Knowledge discovery from patient forums: gaining novel medical insights from patient experiences
Dirkson, A.R.

PART IV:

GIST AS A CASE STUDY

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

Alan Moore, *V for Vendetta*

9

PATIENT FORUMS AS A COMPLEMENTARY DATA SOURCE

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

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

9.1. Introduction

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

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

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

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

9.2. Materials and methods

9.2.1. Data collection

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

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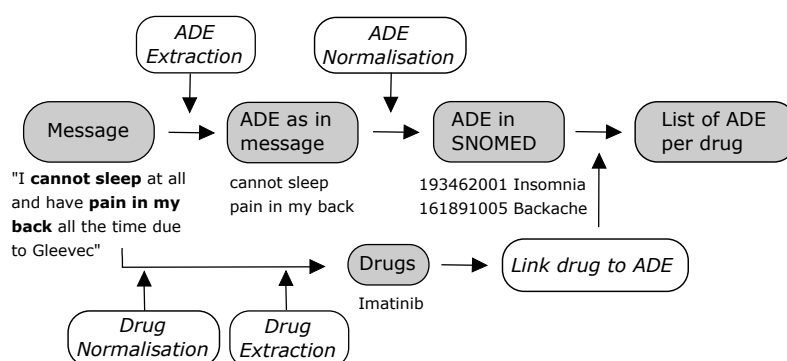

Figure 9.1: An overview of the software pipeline we developed for automatically determining which adverse drug effects (ADE) are mentioned on a patient forum. All italicized parts indicate modules we developed. An example message is provided to clarify each step. ADE: adverse drug events

9.2.2. Machine learning pipeline

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

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

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

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

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

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

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

---

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

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

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

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

9.2.3. Data analysis

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

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

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

---

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

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

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

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

9.3. RESULTS

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

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

---

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

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

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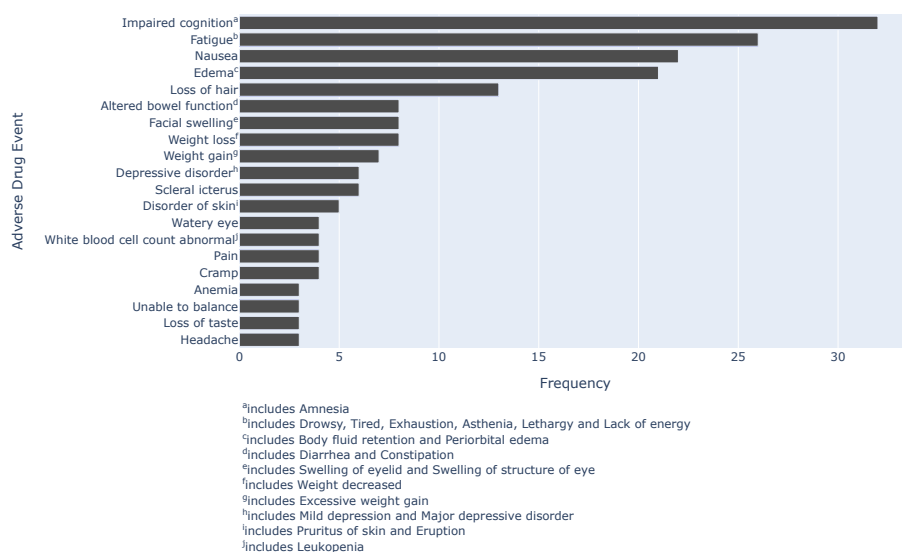

Figure 9.2: The 20 most prevalent adverse drug events reported for avapritinib (formerly BLU-285) on the GIST patient forum
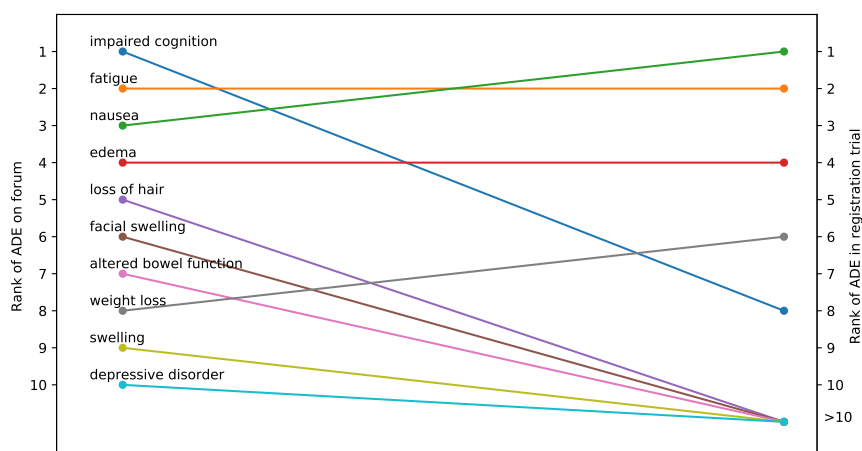
Figure 9.3: The change in rank in terms of prevalence of reporting of the top 10 adverse drug events found for avapritinib on the forum (left) compared to the registration trial (right). ADE: adverse drug events

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

   We also analyze ADEs that occur after long-term use of a drug. Figure 9.4 shows the most prevalent long-term ADEs reported for Imatinib on the GIST patient forum. The most reported are dyspnea, toothache, tremor, vertigo and excessive weight gain. It appears that patients suffer from problems with their teeth (i.e., toothache and tooth disorder), muscles (i.e., tremor, muscle atrophy and muscle fatigue), and skeletal system (i.e., osteoporosis). We acknowledge that these ADEs might be related to other factors such as age, and no definitive causality can be deduced from patient reports. Nonetheless, analysis of long-term ADEs on patient forums can provide valuable hypotheses for future research.
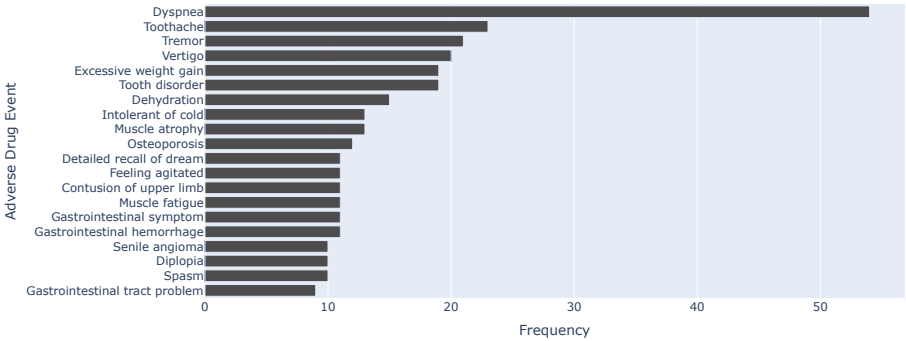
Figure 9.4: The 20 most prevalent long-term adverse drug events reported for imatinib on the forum

ᵃincludes Cramp in foot, Cramp in lower leg and Cramp in limb
ᵇincludes Red eye, Dry eyes, Contusion of eye and Disorder of eye
ᶜincludes Mild depression, Depressed mood and Major depressive disorder
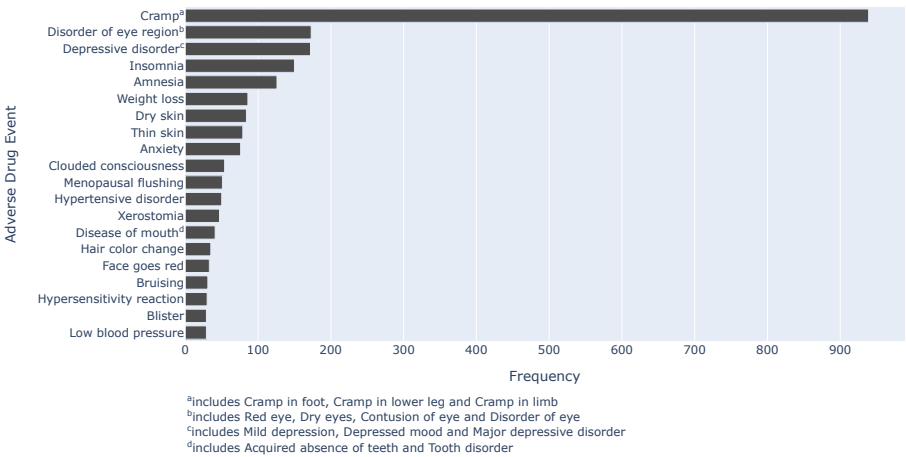ᵈincludes Acquired absence of teeth and Tooth disorder

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

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

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

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

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

9.4. DISCUSSION

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

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

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

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

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

---

[5]https://ohdsi.org/

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

9.5. Conclusion

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

9