



Universiteit
Leiden
The Netherlands

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

Dirkson, A.R.

Citation

Dirkson, A. R. (2022, December 6). *Knowledge discovery from patient forums: gaining novel medical insights from patient experiences*. SIKS Dissertation Series. Retrieved from <https://hdl.handle.net/1887/3492655>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3492655>

Note: To cite this publication please use the final published version (if applicable).

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

¹Available 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 ($\kappa = 0.69$), 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 output². 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 OHDSI³ 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/>

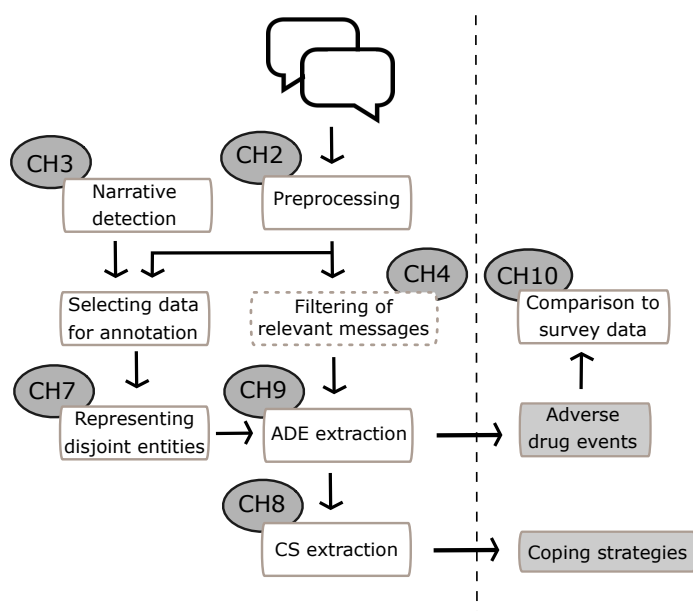


Figure 12.1: An overview of how the thesis chapters interlink and contribute to the extraction and analysis of ADEs and coping strategies. The dotted component (“Filtering of relevant messages”) was not used in the overall pipeline. The output data is indicated in gray boxes to the right of the striped 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 not 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

⁴Normalization 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-CT⁵ (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

⁵https://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 depth⁷ 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 ADEs⁸ 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,

⁶In the CORE subset, we found that concepts range from 1 to 10 levels of depth

⁷This depth was chosen to prevent ADEs becoming too vague or general

⁸SNOMED-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 pre-train 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 Reusability⁹) 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.

⁹Available 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 PatientsLikeMe¹⁰. 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 guidelines¹¹. 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/>

¹¹Annotation 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 course¹². 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

¹²Available at: <https://www.leidraad-ai.nl/>

¹³Available 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.