A Corpus of Drug Usage Guidelines Annotated with Type of Advice

Sarah Masud Preum, Md. Rizwan Parvez, Kai-Wei Chang, John A. Stankovic
University of Virginia; University of California, Los Angeles
Charlottesville VA 22903; Los Angeles CA 90095
preum@virginia.edu, rizwan@cs.ucla.edu, kwchang@cs.ucla.edu, jas9f@virginia.edu

Abstract

Adherence to drug usage guidelines for prescription and over-the-counter drugs is critical for drug safety and effectiveness of treatment. Drug usage guideline documents contain advice on potential drug-drug interaction, drug-food interaction, and drug administration process. Current research on drug safety and public health indicates patients are often either unaware of such critical advice or overlook them. Categorizing advice statements from these documents according to their topics can enable the patients to find safety critical information. However, automatically categorizing drug usage guidelines based on their topic is an open challenge and there is no annotated dataset on drug usage guidelines. To address the latter issue, this paper presents (i) an annotation scheme for annotating safety critical advice from drug usage guidelines, (ii) an annotation tool for such data, and (iii) an annotated dataset containing drug usage guidelines from 90 drugs. This work is expected to accelerate further release of annotated drug usage guideline datasets and research on automatically filtering safety critical information from these textual documents.

Keywords: Drug Usage Guidelines, Annotation Tool, Medical Corpora, Health Informatics

1. Introduction

Drug safety is crucial for ensuring overall health safety of patients and effectiveness of treatment (Yi et al., 2015). In order to ensure drug safety, with each of their prescribed drugs, patients are often provided with drug usage guidelines (DUG) documents [1]. These documents contain essential information regarding drug usage, including, but not limited to, dosage, drug administration, adverse reactions or symptoms, drug storage and disposal, contraindications, drug-drug interactions, and drug-food interactions (MDS, 2017) as illustrated in Table 1. Doctors and pharmacists are expected to educate patients about drug usage guidelines to ensure drug safety (Patel and Dowse, 2015). Such as, educating a patient who is prescribed *Coumadin* to avoid certain herbal products like *St John's Wort* and foods high in Vitamin K due to potential drug-food interaction. But patients are often unaware of such critical information / advice related to their prescription drugs due to several reasons, including, (i) lack of communication with their doctors and/or pharmacists (Liddy et al., 2014), (Patel and Dowse, 2015) (ii) low health literacy (Wolf et al., 2006), and (iii) volume of received information (Savas and Evcik, 2000).

Textual analysis of drug usage guidelines and automatically filtering critical personalized advice can aid a patient to adhere to the drug usage guidelines. Such research can benefit from annotated datasets where the information is annotated according to its topic, e.g., whether a piece of advice from a DUG document is related to *interaction with food or beverage* or *interaction with other drugs*. For example, the annotated advice from DUG documents can be forwarded to a medication reminder app to present advice that indicates potential interaction with (i) daily activities, (ii) exercise, (iii) diet, and (iv) other drugs. Examples of such advice are presented in Table 1. Presenting advice based on their topics can aid users to receive critical advice effectively and increase drug adherence (Tang et al., 2014), (Jimmy and Jose, 2011). Also, often advice are subjected to physiological and/or temporal conditions. Annotated advice can be used to filter out irrelevant information and personalize DUG documents for a patient. For example, removing pregnancy related advice for a male patient. This will reduce the information burden for the patients, which is often identified as a primary barrier to self-management of chronic diseases (Woolley, 2015).

> ...
> **HOW TO USE:** Read the Medication Guide ...
> Take this medication by mouth with or without food as directed by your doctor or other health care professional, usually once a day. It is very important to take it exactly as directed. ...
> It is important to eat a balanced, consistent diet while taking Warfarin. ... Avoid sudden large increases or decreases in your intake of foods high in vitamin K (such as Broccoli, Cauliflower, Cabbage, Brussels sprouts, Kale, Spinach, and other green leafy vegetables, liver, green tea, certain vitamin supplements). If you are trying to lose weight, check with your doctor before you go on a diet. ...
> Since this drug can be absorbed through the skin and lungs and may harm an unborn baby, women who are pregnant or who may become pregnant should not handle this medication or breathe the dust from the tablets.. ...

Text Box 1: An excerpt from the drug usage guideline (DUG) document of *Warfarin* from MedScape (MDS, 2017). Different colors are used to underline advice related to different topics. Text underlined in blue, green, and red indicate advice related to *drug administration*, *food interaction*, and *pregnancy*, respectively. Such DUG documents are also available from FDA drug database(FDA, 2017) and WebMD(WMD, 2017).

Lack of annotated corpus of the DUG documents limits the potential NLP research on extracting critical information from DUG data to increase drug safety. To bridge this knowledge gap, we develop a novel annotation scheme and a novel, interactive annotation tool to annotate textual ad-

[1] Also known as consumer medical information (CMI) and patient handouts.

vice statements from DUG documents according to their topics. This annotation tool is used to annotate a corpus of 90 online DUG documents and 9,831 sentences. The multi-label annotation results in the first annotated corpus of DUG documents containing 1,611 annotated safety critical drug usage guidelines. We make the annotation tool and the annotated corpus available to the community (Preum et al., 2018). These resources can aid the release of more annotated datasets of DUG documents and accelerate NLP research on automatic extraction of safety critical information from these textual documents.

2. Background

In this section we briefly introduce the drug usage guideline (DUG) data and the relevant existing research.

The DUG documents contain a variety of information spanning different topics, which can be broadly categorized in following classes. (i) basic drug information (e.g., alternate names, diseases that are commonly treated with the drug, ingredients, dosage information), (ii) drug administration related advice (i.e., how and when the drug should be administered), (iii) side effects of the drug, (iv) information regarding how the drug can interact with other drugs, foods and activities, and (v) drug storage and disposal related information. This information is the input for the data annotation tool.

The DUG documents are semi-structured documents where the text contents are organized under different section headers. As shown in the text excerpt in Section 1., under the tile *How To Use*, information related to drug administration and potential drug-food interaction are included. The topics described above are not always structurally organized in the textual documents. Often critical information related to drug-food interaction or drug-activity interaction are scattered through the document and patients face difficulties in finding them. Also, the structure and organization of information varies from one source to another.

Such irregular organization and structural variety across sources pose a challenge to automatically extract critical information from the DUG documents. Therefore, there is a need for an annotated corpus for DUG documents. Currently, there are several annotated medical corpora (Saeed et al., 2011), (Aronson and Lang, 2010), (Uzuner et al., 2007), (Pardelli et al., 2012), (Bongelli et al., 2012) that contain clinical notes, bio-medical textual contents, and electronic health records. There are some existing annotation tools and techniques than focus on annotating electronic health records (Roberts and Demner-Fushman, 2016) and clinical practice guidelines (Read et al., 2016) and extracting relations from bio-medical text (Ellendorff et al., 2014). But to the best of our knowledge, there is no existing work that focuses on annotating and analyzing the textual content of such documents. So, we introduce a DUG data annotation scheme, an annotation tool, and an annotated corpus as presented in the following sections.

3. Data Annotation Scheme

While a DUG document contains an array of information, not all information is critical to patients (Yi et al., 2015). We develop the following multi-label data annotation scheme to annotate the critical advice statements of DUG documents in 8 categories.

1. **Activity or lifestyle related advice**: to indicate potential interaction between the corresponding drug and any activity of daily living (e.g., driving). For instance, from Table 1, driving or performing other activities that require alertness might cause fatal accident if the person took Ambien within a day.

2. **Disease or symptoms related advice**: to indicate potential interaction between the corresponding drug and any diseases / symptoms. Such advice statements are crucial for patients suffering from multiple diseases, as multiple diseases can often be conflicting (Kienhues et al., 2011), (Caughey et al., 2013).

3. **Drug administration related advice**: to annotate important advice related to drug administration process, e.g., how the drug should be taken. This type of advice are essential for medication adherence and effective treatment. For example, Topomax should not be taken within 6 hours of drinking alcohol (Table 1).

4. **Exercise related advice**: to indicate potential interaction between the corresponding drug and any exercise (e.g., running outdoors in a hot weather). As exercise is often suggested to patients taking prescription drugs, annotating the special circumstances when (i) exercise should be limited or avoided in certain context or (iii) performed with certain preparation (e.g., checking blood sugar before exercising) is necessary to avoid adverse conditions.

5. **Food or beverage related advice**: to indicate potential interaction between the corresponding drug and any food / beverage. This is critical to health safety. For example, in the text excerpt from DUG document of drug Warfarin presented in Section 1., patients are suggested to avoid dark green vegetables as they can interact with the drug. But these vegetables are widely known as healthy food and often suggested to people for weight loss. High weight is a significant factor of deep vein thrombosis, a disease for which Warfarin is suggested. So, this particular food related advice is critical for the patients who are prescribed Warfarin.

6. **Other drug related advice**: to indicate the advice suggest avoiding / limiting consumption of other drugs. As shown in Table 1, when taking Abilify, certain prescription and over-the-counter drugs (e.g., allergy or cough relief products) should be taken with caution, as they can increase the effect of drowsiness.

7. **Pregnancy related advice**: to indicate whether the advice is for women who are already pregnant, are planning to conceive, or recently gave birth. This advice is critical to pregnant women, nursing mother, and children. Also, as pregnancy related advice statements are applicable for only a small portion of the patient pool, this annotated advice can be filtered according to the personal condition of a patient.

Drug Name	Advice Text	Annotation
Ambien	Do not **drive, use machinery, or do any activities** that require clear thinking after you take this medication and the next day. You may feel alert, but this medication may continue to affect your thinking, making such activities unsafe.	Activity or lifestyle related
Zoloft	Older adults may also be more likely to develop a type of **salt imbalance** (**hyponatremia**), especially if they are taking "water pills" (**diuretics**).	Disease or symptom related Other drug related
Fentanyl	Avoid activities that might cause your body temperature to rise. Such as doing **strenuous work/exercise** in hot weather.	Exercise related
Topamax	not drink **alcoholic beverages for 6 hours before or 6 hours after taking Topamax** extended release capsules, since alcohol may affect this medication works.	Temporal Food or beverage related Drug administration related
Abilify	Ask your pharmacist about using **allergy** or **cough-and-cold products** because they may contain ingredients that cause drowsiness.	Other drug related
Ativan	This medication is not recommended for use during **pregnancy**. It may harm an unborn baby.	Pregnancy related

Table 1: Different types of advice extracted from the online DUG data. The first, second, and third columns contain the name of the drug, an advice statement from the DUG document of that drug, and the annotation of that advice statement, respectively. An advice statement can have multiple tags based on its topics.

8. **Temporal advice**: to indicate the advice suggests an action with temporal condition(s), e.g., when to take a drug, for how long to wait before eating/drink after taking the drug. It suggests duration or frequency of drug usage, and interval between consecutive dosage. Also, it denotes temporal dependency between taking a drug and one of the following events: (i) having a meal, (ii) doing an activity, and (iii) exercising. Such as, as shown in Table 1, when taking Topomax, alcohol should not be consumed with in 6 hours.

Although a DUG document contains a few other categories of advice (e.g., drug dosage, route of drug administration, drug storage, drug disposal), those categories are not considered here. Because, existing studies find patients are educated by their primary health care providers on these types of advice (Yi et al., 2015) when applicable. So, this information can be either filtered out or presented with lower priority. This will reduce the document length and the cognitive overload of the patients (Savas and Evcik, 2000), (Shrank and Avorn, 2007).

4. Data Annotation Tool

Our goal is to develop a drug usage guideline (DUG) document annotation tool that is **interactive** and **generalizable** to DUG data from different sources. Based on empirical evidence found in the DUG documents from different sources, the tool should address following issues:
(i) The tool should parse through the DUG document sentence by sentence, as critical information can be found in different parts of the document. (ii) It should provide the option to annotate advice that spans across multiple sentences. (iii) It should be flexible so that an annotator can add multiple tags to an advice. (iv) It should be interactive so that an annotator can change previous annotations based on his new observation as he/she goes through the document. (v) As the DUG documents often contain redundant headers or text fragments, the tool should allow the annotator to select part of text as advice.

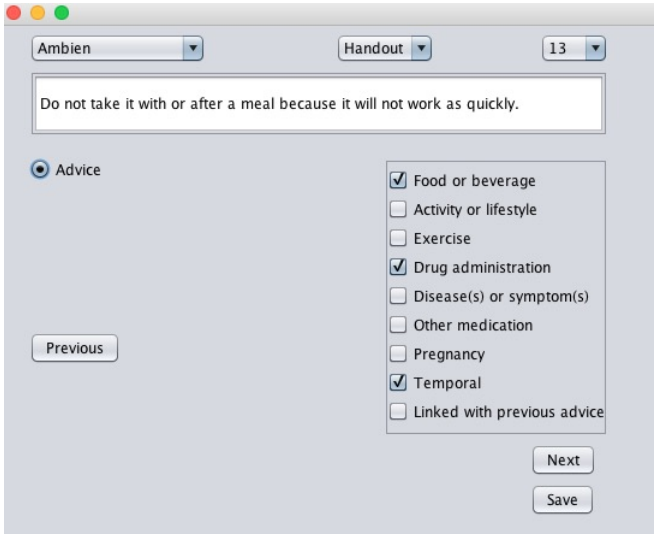

Figure 1: The Data Annotation Tool for annotating advice statements in DUG according to their topics. Here, an annotator tagged the 13-th sentence of the DUG document of the drug named Ambien as advice. He also annotated three topics for the current sentence, namely, food or beverage, drug administration, and temporal.

We have developed a desktop application for annotating DUG documents based on the annotation scheme described in Section 3. that addresses the above mentioned issues. This tool is developed using Java Swing library. A screen shot of this tool is shown in Figure 1. The input of this tool is the textual content of the DUG documents. The output is the annotation of each sentence of the DUG document. The main features of this tool are described below.

An annotator can select a drug to start annotating its DUG document. The tool supports annotating multiple files for the same drug, as often there are multiple overlapping sources of drug usage guidelines for a single drug. It can

Type of Advice	Count	%Gold Label
Activity or lifestyle related	146	100
Disease or symptom related	245	97.5
Drug administration related	224	98.2
Exercise related	40	97.9
Food or beverage related	253	99.6
Other drug related	310	100
Pregnancy related	211	99
Temporal	182	99.45

Table 2: Annotation of Drug usage guideline dataset. Eight types of advice are annotated in the data as shown in column 1. The second column denotes the count of advice for each type of advice. The third column contains the % of advice statements of that received gold label in annotation.

be selected from the drop down menu named *Handout* in Figure 1. The annotator can go to a specific line of the current DUG document. An annotator can tag a sentence as an advice and specify the categories of the advice statements. Also, a sentence can be annotated as an advice without specifying the exact category of the advice. When an advice statement consists of multiple consecutive sentences, it can be annotated using the *linked with previous* option. An annotator can browse forward and backward through a document and update the annotation of each sentence.

Also, the DUG documents often contain section headers, titles, and formatted text in between sentences that adds redundant text fragments in advice sentence(s). This annotation tool allows the annotators to select portion of text as advice and ignore the rest of the text. As it is unlikely for an annotator to annotate the whole corpus at a time, the tool supports session memory, i.e., once the application is relaunched it loads the data from the most recent position of the corpus.

This tool can be easily adapted to annotate more categories of advice. It can also be generalized to annotate DUG documents from different sources and other descriptive health / clinical textual documents (e.g., health articles, websites).

5. Annotated Corpus

The data collection is motivated by self management of patients suffering from multiple chronic diseases. We sampled 34 anonymized prescriptions from MTSamples. This dataset contains anonymized prescriptions of real patients. Each of the sampled prescriptions represents an anonymous patient who is suffering from two or more chronic diseases. The chronic diseases covered in the sampled prescriptions include the most common chronic diseases, e.g., diabetes, hypothyroidism, bipolar affective disorder, alcohol withdrawal, anxiety, depression, lethargy, alcohol dependence, substance abuse, obesity, chronic pain, chronic kidney disease, and coronary vascular disease. Each prescription contains a list of suggested drugs and their corresponding dosages.

From the 34 prescriptions, a total of 166 drugs are found. For each of these drugs, we crawled online drug usage guidelines (DUG) documents from MedScape. We chose MedScape as it is one of the most widely used applications by the physicians. It contains more comprehensive DUG documents when compared to FDA drug database or WebMD DUG documents. Among the 166 drugs, the online drug usage guideline document is available for only 90 drugs in MedScape. We crawled and annotated these 90 online DUG documents.

The collected DUG document corpus contains 9,831 sentences and 170,646 words. It is annotated by three human annotators using the data annotation tool presented in Section 4. For each advice, they also tag the potential categories of the advice based on the annotation scheme presented in Section 3. Each of the annotators have at least a masters degree in computer science. Majority voting is applied to decide the ground truth of annotation. The result of this annotation is presented in Table 2. There are 1,637 advice statements in eight categories. Here, majority of the advice received gold label, i.e., all three annotators agree on the annotation of the advice.

Among the eight categories of our proposed scheme, *other drug related* advice is the most common category (n=336). Advice from *food and beverage related* categories are also common (n=253), as most of the drugs in our constructed corpus interact negatively with alcoholic beverages. As our corpus contains drugs that are used for treating chronic diseases, a major portion of the annotated advice (n=245) describes how a drug can interact with *other diseases* or *cause physiological / psychological syndromes*. *Exercise related advice* are relatively rare (n=40), as exercise is often recommended for most of the chronic diseases. However, it is found from the annotation that exercise in certain contexts (e.g., in a hot or humid weather, within certain time range of drug administration) can negatively affect well-being. Some of the categories show strong correlation, e.g., *drug administration related* advice statements are often *temporal*, and related to *food/beverage, activity* (e.g., sleeping, driving), and *exercise*.

This dataset or the extension of such dataset can be used to automatically extract personalized advice from patient handout or DUG documents to raise patient's awareness and hereby increase medication adherence and effectiveness of treatment. Some potential applications to utilize such datasets are presented below.

Firstly, a medication reminder app can present personalized safety critical drug related advice while reminding the patient to take a medication. Such as, reminding safety-critical pregnancy related advice from a DUG document to only women of child bearing age. As often DUG documents contain numerous advice, filtering them in a personalized manner and prioritizing them according to severity can increase patient's adherence.

Secondly, another application is detecting conflicts between advice from the DUG documents with other health related advice (Preum et al., 2017a; Preum et al., 2017b). Such as, Aspirin is suggested to take with food or milk to avoid stomach upset. On the other hand, an individual with lactose intolerance is also suggested to avoid milk. In this case, there is a conflict between these two advice statements. Such datasets can accelerate automatic conflict detection from drug related advice.

Finally, presenting drug administration related advice in a

context-aware manner can increase effectiveness of treatment. Often drugs are suggested to take within a certain interval of other activities (e.g., one hour after meal, 2 hours before sleep). Individuals suffering from multiple conditions may not be aware of all such temporal, drug administration related advice, as they are often prescribed several drugs. In such cases, presenting relevant drug related advice by inferring the context of their daily lives can be beneficial. Such as, for a drug that should be taken two hours before sleep, an activity tracking app can suggest it's user to take the drug two hours before her frequent bed time.

6. Conclusion

A plethora of textual documents containing crucial information on drug usage guidelines (DUG) are available online. Although analyzing such textual document can aid patient education and promote safe usage of drugs, these resources are underutilized. To bridge this knowledge gap, we introduce a multi-label annotation scheme to annotate advice from the DUG documents in eight categories based on their topics. We develop an interactive data annotation tool for this data that can also be generalized to annotated advice from various other descriptive textual information sources (e.g., DUG documents from other sources, health article). Finally, we share the first annotated corpus on DUG data containing annotated drug usage guidelines of 90 drugs that are used to treat over 30 chronic diseases. The corpus contains 9,831 sentences and 1,611 advice statements on eight safety critical categories. The corpus yields several important insights on instructions regarding safe usage of drugs. The annotation tool and the annotated corpus can aid future research to automatically annotate / classify critical information from DUG documents as well as other textual health documents.

7. Acknowledgements

This work was supported, in part, by NSF grant CPS-1646470.

8. Bibliographical References

Aronson, A. R. and Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Bongelli, R., Canestrari, C., Riccioni, I., Zuczkowski, A., Buldorini, C., Pietrobon, R., Lavelli, A., and Magnini, B. (2012). A corpus of scientific biomedical texts spanning over 168 years annotated for uncertainty. In *LREC*, pages 2009–2014.

Caughey, G., Gilbert, A., Roughead, L., McDermott, R., Ryan, P., Esterman, A., et al. (2013). Multiple chronic health conditions in older people. *Implications for health policy planning, practitioners and patients*.

Ellendorff, T., Rinaldi, F., and Clematide, S. (2014). Using large biomedical databases as gold annotations for automatic relation extraction. In *LREC*, pages 3736–3741.

(2017). Medication guides: Us food and drug administration. `https://www.fda.gov/Drugs/DrugSafety/ucm085729.htm`. Accessed: 2017-04-15.

Jimmy, B. and Jose, J. (2011). Patient medication adherence: measures in daily practice. *Oman medical journal*, 26(3):155.

Kienhues, D., Stadtler, M., and Bromme, R. (2011). Dealing with conflicting or consistent medical information on the web: When expert information breeds laypersons' doubts about experts. *Learning and Instruction*, 21(2):193–204.

Liddy, C., Blazkho, V., and Mill, K. (2014). Challenges of self-management when living with multiple chronic conditions. *Canadian Family Physician*, 60(12):1123–1133.

(2017). Medscape: Search drugs, otcs & herbals. `http://reference.medscape.com/drugs`. Accessed: 2017-04-15.

Pardelli, G., Sassi, M., Goggi, S., and Biagioni, S. (2012). From medical language processing to bionlp domain. In *LREC*, pages 2049–2055.

Patel, S. and Dowse, R. (2015). Understanding the medicines information-seeking behaviour and information needs of south african long-term patients with limited literacy skills. *Health Expectations*, 18(5):1494–1507.

Preum, S. M., Mondol, A. S., Ma, M., Wang, H., and Stankovic, J. A. (2017a). Preclude: Conflict detection in textual health advice. In *Pervasive Computing and Communications (PerCom), 2017 IEEE International Conference on*, pages 286–296. IEEE.

Preum, S. M., Mondol, A. S., Ma, M., Wang, H., and Stankovic, J. A. (2017b). Preclude2: Personalized conflict detection in heterogeneous health applications. *Pervasive and Mobile Computing*.

Preum, S. M., Parvez, M. R., Chang, K.-W., and Stankovic, J. A. (2018). A Corpus of Online Drug Usage Guideline Documents Annotated with Type of Advice. Available at `https://doi.org/10.5281/zenodo.1173345`.

Read, J., Velldal, E., Cavazza, M., and Georg, G. (2016). A corpus of clinical practice guidelines annotated with the importance of recommendations.

Roberts, K. and Demner-Fushman, D. (2016). Annotating logical forms for ehr questions. In *International Conference on Language Resources and Evaluation*, volume 2016, page 3772. NIH Public Access.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011). Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.

Savas, S. and Evcik, D. (2000). Do undereducated patients read and understand written education materials? a pilot study in isparta, turkey. *Scandinavian journal of rheumatology*, 30(2):99–102.

Shrank, W. H. and Avorn, J. (2007). Educating patients about their medications: the potential and limitations of written drug information. *Health Affairs*, 26(3):731–740.

Tang, F., Zhu, G., Jiao, Z., Ma, C., Chen, N., and Wang,

B. (2014). The effects of medication education and behavioral intervention on chinese patients with epilepsy. *Epilepsy & Behavior*, 37:157–164.

Uzuner, Ö., Luo, Y., and Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

(2017). Webmd: Drugs medications a-z. `http://www.webmd.com/drugs/2/index`. Accessed: 2017-04-15.

Wolf, M. S., Davis, T. C., Tilson, H. H., Bass III, P. F., and Parker, R. M. (2006). Misunderstanding of prescription drug warning labels among patients with low literacy. *American Journal of Health-System Pharmacy*, 63(11).

Woolley, K. R. (2015). Enhancing education of medication side effects to improve patient outcomes.

Yi, Z.-M., Zhi, X.-J., Yang, L., Sun, S.-S., Zhang, Z., Sun, Z.-M., and Zhai, S.-D. (2015). Identify practice gaps in medication education through surveys to patients and physicians. *Patient preference and adherence*, 9:1423.