Preclude2: Personalized Conflict Detection in Heterogeneous Health Applications

Sarah Masud Preum*, Abu Sayeed Mondol†, Meiyi Ma‡, Hongning Wang§, and John A. Stankovic¶
Department of Computer Science, University of Virginia
Charlottesville, Virginia 22904
Email: *preum@virginia.edu, †mm5gg@virginia.edu, ‡mm5tk@virginia.edu, §hw5x@virginia.edu, ¶jas9f@virginia.edu

***Abstract*—Conflicting health information is one of the primary barriers of self-management of chronic diseases. This problem is growing with the prevalence of pervasive digital health care applications. Increasing number of people now rely on mobile health apps and online health websites to meet their information needs and often receive conflicting health advice from these sources. This problem is more prevalent and severe in the setting of multi-morbidities. In addition, often medical information can be conflicting with regular activity patterns of an individual. In this work, we formulate the problem of finding conflicts in heterogeneous health applications including health websites, health apps, online drug usage guidelines, and daily activity logging applications. We develop a comprehensive taxonomy of conflicts based on the semantics of textual health advice and activities of daily living. Finding conflicts in health applications poses its own unique lexical and semantic challenges. These include large structural variation between text and hypothesis pairs of advice, finding conceptual overlap between pairs of advice, inference of the semantics of an advice (i.e., what to do, why and how) and activities, and aligning activities suggested in advice with the activities of daily living based on their underlying dependencies and polarity. Hence, we develop *Preclude2*, a novel semantic rule-based solution to detect conflicts in activities and health advice derived from heterogeneous sources. Preclude2 utilizes linguistic rules and external knowledge bases to infer advice. In addition, Preclude2 considers personalization and context-awareness while detecting conflicts. We evaluate *Preclude2* using 1156 real advice statements covering 8 important health topics, 90 online drug usage guidelines, 1124 online disease specific health advice covering 34 chronic diseases, and 2 activity datasets. The evaluation is personalized based on 34 real prescriptions. *Preclude2* detects direct, conditional, sub-typical, quantitative, and temporal conflicts from 2129 advice statements with 0.91, 0.83, 0.98, 0.85 and 0.98 recall, respectively. Overall, it results in 0.88 recall for detecting inter advice conflicts and 0.89 recall activity-advice conflicts. We also demonstrate the effects of personalization and context awareness in conflict detection from heterogeneous health applications.**

I. Introduction

A. Motivation

Chronic diseases are identified as the primary reason of death and disability in the United States. As of 2012, about 117 million people, which amounts for almost half of the adult population, are suffering from at least one chronic disease [1]. One of four adults suffer from multi-morbidities, i.e., they have two or more chronic diseases [1]. People suffering form chronic diseases account for 81% of hospital admissions and 91% of prescriptions [2]. Thus, chronic diseases cause increased demand in health care and the rise of health care cost. So, it is essential to effectively manage chronic disease [2] to improve the general population health. But existing research has identified several challenges that impede self management of chronic diseases, including, undesirable physical and emotional symptoms (e.g., pain, depression), poor access to care, lack of communication with health care providers, and conflicting information [3], [4], [5]. In this research we address the challenge of detecting conflicting information in the context of multiple chronic diseases among adults.

Conflicting health advice[1] refers to the phenomenon where two or more pieces of advice are logically inconsistent. Conflicting health advice frequently occurs in case of chronic diseases as well as in case of other diseases and general health topics (e.g., weight loss, diet, pregnancy). Advice originating from multiple health apps/websites can be conflicting [6], [7] due to three factors. **Firstly**, when two information sources (i.e., app/website) provide advice on different health topics/diseases they might be conflicting. Such as, often advice sources on dyslipidemia and obesity suggest patients to eat grapefruit for its high nutritional content while advice on hypertension suggest to avoid grapefruit as it might interact some hypertension medications. **Secondly**, even when two sources are related to the same topic/disease, conflicts may occur due to conflicting findings from the underlying research corresponding to each source [8] or different interpretations of advice text by advice recipients. **Finally**, an app/website may lack the contextual awareness and/or personalization of a user and suggest an advice that adversely interacts with the physiology, lifestyle, diet, disease, or medications of the user and thus causes a conflict. For instance, a weight loss app suggests a user to reduce her calorie intake while being unaware of the pregnancy of the user and thus her physiological requirement of consuming increased amount of calories. In this case, the weight loss app providing conflicting advice as it is not personalized.

This problem is even more common in case of multi-morbidity due to (i) a fragmented care system and (ii) interaction among clinical or medical guidelines and lifestyle.

- To begin with, the health care system usually provides care on a *per disease* basis. Thus patients with multi-morbidities receive treatments from a number of specialists, which can result in conflicting treatments. The most common case occurs due to using multiple drugs / medications: the use of one may contraindicate or negatively interact with the use of another. As a result, treatment of one disease can be conflicting with the

[1]We refer to actionable health information as health advice.

	Cases	Advice 1	Advice 2
1	Opposite polarity (actions)	Eat citrus fruits and green leafy vegetables as they are rich in Vitamin C.	Be careful about green leafy vegetables if you are on Coumadin or ACE Inhibitors.
2	Opposite polarity (effects)	Pate made from meats may carry the listeria bacteria and cause listeriosis. Avoid eating it while pregnant.	Consume red meat at least two to three times a week to fight anemia.
3	Temporal	Do stretching exercises when you wake up.	Avoid stretching or similar exercises after the end of week 12 of your pregnancy.
4	Conditional	Alcohol may severely affect your baby's development. Avoid alcohol if pregnant or trying to conceive.	Small amounts of alcohol increase the body's metabolic rate, causing more calories to be burned.
5	Sub-typical	Eat calcium-rich foods like milk, cheese and green vegetables.	Use skimmed milk instead of whole milk as dairy products often cause bloating and gas.
6	Quantitative	Limit your caffeine intake to less than 200 milligrams per day during pregnancy.	Up to 400 milligrams (mg) of caffeine a day appears to be safe for most healthy adults.
7	Cumulative effect	Run for at least 30 minutes a day.	Take Salmeterol 1 inhalation (50 mcg) twice daily.

TABLE I: Possible cases of conflict: all advice are taken from real health apps or authentic medical sites

treatment of another disease. Such as, the treatment of diabetes is often conflicting with the treatment of each of the following diseases: Arthritis, Chronic obstructive pulmonary disease (COPD), inflammatory disease, and heart failure [9].

- In addition, treatment guidelines often require making changes in lifestyle [10], [11] and thus interact with activities of daily living. For example, one disease treatment (e.g., obesity / diabetes / heart disease) may require a person to loose weight and perform high intensity exercise on a regular basis. But a prescription drug (e.g., Lopressor, Lithium) may require avoiding extensive exercise temporarily (while the drug is being administered) to maintain stable body temperature and heart rate. Another example is, people often go to sleep / have meal at a certain time. This is specially important for people with multi-morbidities, as they are often required to follow strict daily routine. But some prescription drugs (e.g., Nexium, Levothyroxine, Celebrex) require not eating / sleeping for a certain period of time after the drug is administered.

The adverse effects of conflicting health advice can vary based on whether patients are aware or unaware of the conflicts. When patients are aware of conflicts it causes confusion, frustration and even low adherence to treatment as many report that they resort to inaction when faced with conflicting advice. On the other hand, patients may face even more adverse outcomes in case they are not aware of the conflicts. Because, following conflicting advice may result in (i) short term (e.g., sudden spike in heart rate) and/or long term (e.g., organ failure) health damage and (ii) render treatments ineffective. Eventually, this can increase both health risks and cost of health care. So it is imperative to **automatically detect conflicts** from health advice and activities of daily living (ADL) before they occur. Also, conflicts often occur based on personalized physiological conditions and contexts of an individual. For instance, conflicts of case 2 and case 6 from Table I occur only during pregnancy and thus are not relevant to someone who is not pregnant. The conflict detection system should be **personalized** and **context-aware** to reduce false alarms.

Although people often report finding conflicting health information from online resources or educational materials, it is not realistic for them to detect all conflicts as follows.

- They often lack access to important information [12] due to low health literacy, lack of awareness, or lack of doctor-patient communication.
- They often don't read all sources [13] that are available to them. The most common example of this is how often people skip reading drug usage guidelines (DUG) [14] [2] of prescription drugs. This result in ineffective treatment and low adherence as DUG documents often contain important information on interactions between multiple drugs, drugs and foods, and drugs and activities of daily living.
- Even if they read the available information, they often don't understand the information and misinterpret [15] it.
- People may forget the health information [16] given the high volume and speed of information flow, specially in the context of multiple chronic diseases. This problem is even more serious for patients whose mental agility is affected due to disease, continuous treatment, or medication. Although doctors or pharmacists can aid patients in detecting and resolving potential conflicts in their treatment in an ideal setting; in practice, that's hardly the case due to lack of time [17] and communication gap [15] between professional care providers and patients.

These reasons further underline the need to automatically detect conflicts from health information sources / applications and notify users about it.

In this paper, we develop an automated system to *detect conflicts from heterogeneous health and medical applications by natural language inference and rule based methods in a **personalized** and **context-aware** manner*. Specifically, we consider textual health advice statements originated from health web sites and apps, online drug usage guidelines for prescription medications, and activity logs to detect potential conflicts in case of general health setting and in case of a clinical setting, namely, multi-morbidity. Such a system can alert patients about

[2]Drug usage guideline document is also known as patient handout / consumer medical information / package insert.

potential severe conflicts before they occur and thus increase drug safety and overall health safety.

B. Challenges

Detecting conflicts across multiple health applications poses both lexical and semantic challenges. This task is **lexically challenging** as the lexical structure of advice text can vary significantly in terms of length of advice text and/or tone of advice. Another lexical challenge is processing semi-structured and unstructured textual data (i.e., advice) and aligning them with structured data (i.e., activity log) to detect potential dependencies and conflicts. The **semantic challenges** of conflict detection are multi-fold. **Firstly**, we need to extract the implied action and resulting effects of an advice from the text. **Secondly**, we need to detect whether two or more advice statements have any conceptual overlap (e.g., Kale and cruciferous vegetables). Detecting conceptual overlap often requires inferring the hierarchical relationships between different topics, such as, foods, drugs, and exercise. **Thirdly**, often conflicts are temporal or conditional, i.e., a conflict occurs if a temporal/physiological condition holds true. Hence, it requires thorough inference of the semantics of an advice. **Finally**, to detect potential conflicts between textual advice statements and activity of daily living, we need to identify and understand the semantics of the corresponding activity and advice. This requires formulating the temporal specifications of advice, identifying the potential dependencies among activities, and capturing contextual and personalized information. Context of an activity or suggested advice can be spatial (e.g., outdoor, indoor), temporal (e.g., time of the day) or environmental (e.g., hot weather). We also need to consider personalization of health advice according to the age, gender, physiological conditions, and medical history of an individual while detecting conflicts.

There are some existing works that focus on detecting contradiction in a given pair of sentences/texts [18], [19]. They model this problem as a binary classification task and apply statistical learning models. But detecting a contradiction in a given pair of sentences/texts is different from detecting conflicting advice. Because, the former does not require (i) detecting conceptual overlap to find potential candidates of conflict and (ii) understanding the semantics of an advice, e.g., action, effect, condition. Also, statistical learning models require a lot of labeled training data which is currently unavailable for textual health advice.

C. Contributions

We present Preclude2, a semantic rule based system to detect conflicts in advice derived from health websites, online DUG documents and activities of daily living. The most relevant work in this regard is Preclude [20], our previous conference version of this work. It only considers health advice collected from health apps and websites. As a result, it overlooks potential severe conflict regarding drug usage guideline documents and activities of daily living. Also, it focused on evaluating conflict detection only in the setting of general health and well being. In this work, we develop *Preclude2* by extending Preclude to detect conflicts among textual advice statements and activities of daily living through (i) inferring important health advice statements and (ii) semantics extraction from activities of daily living. The health advice are extracted from online drug usage guideline (DUG) documents and disease specific authentic medical websites. In addition, Preclude2 is evaluated in the setting of multiple chronic diseases among adults. The main contributions of this work beyond Preclude are as follows.

- Preclude2 is the first to formulate and solve the problem of automatic conflict detection from heterogeneous health applications, namely, online/mobile health websites, online drug usage guidelines, and activity logs in a **personalized** and **context-aware** manner. Thus, Preclude2 enhances drug safety and health safety in the imperative scenario of chronic diseases.
- Preclude2 is the first to extract and annotate important advice statements from online DUG documents. We have created a dataset of 1005 advice statements that are extracted from 90 online DUG documents based on real prescriptions of patients suffering from multiple chronic diseases. While we use this data to detect potential conflicts and increase drug safety, it can be further utilized to address several other health care challenges, including but not limited to, improving patient education, medication recommendation, analysis of potential drug interactions, etc.
- Preclude2 is evaluated using multiple real datasets in the setting of multi-morbidity. Specifically, it is evaluated by emulating patients from (i) 34 real prescriptions, (ii) 90 online DUG documents corresponding to 90 prescription medications, (iii) 1124 real disease specific advice statements covering 34 chronic diseases, and (iv) 2 real activity datasets.
- Based on our extensive evaluation, Preclude2 results in a overall 0.88 recall in detecting different types of conflicts in interventions originated from heterogeneous applications. In addition, Preclude2 also addresses personalization and context-awareness in conflict detection from health and medical advice statements and activities of daily living.

II. Problem Formulation

In order to position our solution we first carefully define *conflict* of advice. Each advice involves a set of *action*s and each action results in a set of *effects*. Before formally defining *conflict*, we need to define *object*, *action*, and *effect* of an advice, $advice_i$.

Object (o_i): Each health advice suggests either in favor of or against each in a set of objects. For example, from Table I, objects of advice 1 of case 1 are citrus fruits and green leafy vegetables. Each object of an advice can contain **sub-typical** (s_i) semantics (e.g., *green leafy* vegetable).

Action (a_i): Action is the intervention that is implied by an advice either directly as in an imperative sentence or indirectly as in a declarative sentence. Referring to Table I, the action (i.e., eating) is directly mentioned in case 1 while it is implied in advice 2 from case 4. Each action of an advice is often associated with different semantics that suggest

people when and how to perform the action. An action can specify **quantity** (q_i) (e.g., *200 mg* of coffee) of corresponding object(s). Quantity can be specified using numerical (n_i) or adverbial quantifier (f_i) (e.g., more, few). An action (a_i) can be conditional or temporal. This is specified by one or more **conditional** clauses (c_i) and/or **temporal** clauses (t_i) in an advice text. Such as, in Table I advice 1 from case 3 and case 4 suggest the time and physiological condition of the corresponding action, respectively.

Effect (e_i): An effect refers to the purpose or resulting physiological effect of an action. For example, in case 2 of Table I, a potential effect of consuming Pate made from meats is Listeriosis. Often effect can create a chain of subsequent effects, such as, primary effect, secondary effect, tertiary effect, and so on.

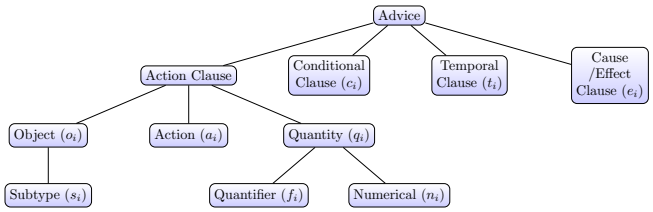

Fig. 1: Semantic decomposition of textual health advice

Thus an advice statement $advice_i$ can be expressed as a tuple of **semantic tokens**: $advice_i$: $<s_i^m$, o_i^m, a_i^m, q_i^m, c_i^m, t_i^m, $e_i^m>$as presented in Figure 1. Here m denotes the index of tuple m. Often a single advice can contain multiple objects. Then each object results in a tuple. Also, q_i^m can be either adverbial quantifier (f_i^m) or numerical (n_i^m) or both. Note that, an action and an effect of an advice can be mapped into positive or negative polarity with respect to the corresponding object. Such as, in case 2 of Table I, action in advice 1 has negative polarity with respect to *meat* while effect in advice 2 has positive polarity with respect to *red meat*.

At first we define pair-wise conflict between two advice statements. This definition can be extended to define conflict among any size set of advice.

Conflict: Two pieces of advice $advice_i$ and $advice_j$ are conflicting with each other if they have at least one common object ($o_i^m = o_j^n$) and at least one of the following is true:

1) Opposite polarity of actions, i.e., a_i^m and a_j^n have opposite polarity (case 1 of Table I).
2) Opposite polarity of effects, i.e., e_i^m and e_j^n have opposite polarity (case 2 of Table I).
3) Opposite polarity of action-effect, i.e., a_i^m and e_j^n have opposite polarity or e_i^m and a_j^n have opposite polarity (case 4 of Table I).
4) Both of the advice have the same polarity but they are quantitatively different from each other, i.e., q_i^m is not compatible with q_j^n (case 6 of Table I).

The above cases demonstrate **direct** and **quantitative** conflicts. In addition, conflicts between a pair of advice statements can be **conditional, temporal, sub-typical** based on the semantics of advice tokens. For example, the fourth pair of Table I have opposite polarity (according to rule 3 presented above) and one advice of the pair has a condition. So it is a conditional conflict. In addition to detecting conflicts, *Preclude2* also extracts such semantic refinements from potential conflicts.

Table I provides concrete examples of different types of conflicts. To begin with, case 1 presents a pair of advice statements that demonstrate conflict due to opposite actions, i.e., eating vs. not eating green leafy vegetables. Case 2 demonstrates conflict due to opposite effects, i.e., causing listeriosis vs. fighting anemia. Sometimes conflicts occur due to a physiological, temporal or contextual condition. For instance, in case 3, the conflict in performing stretching exercise is due to pregnancy. This example also demonstrates how the conflict detection should be aware of the physiological contexts (e.g., pregnancy, disease, medical history) of a user. For the pair of advice statements in case 5, a conflict occurs only for skimmed milk. Thus, it is a sub-typical conflict. Case 6 demonstrates another case of conflict that arise due to quantitative differences. This is also an example of conditional conflict.

In addition, we also provide an ontology of conflicts that can occur between an advice statement and an activity of daily living [3]. Like inter-advice[4] conflict, activity-advice conflict can be direct, conditional, temporal, and sub-typical. In addition, based on the temporal specification of advice statements, temporal conflicts can be further categorized in other classes as described below.

- **Duration:** This conflict occurs due to the violation of duration. For example, the following advice will conflict with the activity *having breakfast* if one does not wait at least 30 minutes for breakfast after taking the drug.
  Take this medication by mouth as directed by your doctor, usually once daily on an empty stomach, 30 minutes to 1 hour before breakfast. [Source: DUG of Synthroid]

- **Frequency**: This conflict occurs due to the violation of suggested frequency of an activity. The following advice will cause an activity-advice conflict unless one adheres to the suggested frequency range.
  Take this medication by mouth as directed by your doctor, usually 1-3 times a day with meals. [Source: DUG of Metformin]

- **Interval**: Interval conflict is caused by violation of suggested interval between repeated activities. Such as, *It is important to take your doses at least 6 hours apart or as directed by your doctor to decrease your risk of having a seizure.* [Source: DUG of Wellbutrin]

- **Temporal Dependency** This conflict occurs if there is a violation in temporal dependency. Such as, the following advice suggests temporal dependency of the activity *exercising* on the activity *taking medication*, i.e., taking the medicine 2 hours before exercise.
  If you are taking this medication to prevent breathing problems during exercise, take your dose at least 2 hours before exercise. [Source: DUG of Singulair]

[3]This is referred as an activity-advice conflict.
[4]Conflict between a pair of advice statements

- **Continuity**: This conflict occurs due to violation of temporal consistency across a continuous time period. While the above conflicts can occur multiple times a day, this conflict occurs over consecutive days. Such as, *Remember to take it at the same time each day.* [Source: DUG of Lipitor]

While all the examples above demonstrate conflicts that are specific to a time window, some temporal conflicts can be indefinite (i.e., does not conform to a specific time or depends on a specific activity / event). Such as,

If you are taking this medication for asthma or for both asthma and allergies, take your dose in the evening. [Source: DUG of Singulair]

Thus, we provide a **comprehensive** and **interpretable** taxonomy of conflicts which can play an important role in understanding the semantics of conflicts.

A. Context of textual health advice:

Context is defined as "*the interrelated conditions in which something exists or occurs*". In this paper, we consider different contexts of textual advice statements that are targeted to people suffering from one or more chronic conditions. In particular, we consider following three types of contexts: temporal context, spatial context, environmental context, and social context. We consider the temporal context of an advice as temporal condition as described earlier. **Spatial contexts** specify the spatial aspect of an advice, e.g., exercising indoor or jogging outside. **Environmental contexts** specify special circumstances of environment that might affect a health related intervention. For instance, the following advice is for people suffering from vestibular migraine: *Exercise-induced migraines are more likely to occur in people who are exercising in hot, humid weather, or at high altitudes.* It demonstrates the effect of the environmental context on exercising. Social contexts refer to the social aspects that might affect a health related intervention. Such as, the following advice adds constraints on social interaction for people who are currently on the drug Cyclosporine. *Avoid contact with people who have recently received live vaccines (such as flu vaccine inhaled through the nose).*

Considering context while detecting conflicts might reduce false positives and thus reduce cognitive burden of the recipients of advice, including, patients and caregivers. Such as, the following advice suggests people who are prescribed a drug named Abilify to not exercise in hot weather: *This medication may make you sweat less, making you more likely to get heat stroke. Avoid doing things that may cause you to overheat, such as hard work or exercise in hot weather, or using hot tubs.* This advice will be conflicting only with other advice statements that suggest to exercise outdoors even in high temperature. So, the people who are on Abilify can still perform exercise in an environment where temperature can be controlled favourably.

B. Personalization of textual health advice:

Health advice are often personalized based on an individual's age, gender, medical history, family medical history, past and current clinical diagnosis, lifestyle, and other physiological factors. In this work, we detect conflicts between a pair of advice in a personalized manner. Specifically, we consider the personalization factors that are mentioned in the prescription and the textual health advice statements, e.g., age, gender, medical history, allergy, drinking habit. Like the context of advice, personalization might reduce false positives while detecting conflicts. In the following advice, targeted to people suffering from Gastroesophageal Reflux Disease (GERD), some foods are discouraged, including, some citrus fruits and certain vegetables.

Fruits and vegetables are important in a healthy diet. But certain fruits can cause or worsen GERD symptoms, especially highly acidic fruits. If you have frequent acid reflux, you should reduce or eliminate your intake of the following foods: oranges, grapefruit, lemons, limes, pineapple, tomatoes, tomato sauce or foods that use is, such as pizza and chili, salsa.

This advice might be conflicting with other advice that suggest to take these citrus fruits and vegetables for their health benefits. But that conflict will be true only for people who have frequent acid reflux. Thus considering personalization factors of an individual (e.g., medical conditions of a patient) for conflict detection can reduce the number of false alarms.

III. Scope of Study

This research focuses on health and medical advice originating from online health sites and drug usage guidelines. In particular, it investigates potential conflicts in the textual health advice coming from these sources in the context of multiple chronic disease or multi-morbidity in adults and aging population, a crucial aspect of health care. The choice of our operational scenario for online health and medical advice is guided by the factors: (i) commonly occurring chronic diseases among adults and aging population, (ii) commonly occurring co-morbidity (i.e., set of multiple chronic conditions) among adults and aging population, and (iii) the potential interactions among lifestyle, diet, and medication of multiple diseases.

As the prevalence of reported chronic diseases vary across reports from different years, sources and countries [21], [22], there is no single standard list of most common chronic diseases and co-morbidity. But across different sources the following chronic diseases appear most frequently: hypertension (i.e., high blood pressure), hyperlipidemia (i.e., high cholesterol), arthritis, diabetes (both type 1 and type 2), chronic kidney disease, depression, coronary artery diseases (including myocardial infarction/heart attack), and different mental disorders. In addition, we consider co-morbidities, i.e., co-existing of more than one chronic conditions. As of center for disease control [21] the most frequent chronic disease pairs among the adults of the USA in 2009 are: (i) hypertension with hyperlipidemia, (ii) Diabetes with hypertension, (iii) Diabetes with hyperlipidemia, and (iv) hypertension with arthritis. They reported hypertension, hyperlipidemia, and diabetes as the most frequent chronic disease triad.

In order to capture realism in disease management, our operational scenario is based on real prescriptions. We collected a sample of 222 prescriptions from MTsamples [23] covering a variety of disease categories including, endocrinology, pain management, allergy / immunology, psychiatry, office visits, and diets and nutrition. These categories are chosen to capture

most common chronic conditions and co-morbidities. Each prescription is anonymized. Among these 222 prescriptions, we found only 34 prescriptions that are (i) prescribed to people suffering from multiple morbidity (including one or more common chronic diseases) and (ii) contain diagnosis of the patient and a list of prescription medications. Several prescriptions are discarded as they don't contain diagnosis and/or medication list. We need the diagnosis to collect relevant disease management advice corresponding to the diseases diagnosed in the prescription. We need the medication list to collect drug usage guidelines corresponding to each medication/drug mentioned in the prescription. Among the 34 prescriptions, we found a set of 34 chronic diseases in total where each prescription contain diagnosis of multiple diseases. We collect advice on these diseases from several authentic health websites and apps as shown in Table VIII. In addition, we found a set of 166 drugs/medications from these 34 prescriptions. We collect the drug usage guideline for each of these drugs from Medscape [24]. We use these data for our evaluation as described in Section V

IV. SOLUTION

Algorithm 1: ***ConflictDetect***(L, A_j)

Input : L, list of advice since time T;
Input : A_j, new incoming advice;
Output : conflictFlag, typeOfConflict;

H: HashMap of each advice in L and corresponding token ;
conflictFlag $\leftarrow$ false ;
$A'_j \leftarrow$ Preprocess(A_j) ;
$S_{A'_j} \leftarrow$ ExtractSemanticClauses(A'_j) ;
$S_{A_j} \leftarrow$ ExtractSemanticTokens$(S_{A'_j})$;
for *each advice statement* A_i *in* L **do**
  $S_{A_i} \leftarrow H$.getValue(A_i);
  $SO_i \leftarrow$ set of objects from S_{A_i};
  $SO_j \leftarrow$ set of objects from S_{A_j};
  if $SO_i \cap SO_j \neq \phi$ **then**
    // to check context metadata
    $Sub_i \leftarrow$ set of subjects from S_{A_i};
    $Sub_j \leftarrow$ set of subjects from S_{A_j};
    if *compatible*(Sub_i, Sub_j) ==*true* **then**
      $CO \leftarrow SO_i \cap SO_j$;
      for *each object o in* CO **do**
        conflictFlag $\leftarrow$ false ;
        $P_i \leftarrow$ AssignPolarity(S_{A_i});
        $P_j \leftarrow$ AssignPolarity(S_{A_j});
        if $P_i \neq P_j$ **then**
          conflictFlag $\leftarrow$ true ;
          DetectRefinedConflict(S_{A_i}, S_{A_j});
        else
          DetectQuantitativeConflict(S_{A_i}, S_{A_j});

Existing textual contradiction detection systems are based on statistical learning [18], [25], [19]. It is not feasible in this case, as statistical learning methods require a significant amount of labeled training data to avoid sparsity of feature space. But, there is no available dataset on conflicting health advice. Also, labelling health advice for potential conflict is intellectually more demanding than labelling potential pair of contradictory sentences.

Hence, we develop ***Preclude2***, a novel system consisting of a collection of semantic rules and a conflict detection algorithm (**Algorithm 1**) that detects conflicting pairs of advice statements and types of conflicts by analyzing the semantics of advice statements. It also detects activity-advice conflicts by considering semantics of activities and inferring advice statements. Unlike statistical learning based contradiction detection systems, *Preclude2* (i) detects conflicts in a personalized and context aware manner while utilizing relatively small amounts of training data and (ii) informs users about potential types of conflicts (e.g., temporal, quantitative) that can aid users' decision making process to resolve the conflict. Our assumption is *Preclude2* runs as a **watchdog** application in personal devices and intercepts health advice to detect conflicts and thus preclude safety risks.

At first we present our solution for inter-advice conflict detection. Later, we describe how *Preclude2* handles activity-advice conflict detection.

A. Conflict Detection in Textual Health Advice

Preclude2 uses a collection of novel semantic parsing rules to extract different semantics of an advice (Sections IV-A1, IV-A2). These rules are empirically extracted from training data and are guided by linguistic inference, e.g., the structure of sentences, co-located words and their Parts Of Speech (POS) tags [26], and grammatical relationships of the words. *Preclude2* keeps track of all previous advice a user received using a list L. Whenever the user receives a new advice (from an app/website), the advice text is parsed and a typed dependency representation of the advice is generated using the Stanford CoreNLP pipeline [27]. Next, potential conflicts between this advice and any previous advice are detected using the semantic rules and the **4-phase** solution (**Algorithm 1**) as follows.

1) Phase 1: Semantic Clause Extraction: In this phase, an advice statement is divided into four types of semantic clauses. Although there are generic clause extraction tools in NLP to extract noun and temporal clauses, we are the first to extract action, effect, and conditional clauses. We develop semantic clause extraction rules by utilizing dependency relationships found in the advice statements from training data and linguistic patterns of standard English language [28].

Action Clause: It contains action verb(s), object(s), and quantitative tokens of each object. It is further decomposed to extract these tokens (Section IV-A2).

Temporal Clause, t_i: It denotes temporal conditions or suggested point of time of an action. They are contained in Prepositional Phrases (PP). Some sample indicators of temporal expressions are: *after, before, as soon as, till, until, when, whenever, while,* and *during*. We create a lexicon of potential temporal expressions by combining lexicons from English grammar [28] and regular expressions from SUTime [29].

Effect/cause clause, e_i: It indicates the purpose of an action in an advice. In case of imperative sentences, the action and object clauses are followed by effect clauses. Here, the effect clause is denoted by prepositions of cause, including *to, as, so, because of, on account of, for, from, out of, due to,* and *in order to*. In addition, by analyzing the training data

we find other phrasal verbs that indicate purpose, e.g., *lead to, make, help in*. We create a lexicon of potential effect/cause indicators from training data and grammatical resources [28]. In addition, a set of rules is created to filter false positives in effect extraction. For example, filtering "to" when it does not indicate effect, e.g., *used to, seem to, have to, according to*.

Conditional clause, c_i: It restricts the action under some specific conditions. Conditional clauses are indicated by subordinate clauses or phrases starting with preposition, such as, *if, when, before, after, without* or verbal phrase like, *make sure*.

2) Phase 2: Semantic Tokenization of Action Clause: In second phase, an action clause is further decomposed into the following tokens: action, object, subject, and quantity. It should be noted that although there are several NLP tools for parts of speech (POS) tagging, extracting semantic token is not the same as POS tagging. For example, in advice 1 of case 3, POS tagging tools identify both "do" and "wake up" as verb, while only one of them is the desired action token, i.e., "do". Here instead of describing the rules verbatim, we present the intuition and overview of the rules for the sake of clarity.

Action, a_i: This token is present in an advice sentence if (i) the sentence is imperative or (ii) the sentence is declarative and starts with a Verb Phrase (VP), such as,

Imperative: *Include* peanut butter in your daily diet.

Declarative starting with VP: *Adding* peanut butter for cooking helps to fight anemia.

It should be noted that in case of the second advice presented above, *Preclude2* tags only *adding* as action token while in case of Parts of Speech (POS) tagging both *adding* and *cooking* are tagged as verb. *Preclude2* includes action verbs (e.g., drink, eat), phrasal verbs (e.g., stick to), and negated verbs (e.g., don't take, avoid eating) as action tokens.

Object, o_i: Extracting objects are crucial for conflict detection, as conceptual overlap (i.e., having a common object) is the precondition of conflict between a pair of advice. Objects are noun or noun phrase.[5] The key challenges in this stage are:

(i) Differentiating objects and other noun phrases (i.e., ignoring noun phrases that are not objects).

(ii) Maintaining object hierarchy: Often one advice refers to a sub-type of an object of another advice, e.g., one advice suggests avoiding dairy and another suggests eating cheese. In this case, object extraction should be aware of that cheese is a sub-type of dairy.

(iii) Finding compound objects: Often objects are compound words or phrases. For finding semantic overlap, *Preclude2* includes both simple and compound objects, e.g., mapping *apple juice* to <apple, juice, apple juice >. This is crucial due to safety critical nature of the problem (i.e., if someone is prescribed to avoid apples due to fructose intolerance, she should avoid apple juice as well).

These challenges are addressed by utilizing external knowledge base and semantic rules. Firstly, for filtering objects from non-object noun phrases we use *MetaMap*, a knowledge base to discover Metathesaurus concepts referred in text [30]. Specifically, MetaMap is customized based on training data to filter only relevant types of objects, e..g, foods, drinks, activities, diseases, and syndromes. Secondly, to maintain object hierarchy, multiple external knowledge bases are used. The topics requiring object hierarchy include, seafood [31], vegetables [32], grains [33], etc. Finally, compound objects are extracted using semantic parsing rules, e.g., if component words in a compound object are nouns, then consider all of them as candidate objects.

Some objects are negated, as in example 2 from case 5 of Table I, whole milk is a negated object. Negated object are contained in prepositional clause starting with *instead* or *rather*. Action verbs corresponding to negated objects are negated.

Often objects are associated with modifiers indicating subtype and quantity. These lead to sub-typical and quantitative conflicts. Sub-typical tokens are mapped in objects. Quantitative tokens are described later.

Subject: This token stores context metadata from apps and advice (or null in case there is no metadata). Context metadata refers to the subject to whom an advice is targeted. Subject can be specified as a header to advice or can appear in advice text. Such as, for the advice: *Men should have 30 to 38 grams fiber a day and women (aged between 18-50) should have 25 grams fiber a day.*, the two subjects are *men* and *women (aged between 18-50)*. The two subjects have different fiber requirements. Before polarity assignment, it is checked whether a pair of advice are compatible in terms of subject.

Quantitative, q_i: Quantitative clauses or phrases indicate a suggested amount of an object, frequency, or duration of suggested action. Quantitative tokens can be specific (i.e., contain numeric) or indefinite (i.e., contain only quantifier like, *few, more, plenty*). Although the coreNLP maps the quantitative tokens as adverbial Quantifier Phrases (QP) and Cardinal Numbers (CD), more level of detail is required for inferring the semantics of text. Such as, range, minimum, maximum, duration, and frequency. *Preclude2* addresses these cases.

In case of such quantitative tokens, the action is often normalized. For example, for advice "Don't sleep more than 8 hours", after finding token <duration: at most 8 hour>, the action token is revised from <Do not sleep>to <Sleep>. This ensures quantitative conflicts caused by numerical mismatch are detected, i.e., they are not overlooked due to negative polarity of action.

3) Phase 3: Assigning Polarity to Action and Effect Tokens: Polarity of an action/effect in an advice indicates whether the set of objects is encouraged or discouraged. Polarity of actions is assigned by building a customized lexicon of verbs from the training data and extending it by using verb synset from WordNet [34]. The initial positive and negative lists developed from the training data contain 18 and 21 verbs, respectively. After extending the lists using WordNet, the positive and negative lists contain 152 and 153 verbs, respectively. Then, for each action found in the test data, it is labeled as *positive* or *negative* based on it's appearance in the positive verb list or negative verb list. If the action does not appear in any list, then the polarity is labeled as *null*. In that case, polarity is assigned to the corresponding effect clause.

[5]In case of intransitive verbs (e.g., run, exercise), often there is no object in the sentence. Then we use verbs to detect conceptual overlap.

Precondition	Rule	Resulting Conflict
(1-4) Same object	$a_i \neq a_j$	Direct Conflict i.e., Opposite Polarity
	$a_j = null, a_i \neq e_j$	
	$a_i = null, e_i \neq a_j$	
	$a_i = a_j = null, e_i \neq e_j$	
(5-7) Direct Conflict	$c_i \neq null, c_j \neq null,$ c_i and c_j are not mutually exclusive	Conditional Conflict
	$c_i = null, c_j \neq null$	
	$c_i \neq null, c_j = null$	
(8-10) Direct Conflict	$t_i \neq null, t_j \neq null$ t_i and t_j are not mutually exclusive	Temporal Conflict
	$t_i = null, t_j \neq null$	
	$t_i \neq null, t_j = null$	
(11-13) Direct Conflict	$s_i \neq null, s_j = null$	Sub-typical Conflict
	$s_i = null, s_j \neq null$	
	$s_i \neq null, s_j \neq null, s_i \neq s_j$	
(14) Same Polarity	$f_i \neq f_j$, $s_i = null, s_j = null$	Quantitative Conflict
(15) Same Polarity	unit(n_i) = unit(n_j), $n_i \neq n_j$	Quantitative Conflict

TABLE II: Rules for detecting conflicts between advice A_i $<s_i^m$, o_i^m, a_i^m, q_i^m, c_i^m, t_i^m, $e_i^m>$and advice A_j $<s_j^n$, o_j^n, a_j^n, q_j^n, c_j^n, t_j^n, $e_j^n>$. The superscripts are dropped for the sake of simplicity.

The default polarity of an effect is positive. A negative effect is denoted by two patterns in the *effect* clauses. Firstly and more commonly, a negative effect is denoted by <Verb Phrase (VP), Noun Phrase (NP)> tuple, where NP is a disease, syndrome, or an unhealthy content (e.g., high calorie, trans fat, salt) and VP is a verb phrase that causes that NP. These specific <VP, NP> tuples are denoted as negative markers. Customized lexicons are built from training data and MetaMap to identify presence of negative markers. Secondly, a negative effect is also denoted by negation of verb/adjective phrases (e.g., not safe). Similar to assigning polarity to action, we build a customized lexicon of verbs and adjectives from the training data and extend it using synsets from WordNet. If any of the two aforementioned patterns is found in an effect clause, then it's polarity is *negative*.

It should be noted that although there are several existing lexicons of positive and negative words, using them results in performance deterioration in our case. Because, empirical observation confirms that this problem demands **domain specific lexicons** (Section V-A5). For example, VPs such as *cause, lead to* are found to have *negative* polarity in our training data, while in traditional settings they are *neutral*.

4) Phase 4: Conflict Detection Among Pairs of Advice: After assigning polarity to the semantic tokens of an advice, the problem is reduced to mapping the token sets to the potential cases of contradiction presented in Section II. A set of rules is developed corresponding to each case as presented in Table II. Upon detecting conceptual overlap from the semantic tokens and assigning polarity, these rules are executed. The temporal order of rule execution is as follows.

Firstly, it is checked whether the polarity of the two advice statements are opposite (rules 1-4). If they are opposite, then it is a direct conflict.

Secondly, upon detecting a direct conflict further rules are executed to check whether this conflict can be refined (lines 1.21-1.23 of **Algorithm 1**). Thus, rules for conditional conflict (rules 5-7), temporal conflict (rules 8-10), and sub-typical conflict (rules 11-13) are executed in parallel. It should be noted that a conflict can satisfy multiple rules simultaneously, e.g., a conflict can be conditional as well as sub-typical.

Thirdly, if the polarity of the two advice are the same (i.e., none of the rules 1-4 holds), then the quantitative tokens of the overlapping object(s) are checked for quantitative conflicts. These conflicts occur when two advice statements have the same polarity about a common object but differ in terms of quantity of the common object. The difference in quantity can be caused by adverbial quantifiers (e.g., few, more) with opposite polarity (e.g., one advice suggests to *eat more kale* while the other suggests to *take less kale to mitigate side effects of a medication*). Such cases are handled by rule 14. In addition, the difference in quantity can also be caused by numerical mismatch (e.g., case 6 of Table I). A pair of advice with the same polarity can be numerically conflicting if the following two conditions hold: (i) both of their quantitative tokens of the common object are numerical with the same unit and (ii) the values of the quantitative tokens are not compatible (i.e., unequal or have different ranges) (rule 15). Currently, *Preclude2* does not handle the case of numerical quantitative tokens with different units.

5) Handling Multiple Sentences: One distinguishing factor between detecting conflicting advice and detecting textual contradiction is the length of the sentences considered. In traditional contradiction detection literature, the pair of text under consideration have the same length, i.e., each of the texts contains a single sentence. But, a pair of potentially conflicting advice statements often have different number of sentences. Different sentences in an advice statement can convey different information as presented below.

(i) A pair of consecutive sentences often contain an action-effect tuple where one sentence contains a suggested action and the other contains the resulting effect(s) of the action. (ii) An action suggested in one sentence is often explained in further detail in subsequent sentence(s). (iii) An action discouraged in one sentence is often followed by one or more sentences containing alternate action(s). (iv) Consecutive sentences often suggest different actions with no common objects.

For the first two cases the semantic tokens are merged, as the sentences suggest the same action. In other cases, each sentence results in a separate tuple of semantic tokens. Thus, *Preclude2* handles multiple sentences by following linguistic intuition derived from the textual health advice domain.

B. Activity-Advice Conflict Detection

Advice statements can be conflicting with activities of daily living in different ways as discussed in Section II. Based on the temporal specification of the advice text, some conflicts are temporally indefinite and some are specific.

To detect indefinite activity-advice conflicts, the advice semantic inference described earlier in this section is used. From each advice, a set of topics covering food, activity, medication, and disease are extracted. Then the polarity of the advice with respect to each extracted topic is assigned. These steps are performed based on rules extracted from the training data. An activity-advice conflict occur when an advice discourages an activity. Thus direct, conditional, and sub-typical activity-advice conflicts are detected based on the polarity of activity topics of advice statements.

Detect temporally specific activity-advice conflicts requires understanding the semantics of advice and activities. We use a context-free grammar based approach to find the temporal specifications of advice text. The set of terminals of this grammar are as follows.

- time of the day, d: morning | evening | noon
- natural number, n: 1 | 2 | 3...
- activity, a: sleeping | eating | taking medication | ...
- prepositions of temporal dependency, p: before | after | for | apart
- unit of time slots, u: hour | minute | day
- time stamp, t: 9 am | 10.30 pm | ...

Here, d represents simple/terminal temporal specification, e.g., taking a medication in the *morning*. Similarly, t exact timestamps as in e.g., taking a medication before *9am*.

Now the temporal dependency of two activities can be expressed using following variable,

V_1: $n.u.p.a \mid p.a$

In other words, V_1 can represent temporal dependencies like *before sleeping, after eating, 2 hours before taking medicine*, etc.

Frequency of a suggested action (e.g., taking a medication 3 times a day) can be expressed using following variable, V_2: n times u

These variables can be combined to form more complex temporal specifications, such as,

V_4: $V_1.V_2 \mid V_1.V_2.V_3$,

where V_3: $n.u.p$

This expression can encode following temporal specification, T_i:

taking a medication 2 hours before meal (V_1), 3 times a day (V_2), 4 hours apart (V_3).

By using these context free grammar rules, advice text are normalized to align time log of activities. Then the temporal specifications are computed to find potential conflicts. For instance, with the normalized expression for the temporal specification T_i, the activity log is checked to detect potential violations of the specification. Potential violations are,

- dependency violation: the medication is taken after a meal
- duration violation: the medication is taken 1 hour before a meal
- frequency violation: the medication is taken 2 times a day
- interval violation: the medication is taken 6 hours apart

Any one of the violations result in an activity-advice conflict.

V. Evaluation

We evaluate conflict detection from textual health advice in two different settings. In the first setting we evaluate conflict detection from online textual health advice on general health care topics. In the next setting, we evaluate conflict detection in online textual health advice as well as activities of daily living (ADL) in case of multi-morbidities. This is based on disease specific real textual advice collected from health websites, online drug usage guidelines, and publicly available ADL datasets. For each setting, at first we describe the data collection and annotation processes. Then we demonstrate the performance of *Preclude2* . For the second setting (i.e., multi-morbidity), we also consider the effect of personalization and context-awareness in conflict detection.

A. Conflict Detection in General Health Advice

1) Data collection: The choice of our operational scenario for online health advice in general heath care is guided by three factors: (i) most popular general health topics (i.e., exercise, diet, and weight loss) [35], (ii) most common conditions and diseases for which people use these online resources (i.e., pregnancy, diabetes) [36], [37], and (iii) potential interactions between different health topics. Hence, we choose 8 health topics as presented in Table III. Although anemia and digestive health (e.g., food allergy/intolerance) do not belong to the most popular health topics, we include them in our study as (i) a significant portion of population world wide suffer from these [38], [39] and (ii) they demonstrate interactions with other topics, i.e., their advice are sometimes conflicting with advice of other selected topics (e.g., diabetes, weight loss).

In the fourth column of Table III we have listed 8 different health apps. Among these 4 are Android apps (Effective Weight Loss Guide, Healthy Nutrition Guide, Health and Nutrition Guide, and Anemia Help) and the other 4 are iOS apps. As the sources of online health advice, we have used WebMD, Yahoo! Health, MayoClinic, and HealthLine, all of which belong to top ten most popular health sites as of 2016 [40]. We have collected advice statements from these sources that primarily relate to food, exercise, life style (sleeping, drinking), and some over-the-counter drugs. Some of these advice statements are aimed at certain context, e.g., advice for lactose intolerant people. These contexts are encoded in the apps as *metadata*.

2) Ground Truth Annotation: For evaluation purposes, we split the dataset of 1156 advice statements into training and testing sets with 380 and 776 advice statements, respectively. We empirically develop the semantic decomposition rules and the conflict detection rules from the training set and evaluate the effectiveness of these rules on the test set. Potential candidate pairs in training and test sets are about (380^2=) 144K and (776^2=) 602K. Among these pairs, conflicts from cases 1-6 of Table I occur if there is at least one common topic/object between the pair of advice statements. So, for efficient ground truth annotation, we filter advice pairs that do not have any common object as follows.

For labeling ground truth, objects are manually extracted from each advice by 3 human annotators. Each object of a sentence is labeled as one of 3 classes: positive, negative, and neutral. Conflicts of cases 1-5 of Table I occur if the polarity of

Health Topic	Number of Advice		Mobile App Name
	Health websites	Mobile Apps	
Anemia	39	6	Anemia Help
Diabetes	81	18	Health & Nutrition Guide
Digestive health	30	35	Food Shopping Essential
Diet	85	256	Healthy Nutrition Guide Health & Nutrition Guide Effective Weight Loss Guide
Exercise	51	60	Health & Nutrition Guide Effective Weight Loss Guide
Pregnancy	48	92	Pregnancy Pregnancy Foods to Avoid
Weight loss	32	323	QuickWeight Effective Weight Loss Guide
Total	366	790	1156

TABLE III: Numbers of advice collected from 8 health topics from authentic health websites (column 2) and mobile health apps (column 3). The rightmost column contains the name of the mobile apps. Advice statements on a single topic were collected from several websites, so the website names are not presented here for the sake of brevity.

validated pairs	1294
pairs with gold label	1266
% of pairs with gold label	97.8
Number of conflicts (combining test and training set)	
Direct conflict	364
Refined conflict	624
Not conflict	306
Fleiss κ	
Direct conflict	0.977
Refined conflict	0.982
Not conflict	0.970
Overall	0.977

TABLE IV: Statistics for the validated pairs. A gold label reflects a consensus of three votes from the three annotators.

the two advice statements with respect to the object is opposite. The other case is quantitative conflict, which occurs when there is at least one common object with the same polarity and the quantitative tokens are incompatible. 336 and 830 pairs of potentially conflicting advice statements are found from the training set and test set, respectively. Each of these pairs has at least one common object with opposite polarity. An additional 128 pairs of advice statements are found as potential candidates for quantitative conflicts. Each of these pairs has at least one common object with the same polarity and each advice of the pair contains a numerical/adverbial quantifier corresponding to that object.

Finally, the filtered (336+830+128)=1294 pairs are annotated by 3 human annotators. The statistics of this annotation is presented in Table IV. Here the *refined* class refers to the temporal, quantitative, sub-typical, and conditional conflicts. Among the 1294 pairs of advice, 364 have direct conflicts, 624 have refined conflicts, and 306 are not conflicting. Here, out of 1294 validated pairs, 1266 pairs obtain gold label (i.e., all three annotators agree on the label). The agreement among annotators are calculated using Fleiss κ statistics [41]. κ is scaled between 0-1 and a higher value of κ indicates higher inter-annotator agreement.

3) Accuracy of Semantic Decomposition and Polarity Assignment: In this section we measure the performance of

	Accuracy
Temporal Clause Extraction	90%
Conditional Clause Extraction	95%
Effect Clause Extraction	88%
Object Extraction	97%
Action Extraction	87%
Quantitative Token Extraction	85%
Polarity Assignment	94%

TABLE V: Accuracy of Semantic Decomposition and Polarity Assignment

Preclude2 on the test data. At first we measure the performance of different components of *Preclude2* as presented in Table V. The ground truth for each token/clause is manually annotated. We measure the performance of token/clause extraction in terms of accuracy. At first, we measure the accuracy of detecting semantic clauses. Accuracy of temporal clause extraction is 90%. Accuracy of detecting conditional clauses is 95%, as most of the indicators of conditional clauses in the test set are present in the training set as well. The structure of effect clause varies widely as discussed in Section IV. The accuracy of effect clause extraction is 88%.

The accuracy of object token extraction includes the accuracy of sub-type token extraction, as a sub-type token is part of an object. Object extraction achieved an accuracy of 97%. This is because (i) MetaMap is customized to filter irrelevant objects, and (ii) the training set was a balanced representation of the test set in terms of object relation patterns. Accuracy of action token extraction is 87%. Although action extraction has fewer challenges than object extraction, more error is introduced here from parsing (System error) as the parser generated wrong labels for some of the verbs in the test set. As mentioned earlier, quantitative tokens include numerical as well as adverbial and adjective quantifiers. We find the overall accuracy of detecting different types of quantity tokens is 85%. In this case, the lexicon collected from training data was extended by adding synonyms and antonyms. However, our approach missed some unit tokens and thus resulted in comparatively lower accuracy in token extraction. Finally, the overall accuracy of polarity assignment is 94%.

4) Performance of Conflict Detection: We present the performance of *Preclude2* across different classes of conflicts in Table VI. Direct conflicts (i.e., conflicts corresponding to

Conflict types	Total Number of actual conflicts	Number of detected conflicts	Recall
Direct Conflict	254	228	0.90
Conditional Conflict	182	173	0.95
Temporal Conflict	97	78	0.80
Sub-typical Conflict	239	227	0.94
Quantitative Conflict	39	29	0.74
Numerical Conflict	19	15	0.79

TABLE VI: Total number of different types of conflicts and recall of detecting those conflicts in the test set. It should be noted a pair of advice can have multiple conflicts.

rules (1-4) in Table II) are detected with 0.9 recall. The rest are the refinements of direct conflicts (rules (5-15) in Table II). As extracting conditional clauses receives high accuracy, conditional conflicts are detected with 0.95 recall. Recall of detecting temporal and sub-typical conflicts are 0.80 and 0.94, respectively. In the last two rows of Table VI, recall of detecting two types of quantitative conflicts are shown. Recall of detecting conflicts due to adverbial quantifier and numerical quantifier mismatches are 0.74 and 0.79, respectively. As our approach is a pipeline approach, error from the quantitative clause extraction is propagated to the later phase (i.e., quantitative conflict detection).

5) Comparison with a Baseline: Considering the health application domain, the most relevant work is presented in [8] by Alamari et al., where they focus on finding contradictory claims from abstracts of medical research papers. They group the research claims together based on the topic of the claims. Claims within a group are labeled as YES or NO to denote the polarity of the proposition of a claim. Thus they reduce the problem to binary classification of claims from the same group as YES or NO, where claims from different classes in a group indicate a conflict. For classification they used unigram, bigram, sentiment, directionality (e.g., increase vs. decrease) and negation features. Unlike us, they took a statistical approach to learning features.

A binary classification is performed for baseline compatibility. As the authors in [8] consider direct contradictions only, we use only the direct conflicts from our dataset to compare *Preclude2* with the baseline method. In their evaluation they used linear Support Vector Machine (SVM) for classification. We try both linear and polynomial SVM (while varying the cost parameter 20 times ranging from 0.1 to 10) and report the best results only. For both baselines linear SVM outperforms polynomial SVM. Three standard performance metrics of classification are used here, namely, precision, recall, and F1 score (i.e., harmonic mean of precision and recall) [42]. The results are shown in Table VII.

Also in [8], they create directionality, sentiment and negation lexicon sets from the training data and use them as features for classification. Two versions of the baseline method are compared against *Preclude2* . In *Baseline1*, the original negation, sentiment, and directionality lexicon sets used in [8] are used. It results in very low recall. In *Baseline2*, additional negation, sentiment, and directionality lexicon sets constructed from our training data are combined with their original lexicon set. This results in significant increase in recall and F1 scores from Baseline1 to Baseline2. This implies the significance of using a **domain adapted lexicon set**.

Preclude2 increases the accuracy, recall, and F1 score of *Baseline2* by 1.5 times, 4.5 times, and 3 times, respectively. This is because finding conflicts in health advice requires linguistic semantics that are not used in the baseline method. *Preclude2* captures these semantics through semantic decomposition and heuristics developed from the training data. Also, the statistical method requires a larger amount of training data to reduce the sparsity of feature space.

Method	Accuracy	Precision	Recall	F1
Baseline1	58%	0.52	0.10	0.17
Baseline2	60%	0.63	0.21	0.31
Preclude2	90%	0.85	0.93	0.89

TABLE VII: Comparing our proposed solution with baseline methods: *Preclude2* increase accuracy and F1 of Baseline2 by about 1.5 times and 3 times, respectively.

B. Conflict Detection in Health Advice Related to Multimorbidities

1) Data Collection: The system is evaluated in the context of safety among patients suffering from one or more chronic conditions, i.e., how often these patients receive advice statements / interventions that are conflicting with other medical / clinical advice, health advice, and their regular activities of daily living.

The evaluation is centered around real prescription data collected from MTSamples [23]. This dataset contains anonymized prescriptions of real patients. Each prescription contains patients demographic and physiological information, medical history, symptoms, suggested treatments, and prescribed drugs, diet and lifestyle. We sampled 34 prescriptions from this source. This dataset serves as the premise of simulating an user / patient in the evaluation of Preclude2. The sampling process is mentioned earlier in Section III. The sampled data contain prescriptions suggested to patients suffering from multiple chronic diseases, including, but not limited to, endocrinological disease (e.g., diabetes, hypothyroidism), psychological conditions (e.g., bipolar affective disorder, alcohol withdrawal, anxiety, depression, lethargy, alcohol dependence, substance abuse), obesity hypoventilation syndrome, chronic pain (e.g., headache, hip 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) document from MedScape [6] [43]. Among the 166 drugs, the online drug usage guideline document is available for only 90 drugs in MedScape. We have crawled and annotated these 90 online DUG documents. The patient information found in each prescription are used for personalization of advice (e.g., finding advice that are targeted to a specific group of patients in terms of age range, gender, physiological and clinical conditions).

Subsequently, based on the list of diseases corresponding to the online DUG data, disease specific health advice statements are collected from authentic health websites, including, webMD, MayoClinic, HealthLine, NIH, CDC, and NHS, etc.

[6]Medscape is a free medical reference application available for both iOS and android devices [24]. As of, April 2017, it is also one of the most popular applications used by physicians.

Chronic Disease Name	Count of Advice	**Chronic Disease Name**	Count of Advice
Acute cystitis	22	Hypertension	81
Attention deficit hyperactivity disorder (ADHD)	23	Hyperthyroidism	33
Alcohol dependence	9	Hypothyroidism	20
Alcohol withdrawal	18	Lethargy	18
Anxiety disorder	34	Management of pain medications	6
Atherosclerotic coronary vascular disease	50	Mood swing	4
Bipolar disorder	23	Morbid obesity	172
Chronic kidney disease	35	Old myocardial infarction	72
Delirium	11	Osteoarthritis	32
Dementia	27	Pain management	25
Depressive disorder	57	Post-traumatic stress syndrome (PTSD)	22
Type 2 diabetes	37	Schizoaffective disorder	16
Dyslipidemia	27	Severe backache/ backpain	46
Gastroesophageal reflux disease	35	Obstructive sleep apnea	20
Headache	17	Substance abuse	3
Hip avascular necrosis	21	Transient ischemic aphasia	50
Hyperlipidemia	28	Vestibular migraine	30

TABLE VIII: Numbers of advice statements collected from different diseases from different authentic websites.

Dataset Name / Span	Relevant Activities	Occurrences of Activities
CASAS: Milan / 84 days	Meal	22
	Chores	23
	Read	314
	Sleep	96
	Take medicine	60
CASAS: Cairo / 57 days	Sleep	52
	Breakfast	48
	Lunch	37
	Dinner	42
	Take medicine	44

TABLE IX: Activity of Daily Living datasets used in the evaluation, namely, [44], CASAS dataset from Milan, and CASAS dataset from Cairo [45]. The first column contains name of the dataset and the duration of data collection. The second column contains the activities from the dataset that are being considered for potential conflicts with textual health advice. The numbers in parenthesis in the second column indicate the occurrences of the corresponding activity. For instance, the second row represents the dataset that CASAS dataset that was collected from Milan for 84 days. In this dataset, there are 22 instances of *meal* activity and 60 instances of *taking medicine* activity.

In total 1124 pieces of advice are collected for 34 diseases. Table VIII shows number of disease specific advice collected for each of the 34 diseases found in our sampled prescription data.

The advice suggested in online DUG documents or websites can be conflicting with activities of daily living. Such as, the DUG document for Wellburtin suggests "*To avoid trouble sleeping, do not take this medication too close to bedtime.*". Now if the patient take this medication at 8pm and falls asleep immediately after that there may be a conflicting condition. As we don't have the activity logs of the patients corresponding to the anonymized prescription dataset, we emulate patients' activity using publicly available activity log dataset. Namely, we use 2 public datasets as described in Table IX. Although there are other public datasets containing activity log, here we use only those datasets that contains the "*taking medication*" activity. Because, often activity-advice conflicts involves advice that suggests time of an activity with respect to the time when the medication is taken. Such as, a drug may suggest to *take a medication before 30 minutes of a meal*. In this case, there will be conflict between the advice and the meal activity if the user takes the medication less than 30 minutes before his usual meal time. So, we need to have datasets that have instances of the activity *taking medicine*.

2) Data Annotation: The textual advice data collected from different sources are annotated to extract different ground truth. Specifically, the online DUG data are annotated to identify type of advice, topic of advice, and polarity of the advice with respect to the topic. The health and medical advice data are annotated to identify topic of advice, and polarity of the advice with respect to the topic. Finally, all textual advice data are annotated to identify potential pairwise conflicts among them. All annotations are performed by three human annotators and the inter annotator agreement is measured in terms of kappa statistics (κ) [41]. κ is scaled between 0-1 and a higher value of κ indicates higher inter-annotator agreement. The details are presented below.

Extracting advice and annotating type of advice: The online DUG documents are large blocks of text separated in multiple paragraphs. So, at first these data are annotated to extract advice statements. We developed a tool to annotate the DUG data collected for this project. The tool presents advice sentences to the annotator and asks his/her feedback on whether the sentence is an advice or not. The annotator can specify whether the sentence is an advice. As advice text can span across multiple sentences, the annotator can link consecutive sentences that are part of a single advice through the tool. If the annotator selects the sentence as an advice, then the tool prompts for the type of advice. In total, seven types of advice are identified as follows: food or beverage related advice, activity or lifestyle related advice, exercise related advice, drug administration related advice, pregnancy related advice, disease or symptom related advice, and other medication related advice. From the 90 online DUG documents, 1005 advice are extracted and annotated. Table XI shows examples of different types of advice extracted from these documents. The details of this annotation is shown in Table X.

Annotating topics/objects of advice: All textual advice are manually annotated for advice topics by three human

Type of Advice	Count of advice
Activity or lifestyle related advice	148
Disease or symptom related advice	245
Drug administration related advice	224
Exercise related advice	49
Food or beverage related advice	253
Other medication related advice	310
Pregnancy related advice	211
Total count of advice	1005

TABLE X: Annotation of Drug usage guideline dataset. In total 90 online DUG documents are annotated where the drugs correspond to the 34 real anonymous prescriptions. Seven types of advice are found in the data as shown in column 1. The second column denotes the count of advice for each type of advice. It should be noted that a single advice can belong to multiple categories.

annotators. Each annotator extracted the topic(s) of each advice statements. The advice topics are also categorized in four classes, namely, food or beverage, medicine or drug, activity or exercise, and disease or symptoms. If all three annotators agree on the topic of an advice statement, the annotation receive a gold label. Other wise, the annotation is decided based on majority voting. The inter annotator agreement of topic annotation is 0.78 while 86% advice statements received gold labels. In addition, the three annotators labelled the polarity of each topic in three categories: positive, negative, and neutral. This annotation is used to assign the polarity of an advice with respect to its topic.

Annotating polarity of advice: Once, the topics of an advice are annotated, the polarity of the advice is annotated with respect to each topic. The polarity of an advice with respect to a topic can be positive, neutral, or negative. The inter annotator agreement of polarity annotation is 0.84 while 92% advice statements received gold labels.

Annotating conflicting advice: For evaluation we consider pairwise conflicts, i.e., conflicts between a pair of advice statements. Now, there are 1005 and 1124 advice statements collected from online DUG documents and health websites, respectively. Then the total number of textual health advice is 2129. So, the potential candidate pairs of conflicting advice is (2129^2=) 4532641. Among these pairs, conflicts occur if there is at least one common topic/object between the pair of advice statements. So, for efficient ground truth annotation, we filter advice pairs that do not have any common object as described in Section V-A2. In addition, although there 90 different drugs and 34 different chronic disease, we don't consider all possible tuples of drugs and diseases. Because, these data is based on real prescriptions, we only consider the set of drugs and diseases for potential conflicting pair that are found in one of the 34 prescriptions that we collected. Thus, the potential pairs of conflicting advice represent realism based on the prescriptions. Thus, 3346 pairs of potentially conflicting advice statements are found from the 2129 textual advice corresponding to the 34 prescriptions. Each of these pairs has at least one common topic / object with opposite polarity.

Finally, the filtered 3346 pairs are manually annotated for potential conflicts. We annotated whether there is any conflict or not. Also, if there is a conflict, we annotate the type(s) of conflict as well. The types of conflict are, direct, conditional, temporal, sub-typical, and quantitative. In total, there are 1024 pairs of conflicting advice statements out of the 3346 pairs of advice statements. Among these 1024 pairs of conflicts, there are 199 direct conflicts, 381 conditional conflicts, 59 quantitative conflicts, 566 sub-typical conflicts, and 296 temporal conflicts. The results of conflicting pair annotation is summarized in Table XII. It should be noted that the potential pairs of conflicts for DUG data are relatively much lower than the potential pairs of conflicts for online disease advice data. This is reasonable as we only consider the potential pairs of drugs that are from the same prescription and the prescriptions are prepared by professional medical care providers.

3) Performance of Conflict Detection: In this section we present the performance of detecting conflicts in textual advice and ADL using the 3 datasets presented above. At first we describe the performance of inter advice conflict detection (i.e., health websites and online DUG documents). Then we describe the performance of activity-advice conflict detection.

Inter Advice Conflict Detection For detecting conflicts between a pair of textual advice, we use the dataset described is Section V-A as the training data. Thus the training data includes the 1156 textual health advice statements. In addition, we include 10 online DUG documents such that this set of DUG documents is mutually exclusive of the set of 90 DUG documents corresponding to the prescription dataset. The test data consists of the 2129 textual advice statements corresponding to the prescriptions as described earlier in this section. Table XIII presents the result of this evaluation.

Overall, from 2129 advice statements, we find 1024 pairs of advice that are conflicting. Out of these 1024 conflicts, the number of direct, conditional, quantitative, sub-typical, and temporal conflicts are 199, 381, 59, 567, and 296, respectively. As shown in Table XII, the number of disease-disease advice conflicts are higher than both the number of drug-disease advice conflicts and the number of drug-drug advice conflicts. Preclude2 achieves higher recall for disease-disease advice conflicts across all types of conflicts. For conflicts that include drug advice (e.g., drug-drug advice conflicts and drug-disease advice conflicts), there are some advice statements whose polarity are not detected correctly by Preclude2. So, the recalls of those advice conflicts are relatively lower than the recall of conflicts among disease-disease advice statements.

Activity-Advice Conflict Detection Textual health advice statements are conflicting with activities of daily living (ADL) when they discourage a set of ADLs for health benefits. During the data annotation phase, the activities mentioned in each textual advice statement are annotated as "activity" object / topic. Later, during the polarity assignment phase their polarities are assigned as either positive or negative. The activity of daily living that has negative polarity in an advice results in a conflict with the advice.

As mentioned in section II, activities can be conflicting with advice in different ways, such as, direct, conditional, temporal. The total number of activity-advice conflicts in disease specific advice and drug specific advice are 46 and 126, respectively. The most frequently occurring activity-advice conflicts for disease specific advice statements include strenu-

Drug Name	Advice Text	Annotation
Abilify	This drug may make you dizzy or drowsy or cause blurred vision. Do not **drive, use machinery**, or do **any activity** that **requires alertness or clear vision** until you are sure you can perform such activities safely.	Activity or lifestyle related Exercise related
Actos	It is a good habit to carry **glucose tablets or gel** to treat **low blood sugar**. If you don't have these reliable forms of glucose, rapidly raise your blood sugar by eating a quick source of **sugar**, such as, **table sugar, honey, or candy, or drink fruit juice** or **non-diet soda**.	Other medication related Disease related Food or beverage related
Topamax	Do not drink **alcoholic beverages** for 6 hours before or 6 hours after **taking Topamax** extended release capsules, since alcohol may affect how well this medication works.	Food or beverage related Drug administration related
NovoLog	Check your blood sugar levels before and after **exercise**. You may need a **snack** beforehand.	Exercise related Food or beverage related
Glimepiride	**Pregnancy** may cause or worsen **diabetes**. Discuss a plan with your doctor for managing your blood sugar while pregnant.	Pregnancy related Disease related

TABLE XI: Different types of advice extracted from the online DUG data. For instance, in the second example, glucose *tablet* is one kind of medication, *low blood sugar* is a disease, and *sugar* is a food. So this advice receive the three tags: *other medication, disease, and food or beverage* related advice.

Sources of each pair of advice	Number of conflicting pairs	Number of non-conflicting pairs	Total
Drug-Drug advice	**28**	20	48
Disease-Disease advice	**683**	1940	2623
Drug-Disease advice	**313**	362	675
Total	**1024**	2322	3346

TABLE XII: Statistics for the validated pairs for the advice data related to multi-morbidities. The first column contains the sources of advice statements and the two consecutive columns contain the numbers of (i) validated conflicting pairs and (ii) validated non-conflicting pairs. For instance, the the second cell of the third row contains the number of conflicting pairs of advice statements where each advice statement of the pairs are from two different chronic diseases.

Type of Conflicts	Overall	Drug-Drug	Disease-Disease	Drug-Disease
Direct	0.91	0.67	0.88	0.96
Conditional	0.83	0.92	0.99	0.53
Quantitative	0.98	none	0.98	none
Sub-typical	0.85	0.89	0.96	0.78
Temporal	0.98	0.93	0.98	0.9

TABLE XIII: Performance of conflict detection from textual advice in terms of recall. The second column represents the overall recall for different types of conflicts. The subsequent three columns represent the recall for conflict detection in advice statements that are from only DUG documents, only disease specific websites, and both of (i) and (ii), respectively. The cells containing *none* indicate there is no conflict in the ground truth of the corresponding advice source.

ous exercising, smoking, taking long bath, and lying down after meal. The most frequently occurring activity-advice conflicts in drug specific advice statements include (i) *driving, using machinery, any activity that requires alertness* immediately after taking certain drugs, (ii) smoking, (iii) strenuous exercise, and (iv) prolonged sun exposure.

Preclude2 detects 41 activity-advice conflicts out of 46 conflicts from the disease specific advice statements. It finds 113 activity-advice conflicts out of 46 conflicts from the disease specific advice statements. Overall, Preclude2 achieves 0.89 recall for detecting activity-advice conflicts.

As mentioned in section II, some activity-advice conflicts contain temporal specification (e.g., frequency, duration, interval). For this we combine textual advice data with real activity dataset and detect such temporally specific conflicts in a personalized manner. The details of this evaluation is presented in Section V-B4.

4) Effect of Personalization: This section describes the effect of personalization in conflict detection across heterogeneous health applications. At first, the effect of personalization is considered for inter advice conflict detection. Next, the effect of personalization is considered for activity-advice conflict detection.

Personalized inter-advice conflict detection: In this section, we consider the effect of personalization in conflict detection. Personalization information from each advice statement are manually extracted during the data annotation phase. Then this information is compared against the personalization information provided in the anonymized prescriptions corresponding to each user/patient. Personalization information extracted from the textual health advice datasets used in this paper includes the following: age, gender, demographics, lifestyle (e.g., sedentary, active), substance usage habits (e.g., smoking, alcohol usage, caffeine usage), history of drug and alcohol abuse, medical diagnosis, usage of non-prescription drugs (antacids, NSAIDs, vitamin supplements), and food allergies.

The results are presented in Table XIV. We have in total 34 prescriptions. Among them, 17 contain personalized conflicts. For the other 15 prescriptions, although there are some conflicts, those conflicts are not affected by the personalization information used here. For the 17 prescriptions that have some personazlized conflicts, we rank them according to the number of total conflicts and present the top 5 prescriptions in Table XIV. It should be noted that the number of personalized conflicts does not only depend on the number of chronic conditions one has. It also depends on the overall diagnosis, the prescription medication, and disease progression.

There are in total 1595 conflicts across the 34 prescriptions used in this paper. The total number of conflicts that are affected by personalization information is 163. Among these 163 personalized conflicts, the number of drug-drug conflicts, disease-disease conflicts, and drug-disease conflicts are 9, 98, and 56, respectively. As shown in Table XII, like the whole dataset, the number of personalized disease-disease advice conflicts is relatively higher than the number of personalized drug-disease advice conflicts and number of personalized drug-drug advice conflicts. This is reasonable as the drugs are prescribed by doctors who are aware of the personalization information of the patients. Thus the advice statement related

to a prescribed drug cause fewer number of conflicts with the advice statements related to the other drugs and diseases of the prescription. On the other hand, the online disease specific advice statements are often unaware of the personalization information of the patients and thus result in higher number of conflicts.

Personalized activity-advice conflict detection This section demonstrates how personalized behavior patterns of an individual contribute to potential conflicts between textual advice and activities of daily living. In particular, we demonstrate how Preclude2 detect temporal activity-advice conflicts in a personalized manner. This experiment is based on two different types of data, (i) the textual health advice data that are personalized for each prescription and (ii) the two ADL datasets mentioned in Table IX. As most of the disease specific textual advice statements lack the temporal specifications of activities, only the drug specific textual advice statements are considered in this experiment.

The advice and activity datasets are normalized as described in Section IV-B. From the drug specific advice dataset a wide range of temporal specifications/conditions are extracted, including, (i) taking a medication before (or after) eating (or sleeping), (ii) taking a medication n minutes before (or after) eating (or sleeping), (iii) taking a medication in morning (or evening), (iv) taking a medication n times a day, and (v) taking a medication the same time of each day. Among these specifications, the frequency condition (i.e., taking a medication n times a day) is not considered. Because, in the advice dataset, n lies within the range of 2 to 4 per day, while in both of the activity datasets n lies within the range of 0 to 1 per day. So, including this specification will result in conflict for every day data was collected.

The result of considering personalization in activity-advice conflict detection is presented in Figure 2. It depicts the total number of 5 types of temporal activity-advice conflicts for both the Cairo and Milan datasets. Hence, it represent total number of 5 types of temporal activity-advice conflicts that occurred during 84 days and 57 days in Milan and Cairo datasets, respectively. Although there are in total 34 prescriptions, Figure 2 presents results for only 4 randomly selected prescriptions due to space constraints. The selection of temporal specifications shown here are based on their frequency in the advice datasets, i.e., the most frequently appearing specifications are presented here. For example, the temporal specification "same time each day" appear in 47 times in the drug specific advice dataset. It also results in higher number of conflicts, as this specification appears for majority of the drugs of each prescription.

Overall, the number of conflicts vary based on the individual activity or behavior pattern. Hence, for the same set of prescriptions the number of conflicts vary in Milan and Cairo datasets. For instance, the number of conflicts corresponding to the *same time each day* condition in Cairo dataset is over two times than the number of conflicts in Milan dataset for all four prescriptions. Another aspect of personalization consists of the (i) diagnosis of disease, (ii) prescribed drugs, and (iii) physiological conditions of an individual. Hence the prescriptions 1 and 2 (P1 and P2) result in lower number of activity-advice conflicts when compared to the other two prescriptions.

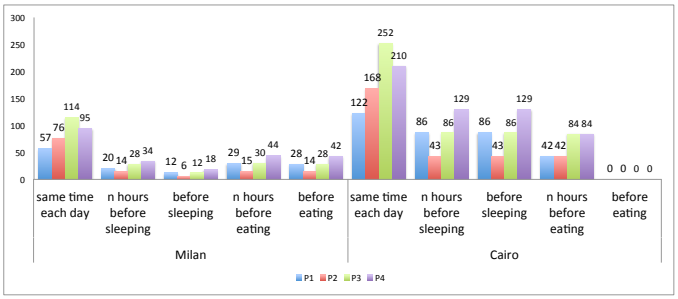

Fig. 2: Total numbers of conflicts for 5 different temporal specifications for both Milan and Cairo datasets. P1, P2, P3, and P4 represent 4 randomly selected prescriptions for which the conflicts are personalized. For instance, for the individual corresponding to prescription P3 (shown in green), 114 conflicts occur due to not taking the medication the same time over a period of 84 days in the Milan dataset. The number of conflicts vary according to the personal activity pattern or behavior.

5) Effect of Context-awareness in Conflict Detection: In this section, we consider the effect of context on conflict detection. Context information is manually annotated with each textual advice during the data annotation phase. As mentioned in Section II-A, we consider spatial, environmental, and social contexts. Context can impact both the conflicts among textual advice (i.e., drug-drug conflicts, disease-disease conflicts, and drug-disease conflicts) and conflicts among textual advice and activities. But the ADL datasets used in this evaluation do not contain contextual information, e.g., the location of the individuals when they are outside home (i.e., a spatial context), the temperature of outdoors (i.e., an environmental context), their outdoor activities, etc. So, we present the effect of context in textual advice only. But our approach can be generalized to other ADL datasets that have context information.

The set of contexts (S_c) in the advice statements that result in conflicts (n=473) is a strict subset of the set of contexts (S_a) in the entire textual advice dataset (n=2129). In other words, S_c is the set of contexts extracted from the 473 textual advice statements that result in conflicts. S_a is the set of contexts extracted from the entire textual advice dataset containing 2129 textual advice statements. S_a has higher number of contexts than S_c. As we are considering context for potential conflicts, only the contexts present in S_c are considered in this experiment. S_c includes (i) exercising in hot weather, (ii) exercising in humid weather, (iii) exercising in high altitude, and (iv) dining out. Here, (i)-(iii) discourage exercising in certain contexts and (iv) discourages consuming certain foods while dining out. Table XV shows the effect of considering these contexts in conflict detection. In this case, context has limited impact on the potential conflicts in advice that are from the same domain, i.e., drug usage guideline or disease specific online advice. The effect of context is increased when advice statements are from different domains. Overall, the effects of context in these datasets are limited as the prescriptions lack contextual information. Longitudinal context data collected over multiple dimensions of an individuals life (e.g., spatial, social, environmental) can better highlight the effect of context in conflict detection.

Prescription Item	Number of Drug-drug conflicts	Number of Disease-disease conflicts	Number of Drug-disease conflicts	Number of Personalized conflicts/total conflicts	Chronic Conditions
1	0	1	7	8 / 76	Hypertension; Hyperlipidemia;
2	0	13	11	24 / 131	BipolarDisorder; Anxiety; Post-TraumaticStressDisorder;
3	1	8	2	11 / 181	Hypertension; Osteoarthritis; MorbidObesity; SleepApnea; SevereBackpain;
4	1	20	8	29 / 193	Type2Diabetes; Hypertension; Hyperlipidemia; Headache
5	0	19	0	19 / 298	CoronaryVascularDisease; PainMedicationManagement; Type2Diabetes;

TABLE XIV: Number of Conflicts that can be affected by Personalization (columns 2-4). Here each row represents a prescription. Column 5 contains the total number of conflicts that can be affected by personalization information. Column 5 contains the total number of personalized conflicts and the total number of conflicts found in the prescription. Column 6 contains the chronic disease that are diagnosed in the prescriptions.

Context	drug-drug conflict	disease-disease conflict	drug-disease conflict
hot weather	3/28	0/683	**34/313**
humid weather	0/28	0/683	3/313
high altitude	0/28	0/683	3/313
dining out	0/28	2/683	0/313

TABLE XV: Effect of considering different contexts in conflict detection. Here, the first column lists the contexts that are considered and the first row lists the different combination of advice sources. Each cell contains the number of conflicts that are affected by the context and the total number of conflicts. For example, *hot weather* affects 34 conflicts out of the 313 conflicts in advice from drug usage guidelines and disease specific health websites.

VI. Related Works

Although we are the first to define and solve the problem of detecting conflicting health advice, there are some relevant research in Natural Language Processing (NLP) and human-in-the-loop Cyber Physical Systems (CPSs).

A. Textual Contradiction Detection

A relevant research topic in NLP is *textual entailment*, where the goal is to determine whether a given text fragment follows from another given text fragment. In existing NLP research, textual contradiction is usually defined as negative entailment. Given two pieces of text, they can be either textually entailed, contradictory, or neutral. However, most of the existing works formulate the textual contradiction detection as a binary classification task to distinguish the contradictory pairs from the non-contradictory pairs [18], [25], [19].

De Marneffe et al. provide a comprehensive taxonomy of contradiction in text that can be detected from linguistic evidence (e.g. negation, antonym, and structural or lexical disagreements) [18]. To solve the problem, they adapt a feature extraction based supervised technique where the features represent different linguistic notions of contradictions, e.g., negation, antonym, numerical mismatch, opposite polarity, etc.. Both of these above mentioned works evaluate their solution on hand-crafted balanced datasets (i.e., number of pairs of contradictions are comparable to number of pairs of non-contradictions). But in reality, contradicting pairs are very rare in a general corpus as pointed out in [19]. In another work, the authors compare the structural similarity of two sentences and detect contradiction based on minimum alignment cost between the pair [25]. A limitation of these works is they do not utilize any external knowledge which plays a vital role in detecting true contradictions as pointed by the authors in [19], [46].

In contrast to previous works, Ritter et al. perform contradiction detection on automatically extracted web data [19] and demonstrate the importance of using external knowledge base for accurate contradiction detection. They find that most seeming contradictions (99%) are not genuine contradiction at all. It requires resolving linguistic ambiguity and utilizing external knowledge to distinguish these false positives from the true positives. They define contradiction in terms of consistency: only one of the two statements are mutually consistent with world knowledge. They take a functional relation based approach where they extract a relation between the subject and the object of a sentence. However, they overlook contradictions caused by negation, numerical mismatch. Also, the accuracy of detecting contradictions largely depends on the accuracy of detecting whether a phrase is functional. But in reality, contradictory phrases are not always functional. A recent relevant research is identifying potentially contradictory claims from medical research papers [8] as presented in Section V-A5.

Shih et al. demonstrate the impact of external knowledge in contradiction detection [46]. In order to obtain essential background knowledge, they check online co-mention pattern of potential contradictory phrases. They formulated the problem as classifying claims corresponding to a specific research question as either positive or negative. They isolate functional relations from a claim that align with the statement of corresponding question. Then they extract negation, n-gram, sentiment, and directionality features from the aligned claim. As their text are highly structured and well aligned, they achieve high accuracy even when using only negation feature.

Although the existing methods utilize several important linguistic features for detecting textual contradiction, they have limited applicability in conflict detection from textual health advice. Because, **none** of the existing works provide appropriate taxonomy of conflicts that can arise while running multiple medical apps. For example, none of the existing works

defines the conflicts caused by the **cases 2-5** of Table I as conflicts.

B. Safety critical medical systems

Another related area is emerging from the domain of medical CPS. Existing medical devices are developed as monolithic stand-alone units and no widely used device interoperability standard is available. Hence, a safety and security critical real time time computing platform, named as medical application platforms (MAP), is proposed. Its goal is to (i) integrate heterogeneous devices and information systems, (ii) coordinate their actions as a system of systems, and (iii) provide execution environment for medical apps. Larson et al. provides developing and formatting requirements for MAP in the context of a smart alarm app for pulse oximetry monitoring [47]. Hatcliff et al. presents some guiding principles of MAP, e.g., interoperability, integration at run-time after deployment, extensibility, safety critical, security critical, component wise regulation, etc. [48]. *Preclude2* can be used as a component of MAP to provide conflict-free operation of multiple medical applications across multiple devices and thus ensure user safety.

C. Conflict Detection in Human Centric CPS Apps

Munir et al. focus on detecting dependencies across interventions generated by different human-in-the-loop apps (e.g., health apps, safety app) [49]. Unlike our work, they use simulated apps and structured metadata from each app. Metadata contain (i) interventions performed by each app and (ii) corresponding potential physiological parameters that might be affected by each intervention. They rely on HumMod, a physiological simulator [50], to approximate the potential effects of an intervention. HumMod uses over 7800 variables to capture cardiovascular, respiratory, renal, neural, endocrine, skeletal muscle, and metabolic physiology. But HumMod can simulate the effects of only a small set of interventions, such as, effects of only 4 drugs, effects of 2 types of exercises, and effects of taking basic nutrients (e.g., carbohydrates, protein, etc.). Also, currently it estimates the potential effects of an intervention only for a 37 year old healthy male whose weight and height are 159 lbs and 70.1 inches, respectively. It can not be personalized to any other age, gender, height, weight. Also, it does not consider the user's context (e.g., disease, physiological condition). Thus the current capability of the simulator is limited. On the other hand, *Preclude2* focuses on detecting conflicts in textual health advice using external knowledge bases and linguistic features. Although *Preclude2* currently does not detect cumulative effect conflicts (case 7 of Table I), it can be extended to detect such conflicts using advanced version of HumMod or similar sophisticated simulators that can model effects of more health interventions in a personalized manner.

VII. Conclusion

Conflicting health information is a common barrier to self management of diseases and conditions in general as well as clinical health care setting. Automatic detection of conflicts in a personalized, context-aware manner can reduce the cognitive burden of people seeking critical health information and increase health safety. So, we develop *Preclude2* a semantic rule based system to detect conflicts from an array of health applications in a comprehensive, personalized, and context-aware manner. Detecting conflicts in health advice poses syntactic as well as semantic challenges. The syntactic challenges include dealing large variation in structure and length of a pair of advice and normalizing activity and advice data according to their temporal specifications. The semantic challenges include detecting conceptual overlap between a pair of advice statements, inferring the meaning of an advice, and assigning polarity to an advice with respect to its topics. To address the syntactic challenges, *Preclude2* decomposes a given advice statement using a set of linguistic rules that are extracted from training data empirically based on linguistic references. In addition, to normalize temporal expressions of advice, it develops context free grammar based techniques. To address the semantic challenges, it utilizes semantic decomposition of advice and isolates critical components of advice as meaningful tokens. Furthermore, *Preclude2* utilizes ontologies of common health topics and linguistic concepts from multiple rich external knowledge bases to identify semantic tokens and assign their polarities.

Preclude2 is evaluated extensively by collecting and annotating real datasets from multiple sources, e.g., 1156 general health related advice from health apps and websites, 1124 chronic disease specific advice from websites, 90 online drug usage guidelines from MedScape corresponding to drugs used to treat chronic diseases, and 2 real activity log datasets. The evaluation is personalized using 34 real prescriptions of people suffering from multiple chronic diseases. Our thorough evaluation using these datasets demonstrates the effectiveness of *Preclude2* in detecting potential inter advice and activity-advice conflicts. Overall, *Preclude2* achieves 0.88 recall in detecting different types of conflicts from disease specific health advice and drug usage guidelines. It detects activity-advice conflicts with 0.89 recall. Our results also demonstrate that advice from personalized information sources (e.g., drugs prescribed by doctors) result in much lower number of conflicts with other advice. This reflects the importance of including personalization information on conflict detection, such as, physiological condition, medical diagnosis, prescription medications, and activity pattern of an individual.

We envision *Preclude2* to act as a building block for safety aware health applications. Specifically, it can be integrated in a search engine to filter / present conflicting information. Also, it can be part of a smart medication reminder system to notify users/patients about potential conflicts with activities of daily living, diet, over-the-counter medications based on their medication.

References

[1] B. Ward, J. Schiller, and R. Goodman, "Multiple chronic conditions among us adults," *Preventing Chronic Disease: Public Health Research, Practice, and Policy*, vol. 11, pp. 1–4, 2012.

[2] "The growing crisis of chronic disease in the united states," http://www.fightchronicdisease.org/sites/default/files/docs/GrowingCrisisofChronicDiseaseintheUSfactsheet_81009.pdf, accessed: 2017-05-07.

[3] C. Liddy, V. Blazkho, and K. Mill, "Challenges of self-management when living with multiple chronic conditions," *Canadian Family Physician*, vol. 60, no. 12, pp. 1123–1133, 2014.

[4] I. Poureslami, L. Nimmon, I. Rootman, and M. J. Fitzgerald, "Health literacy and chronic disease management: drawing from expert knowledge to set an agenda," *Health promotion international*, p. daw003, 2016.

[5] S. Fox and K. Purcell, *Chronic disease and the Internet*. Pew Internet & American Life Project Washington, DC, 2010.

[6] K. Hämeen-Anttila, H. Nordeng, E. Kokki, J. Jyrkkä, A. Lupattelli, K. Vainio, and H. Enlund, "Multiple information sources and consequences of conflicting information about medicine use during pregnancy: a multinational internet-based survey," *Journal of medical Internet research*, vol. 16, no. 2, p. e60, 2014.

[7] D. Kienhues, M. Stadtler, and R. Bromme, "Dealing with conflicting or consistent medical information on the web: When expert information breeds laypersons' doubts about experts," *Learning and Instruction*, vol. 21, no. 2, pp. 193–204, 2011.

[8] A. Alamri and M. Stevensony, "Automatic identification of potentially contradictory claims to support systematic reviews," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 930–937.

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

[10] V. Nobili, C. Carter-Kent, and A. E. Feldstein, "The role of lifestyle changes in the management of chronic liver disease," *BMC medicine*, vol. 9, no. 1, p. 70, 2011.

[11] M. Deacon-Crouch, I. Skinner, M. Connelly, and J. Tucci, "Chronic disease, medications and lifestyle: perceptions from a regional victorian aboriginal community," *Pharmacy practice*, vol. 14, no. 3, 2016.

[12] S. Patel and R. Dowse, "Understanding the medicines information-seeking behaviour and information needs of south african long-term patients with limited literacy skills," *Health Expectations*, vol. 18, no. 5, pp. 1494–1507, 2015.

[13] S. Savas and D. Evcik, "Do undereducated patients read and understand written education materials? a pilot study in isparta, turkey." *Scandinavian journal of rheumatology*, vol. 30, no. 2, pp. 99–102, 2000.

[14] F. Tang, G. Zhu, Z. Jiao, C. Ma, N. Chen, and B. Wang, "The effects of medication education and behavioral intervention on chinese patients with epilepsy," *Epilepsy & Behavior*, vol. 37, pp. 157–164, 2014.

[15] M. S. Wolf, T. C. Davis, H. H. Tilson, P. F. Bass III, and R. M. Parker, "Misunderstanding of prescription drug warning labels among patients with low literacy." *American Journal of Health-System Pharmacy*, vol. 63, no. 11, 2006.

[16] J. J. Mira, S. Lorenzo, M. Guilabert, I. Navarro, and V. Pérez-Jover, "A systematic review of patient medication error on self-administering medication at home," *Expert opinion on drug safety*, vol. 14, no. 6, pp. 815–838, 2015.

[17] D. Gorgos, "Nurses identify barriers to teaching patients about their medications," *Dermatology Nursing*, vol. 15, no. 6, pp. 555–557, 2003.

[18] M.-C. De Marneffe, A. N. Rafferty, and C. D. Manning, "Finding contradictions in text." in *ACL*, vol. 8, 2008, pp. 1039–1047.

[19] A. Ritter, D. Downey, S. Soderland, and O. Etzioni, "It's a contradiction—no, it's not: a case study using functional relations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 11–20.

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

[21] J. J. Ashman and V. Beresovsky, "Multiple chronic conditions among us adults who visited physician offices: Data from the national ambulatory medical care survey, 2009," *Prev Chronic Dis*, vol. 10, p. E64, 2013.

[22] M. H. Fox and A. Reichard, "Peer reviewed: Disability, health, and multiple chronic conditions among people eligible for both medicare and medicaid, 2005–2010," *Preventing chronic disease*, vol. 10, 2013.

[23] "Transcribed medical transcription sample reports and examples," http://www.mtsamples.com/, accessed: 2017-03-15.

[24] "Medscape: Search drugs, otcs & herbals," http://reference.medscape.com/drugs, accessed: 2017-04-15.

[25] D. Andrade, M. Tsuchida, T. Onishi, and K. Ishikawa, "Detecting contradiction in text by using lexical mismatch and structural similarity," in *Proceedings of the 10th NTCIR Conference*, 2013.

[26] "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 173–180.

[27] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[28] J. Eastwood, *Oxford Guide to English Grammar.* Oxford University Press, 2003.

[29] A. X. Chang and C. D. Manning, "Sutime: A library for recognizing and normalizing time expressions." in *LREC*, 2012, pp. 3735–3740.

[30] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[31] "List of types of seafood," en.wikipedia.org/wiki/List_of_types_of_seafood, accessed: 2016-03-15.

[32] "List of vegetables," simple.wikipedia.org/wiki/List_of_vegetables, accessed: 2016-03-15.

[33] "List of grains," vegetablesfruitsgrains.com/list-of-grains/, accessed: 2016-03-15.

[34] C. Fellbaum, "A semantic network of english verbs," *WordNet: An electronic lexical database*, vol. 3, pp. 153–178, 1998.

[35] P. Krebs and D. T. Duncan, "Health app use among us mobile phone owners: a national survey," *JMIR mHealth and uHealth*, vol. 3, no. 4, 2015.

[36] H. J. Seabrook, J. N. Stromer, C. Shevkenek, A. Bharwani, J. de Grood, and W. A. Ghali, "Medical applications: a database and characterization of apps in apple ios and android platforms," *BMC research notes*, vol. 7, no. 1, p. 573, 2014.

[37] V. Obiodu and E. Obiodu, "An empirical review of the top 500 medical apps in a european android market," *Journal of Mobile Technology in Medicine*, vol. 1, no. 4, pp. 22–37, 2012.

[38] P. Lam, "Anemia: Causes, symptoms and treatments," 2015. [Online]. Available: http://www.medicalnewstoday.com/articles/158800.php

[39] "Nutrition digest: Digestive issues," 2015. [Online]. Available: http://americannutritionassociation.org/newsletter/digestive-issues

[40] "Top 15 most popular health websites — june 2016," 2016. [Online]. Available: http://www.ebizmba.com/articles/health-websites

[41] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[42] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[43] "Capterra medical software blog: The top 7 medical apps for doctors," http://blog.capterra.com/top-7-medical-apps-for-doctors/, accessed: 2017-04-15.

[44] E. M. Tapia, S. S. Intille, and K. Larson, "Activity recognition in the home using simple and ubiquitous sensors," in *Pervasive*, vol. 4. Springer, 2004, pp. 158–175.

[45] D. J. Cook and M. Schmitter-Edgecombe, "Assessing the quality of activities in a smart environment," *Methods of information in medicine*, vol. 48, no. 5, p. 480, 2009.

[46] C. Shih, C. Lee, R. T. Tsai, and W. Hsu, "Validating contradiction in texts using online co-mention pattern checking," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 11, no. 4, p. 17, 2012.

[47] B. Larson, J. Hatcliff, S. Procter, and P. Chalin, "Requirements specification for apps in medical application platforms," in *Proceedings of the 4th International Workshop on Software Engineering in Health Care*. IEEE Press, 2012, pp. 26–32.

[48] J. Hatcliff, A. King, I. Lee, A. Macdonald, A. Fernando, M. Robkin, E. Vasserman, S. Weininger, and J. M. Goldman, "Rationale and architecture principles for medical application platforms," in *Cyber-Physical Systems (ICCPS), 2012 IEEE/ACM Third International Conference on*. IEEE, 2012, pp. 3–12.

[49] S. Munir, M. Ahmed, and J. Stankovic, "Eyephy: Detecting dependencies in cyber-physical system apps due to human-in-the-loop," in *12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2015, pp. 170–179.

[50] R. L. Hester, A. J. Brown, L. Husband, R. Iliescu, D. Pruett, R. Summers, and T. G. Coleman, "Hummod: a modeling environment for the simulation of integrative human physiology," *Frontiers in physiology*, vol. 2, 2011.