapp.com/` accessed on July 14, 2021. Prevalence is based on the percentages of each symptom out of the total number of symptoms for each TKI were calculated. [a]includes: chronic pain and generalized aches and pains [b]includes: constipation and diarrhea [c]includes: any pain in upper or lower limb, excludes: cramp, muscle pain, hand-foot syndrome

Rank	Survey	Rank	Forum
1.	Fatigue	1.	Fatigue
2.	Pain or soreness in mouth*	2.	Painful mouth
	Indigestion or heart burn*	3.	Pain in limb
	Skin problems *	4.	Altered bowel function
	Nausea*	5.	Nausea
	Food and drink tasting different than usual*	6.	Pain
	Feeling weak*	7.	Alopecia
8.	Muscle aches, pains or cramps #	8.	Hypertensive disorder
	Aches and pains in joints#	9.	Taste sense altered
	Pain#	10.	Disorder of skin
	Appetite loss #		
	Diarrhea#		

Table C.4: Ranking of prevalence of symptoms related to sunitinib in survey study and forum study. * same prevalence (83%) # same prevalence (67%)

Rank	Survey	Rank	Forum
1.	Fatigue	1.	Fatigue
2.	Muscle aches, pains or cramps*	2.	Pain in limb
	Shortness of breath*	3.	Pain
	Skin problems *	4.	Alopecia
	Feeling weak*	5.	Hand-foot syndrome
	Food and drink tasting different from usual*	6.	Altered bowel function
7.	Hand-foot syndrome#	7.	Weight loss
	Pain #	8.	Disorder of skin
9.	†	9.	Nausea
		10.	Cramp

Table C.5: Ranking of prevalence of symptoms related to regorafenib in survey study and forum study. * same prevalence (67%) # same prevalence (50%) † six symptoms with same prevalence (33%)

CURRICULUM VITÆ

Anne DIRKSON

2022–present	Scientist Netherlands Forensics Institute
2018–2022	PhD. Computer Science Leiden Institute of Advanced Computer Science Universiteit Leiden
2016–2017	Junior Teaching Fellow University College Maastricht
2014–2016	MSc Neuroscience Vrije Universiteit Amsterdam
2011–2014	BA Liberal Arts & Sciences University College Maastricht
2005–2011	Secondary education (Bilingual VWO) Stedelijk College Eindhoven

LIST OF PUBLICATIONS

14. **Anne Dirkson**, Suzan Verberne, Gerard van Oortmerssen, Hans Gelderblom and Wessel Kraaij (2022). *How do others cope? Extracting coping mechanisms for adverse drug events from social media.* Journal of Biomedical Informatics.

13. **Anne Dirkson**, Dide den Hollander, Suzan Verberne, Ingrid Desar, Olga Husson, Winette T.A. van der Graaf, Astrid Oosten, An Reyners, Neeltje Steeghs, Wouter van Loon, Hans Gelderblom and Wessel Kraaij (2022). *Sample bias in online patient generated health data of Gastrointestinal Stromal Tumor patients: Survey study.* JMIR Formative Research.

12. **Anne Dirkson**, Suzan Verberne, Wessel Kraaij, Gerard van Oortmerssen and Hans Gelderblom (2022). *Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers.* Scientific Reports, 12 (10317).

11. Dide den Hollander, **Anne Dirkson**, Suzan Verberne, Wessel Kraaij, Gerard van Oortmerssen, Hans Gelderblom, Astrid Oosten, Anna K.L. Reyners, Neeltje Steeghs, Winette T.A. van der Graaf, Ingrid Desar and Olga Husson (2022). *Symptoms reported by Gastrointestinal Stromal Tumour (GIST) patients on imatinib treatment: combining questionnaire and forum data.* Supportive Care in Cancer.

10. Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, **Anne Dirkson**, Suzan Verberne, Davy Weissenbacher & Graciela Gonzalez-Hernandez (2021). *DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter.* Journal of the American Medical Informatics Association 28 (10). 2184–2192.

9. **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2021). *FuzzyBIO: a proposal for fuzzy representation of discontinuous entities.* Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis. 77–82.

8. Johan de Graaf, Friso de Vries, **Anne Dirkson**, Olaf Hiort, Alberto Pereira, Márta Korbonits & Martine Cools on behalf of Research and Science Work Package of Endo-ERN (2021). *Patients with rare endocrine conditions have corresponding views on unmet needs in clinical research.* Endocrine 71. 561–568

7. **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2020). *Conversation-aware Filtering from Online Patient Forums.* Proceedings of the Fifth Social Media Mining for Health Applications (SMM4H) Workshop @ COLING.

6. Gautam Kishari, **Anne Dirkson** & Tim Majchrzak (2020). *An Exploratory Study of COVID-19 Misinformation on Twitter.* Online Social Networks and Media.

5. **Anne Dirkson**, Suzan Verberne, Abeed Sarker & Wessel Kraaij (2019), *Data-Driven Lexical Normalization for Medical Social Media*, Multimodal Technologies and Interaction 3(3): 60.

4. **Anne Dirkson** & Suzan Verberne (2019), Transfer Learning for Health-related Twitter Data. Proceedings of the Fourth Social Media Mining for Health Applications (SMM4H) Workshop & Task. Association for Computational Linguistics. 89-92.

3. **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2019), *Lexical Normalization of User-Generated Medical Text: Association for Computational Linguistics.* Proceedings of the Fourth Social Media Mining for Health Applications (SMM4H) Workshop & Task. Association for Computational Linguistics. 11-20.

2. **Anne Dirkson** (2019), *Knowledge Discovery and Hypothesis Generation from Online Patient Forums: A Research Proposal.* Student Research Workshop, Association for Computational Linguistics. 64-73.

1. **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2019), *Narrative Detection in Online Patient Communities.* Proceedings of Text2Story — Second Workshop on Narrative Extraction From Texts co-located with 41th European Conference on Information Retrieval (ECIR 2019). 21-28.

SIKS DISSERTATION SERIES

2011
01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
09 Tim de Jong (OU), Contextualised Mobile Media for Learning
10 Bart Bogaert (UvT), Cloud Content Contention
11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16 Maarten Schadd (UM), Selective Search in Games of Different Complexity
17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18 Mark Ponsen (UM), Strategic Decision-Making in complex games
19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles
20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach
21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22 Junte Zhang (UVA), System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

29 Faisal Kamiran (TUE), Discrimination-aware Classification
30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions
31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU), Arguing to Motivate Decisions
34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control
42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44 Boris Reuderink (UT), Robust Brain-Computer Interfaces
45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012
01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda
02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications
05 Marijn Plomp (UU), Maturing Interorganisational Information Systems
06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices
19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data

49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013
01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29 Iwan de Kok (UT), Listening Heads
30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search

35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36 Than Lam Hoang (TUe), Pattern Mining in Data Streams
37 Dirk Börner (OUN), Ambient Learning Displays
38 Eelco den Heijer (VU), Autonomous Evolutionary Art
39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014
01 Nicola Barile (UU), Studies in Learning Monotone Models from Data
02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
06 Damian Tamburri (VU), Supporting Networked Software Development
07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10 Ivan Salvador Razo Zapata (VU), Service Value Networks
11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
14 Yangyang Shi (TUD), Language Models With Meta-information
15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop

27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33 Tesfa Tegegne (RUN), Service Discovery in eHealth
34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015
01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs

18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27 Sándor Héman (CWI), Updating compressed colomn stores
28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31 Yakup Koç (TUD), On the robustness of Power Grids
32 Jerome Gard (UL), Corporate Venture Management in SMEs
33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34 Victor de Graaf (UT), Gesocial Recommender Systems
35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web

19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TUE), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TUE), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TUE), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VU), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs
38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TUE), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31 Gongjin Lan (VU), Learning better – From Baby to Better
32 Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
34 Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
02 Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), BIMODAL EMOTION RECOGNITION FROM AUDIO-VISUAL CUES
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbò Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (LIACS), Robust rules for prediction and description
24 Kaijie Zhu (TUE), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI & VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022
1 Judith van Stegeren (UT), Flavor text generation for role-playing video games
2 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
3 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
4 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
5 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
6 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
7 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
8 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
9 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna Latour (UL), Optimal decision-making under constraints and uncertainty