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

Dirkson, A.R.

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

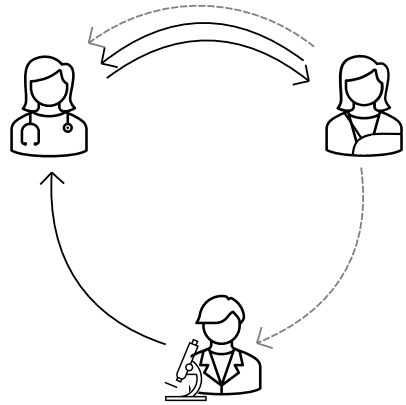

Figure 1.1: The knowledge cycle between medical professional, patient and researcher. The black lines indicate the current knowledge transfer; Researchers inform medical professionals who in turn share their knowledge with patients. Individual patients also share their knowledge with individual medical 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

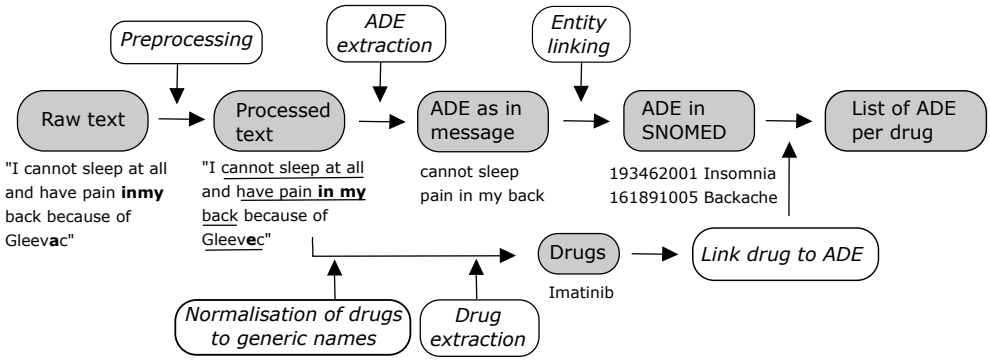

Figure 1.2: An overview of the pipeline for harvesting ADEs from social media. Italicized parts indicate technical steps. An example message is provided to clarify each step

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.