2018

Signal Fusion and Semantic Similarity Evaluation for Social Media Based Adverse Drug Event Detection

Hameeduddin Irfan Khaja

Signal Fusion and Semantic Similarity Evaluation for Social Media Based Adverse Drug Event Detection

Hameeduddin Irfan Khaja

Thesis submitted to the

Benjamin M. Statler College of Engineering and Mineral Resources

at West Virginia University

in partial fulfillment of the requirements

for the degree of

Master of Science

in

Computer Science

Donald A. Adjeroh, Ph.D., Chair

Yanfang Ye, Ph.D.

Saiph Savage, Ph. D.

Lane Department of Computer Science and Electrical Engineering

West Virginia University

Morgantown, West Virginia

2018

Keywords: Social Media, Adverse Drug Events, Granger Causality, Semantic Similarity

Abstract

Signal Fusion and Semantic Similarity Evaluation for Social Media Based Adverse Drug Event Detection

Hameeduddin Irfan Khaja

Recent advancements in pharmacovigilance tasks have shown the usage of social media as a resource to obtain real-time signals for drug surveillance. Researchers demonstrated a good potential for the detection of Adverse Drug Events (ADEs) using social media much earlier than the traditional reporting systems maintained by official regulatory authorities like the United States Food and Drug Administration (FDA). Existing automated drug surveillance systems have used various types of social media channels and search query logs for monitoring ADE signals.

In this thesis, we address two key performance issues related to automated drug surveillance systems. The first is to improve the ADE signal detection by analyzing signals from multiple social media channels, and the second is usage of semantic similarity to evaluate ADE narratives detected by drug surveillance systems. Most current approaches for detecting ADEs from social media rely on a single channel: forums or microblogs or query logs. In this study we propose a new methodology to fuse signals from different social media channels. We use graphical causal models to discover potentially hidden connections between data channels, and then use such associations to generate signals for ADEs. Further, prior work have not emphasized much on the language of healthcare consumers, which is often casual and informal in expressing health issues on social media. There is a high potential to miss the semantic similarity between ADE terms extracted from social media and terms from formal official narratives when the two sets of terms do not share exact text. Thus, we exhibit the usage of semantic similarity to enhance accuracy of detected ADEs, and evaluated similarity measurement algorithms developed over biomedical vocabularies in ADE surveillance domain. We experimented on a dataset of drugs which had FDA black box warnings with a retrospective analysis spanning years 2008 to 2015. The results show a better detection rate and an improved performance in terms of precision, recall and timeliness using our proposed methods.

Acknowledgements

I take this opportunity to thank my advisor and committee chair Dr. Donald Adjeroh for his valuable guidance and relentless support. He inspired me greatly in working through this project with his innovative ideas and exceptional patience. I am grateful to him for believing in me and providing me an opportunity to work in his enthusiastic research group and making my Masters program a memorable experience. I thank Dr. Yanfang Ye and Dr. Saiph Savage for being a part of my committee, and the courses they taught. They were always inspiring and the knowledge they shared helped me a lot throughout my studies.

I am thankful to Tim Mitchem for providing me an opportunity to work as a Graduate Service Assistant at CEHS, WVU. I also thank the Lane Department of Computer Science and Electrical Engineering for providing me resources and financial support through teaching and research assignments.

Also, I would like to thank my family and friends for their love and support. They were always cordial and encouraged me in my work. Last but not the least, I give thanks to the Almighty Allah for answering my prayers and giving me strength at difficult times.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Problem and Motivation

According to the United States Federal Drug Administration (FDA), an Adverse Drug Event (ADE) is defined as any sign or symptom or disease which is unintended and harmful and happens for the normal dosage of the drug [1]. The two main approaches to discover ADEs are premarketing review and postmarketing surveillance. The premarketing review is carried out before the drug is released into the market to detect any potential adverse events. In premarketing review potential risks are identified and they are communicated to the prescribers. Unfortunately, the premarketing review process does not completely identify or address all possible adverse events due to the shortcoming of duration and size, thus mostly insufficient of detecting all adverse events caused by the drug [2]. Postmarketing surveillance is carried out by pharmacovigilance teams for reporting ADEs after the drug has been released into the market.

High morbidity and mortality rates are associated with adverse drug events, and hence, pharmacovigilance serves a critical task in postmarketing surveillance [3], [4]. Existing traditional approach of reporting adverse events in postmarketing surveillance includes a centralized voluntary reporting system like U.S. FDA Adverse Event Reporting System (FAERS) [5], the Yellow Cards from the UK Medicines Agency (MHRA) [6], and VigiBase – the World Health Organization's (WHO's) ADE reporting system [7], [8]. Researchers have been working on finding and improving novel approaches for pharmacovigilance tasks besides the traditional approach by focusing on capturing real-time health data through healthcare content over Web 2.0 [9], [10]. Over the past decade many studies have used publicly available information sources: Web forums, chat rooms, blogs, social networking sites, news websites, personal webpages, and so on to detect ADEs [3].

Recent advancements in pharmacovigilance tasks have shown that the usage of social media data as a good resource to obtain real-time signals for drug surveillance [8]–[13]. Researchers have shown a good potential for the detection of ADEs using social media much earlier than the traditional reporting systems [8], [14]. In this thesis, we address two key performance issues related to automated drug surveillance systems, the first is to improve the ADE signal detection by analyzing signals from multiple social media channels, and the second is usage of semantic similarity to evaluate ADE narratives detected by drug surveillance systems against official narratives, such as those from the FDA.

Most automated drug surveillance systems detecting adverse events from social media or Web 2.0 relied on single channels [3]. One exception is Adjeroh et al., which proposed that fusing heterogeneous signals from social media channels could generate good detection rate for adverse drug events [8]. Their results were quite promising as the signal fusion system they developed utilizing Twitter and search query log signals could detect drug alerts much earlier than the FDA. In this study we propose a new methodology to fuse signals from three different social media channels: Twitter, discussion forums and FDA Adverse Event Reporting System (FAERS) based on a causality model for ADE surveillance. Many studies have exhibited the usefulness of causality models in solving similar identification problems in economics [15]. We used graphical causal models to discover potentially hidden connections between data channels and use such associations to generate signals for adverse drug events.

Also, we note that most of the work have not emphasized the issue of language usage. It is well-known that the language healthcare consumers use in expressing health issues on social media forums and microblogging websites like Twitter is often very casual and informal [16]. On the other hand, warning labels and notifications from official regulatory agencies (such as the FDA in the US) are formal documents and usually described in a language that is very carefully selected by biomedical experts. This raises a major concern as the words detected from social media channels by the surveillance systems do not exactly match with the contents of a typical FDA Black Box Warning (BBW) label or alert notification.

For many pairs of terms, there is a potential to miss the semantic similarity between social media extracted ADE terms and terms from FDA notification when two sets of terms do not share exact text. More specifically, the problem is as follows: given a formal FDA ADE narrative: $X= \{x_1, x_2, \ldots x_n\}$, and an informal ADE narrative from social media $Y= \{y_1, y_2, \ldots y_m\}$, determine the semantic similarity between X and Y. The three major issues related to semantic similarity in automated drug surveillance are: 1) How to measure semantic similarity between social media narratives and official formal documents, 2) How to use semantic similarity to evaluate the accuracy of detected ADEs, and 3) How to use semantic similarity to improve ADE signal detection. This work focuses on the first two problems. In general, X and Y could represent any two documents with words from a given language. Thus, semantic similarity can have applications in other fields such as general healthcare, automobile industry, medical devices, ecommerce, etc.

Previously, Yang et al. [11] attempted to address the problem of health consumers' language over the Internet by generating adverse drug reaction (ADR) lexicons using Consumer Health Vocabulary (CHV) – developed by Zeng et al. [16]. However, this did not address the issue comprehensively, as there are over 200 biomedical vocabularies in just UMLS (Unified Medical Language System), which also includes CHV [17]. Here, we use UMLS-Similarity program developed by McInnes et al. [18], for computing semantic similarity. It incorporates well-known

semantic similarity and semantic relatedness measures. The prominent ones include path finding measures (such as Rada et al. [19], and Wu & Palmer [20]) as well as information content (IC) measures (such as Jiang and Conrath [21], and Sánchez et al. [22]). In prior work, Park et al. evaluated vocabularies from UMLS based on diabetes-related terms extracted from social media [23]. However, it confines itself to only one subset of the vast healthcare domain. In this work we focused on evaluating all the measures listed in UMLS-Similarity and vocabularies in UMLS to determine the best combination of measures and vocabulary in computing semantic similarity for evaluating adverse drug event narratives.

1.2 Thesis Contributions

The contributions of the thesis are summarized as follows:

- A detailed study conducted on automated drug surveillance systems developed for detecting adverse drug events from social media.
- Proposed a causality-based signal fusion scheme to generate adverse drug event signals from potentially hidden connections between social media channels.
- Proposed methodology to use semantic similarity for evaluating the performance of automated drug surveillance systems against the gold standard FDA alerts.
- The results reported in this thesis have crucial implications for various stakeholder groups, including regulatory agencies like FDA, health institutes, postmarketing monitoring teams, pharmaceutical companies and consumer advocacy groups.

1.3 Thesis Outline

Chapter 2 presents a brief background and prior work related to this thesis. It is organized in two parts. The first section discusses existing automated methods for adverse drug events surveillance using social media, and the characteristics of various social media channels based on the CRUFS methodology presented by Abbasi and Adjeroh [9]. The second section describes measures of semantic similarity developed over biomedical vocabularies, an overview of biomedical vocabularies in the UMLS Metathesaurus, and related work which demonstrated the use of semantic similarity in biomedical domain.

In Chapter 3 we propose a novel methodology of signal fusion based on causality. Here, we introduce the dataset we used to generate signals followed by the methodology that explains the graphical causal model for signal fusion. We also describe the experiment setup to implement

our strategy, and finally a discussion on the results we obtained in comparison with prior work. In Chapter 4 we present our evaluation of semantic similarity measures and biomedical vocabularies for comparing the ADE narratives. We discuss various aspects of selecting and refining biomedical vocabularies to be used with similarity measures, and finally evaluating their combinations against human ratings to get the best vocabulary and measure configuration for our ADE surveillance problem domain. Lastly, concluding remarks and future directions are offered in Chapter 5.

1.4 Publication Resulting in part from this Thesis

H. I. Khaja, M. Abate, W. Zheng, A. Abbasi, and D. Adjeroh, “Evaluating Semantic Similarity for Adverse Drug Event Narratives,” in Proceedings - 2018 International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS) 2018. [Accepted]

Chapter 2

Background and Related Work

The traditional systems for ADE reporting includes MedWatch from U.S. FDA Adverse Event Reporting System [5], the Yellow Cards from the UK Medicines Agency (MHRA) [6], and the VigiBase – the World Health Organization's (WHO's) ADE reporting system [7]. Each of these system work in a very similar fashion, that includes visiting their official website and reporting the ADE in detail by filling ADE reporting forms. The ADEs submitted to MedWatch becomes part of world's largest ADE database, FDA's Adverse Event Reporting System (FAERS). This database includes 6.2 million ADE records and around 400,000 records are added each year. However, only 20,000 reports are submitted voluntarily by providers and patients each year. This varying extent of voluntary reporting is because of the lengthy, formal process which requires filling of an extensively detailed ADE reporting form which is time consuming and difficult [24]. Above all, FDA generally requires up to 44 months to detect an ADE associated with a drug [25], whereas, automated drug surveillance systems were able to successfully detect many ADEs at least 15 months earlier (with some detected 2 to 3 years beforehand) [9].

2.1 Automated Adverse Drug Event Surveillance Systems

Karimi et al. in their survey on postmarketing drug surveillance classified prior works into two main categories. The first class of methods uses social media resources to identify ADR mentions. The second delve into detection of adverse events using signal detection techniques, with the aim of reporting ADEs earlier than FDA [26].

It has been observed that existing automated postmarketing drug surveillance systems have used various social media channels including forums like: DailyStrength [13], MedHelp [11], PatientsLikeMe [14] etc., search query logs from major search engines like Google, Bing, or Yahoo!, [14], [26], the advancement of Twitter as a superior micro-blogging website, many studies have demonstrated it as a good channel for monitoring drug signals [3], [8], [13], [14]. These channels exhibit different characteristics with respect to Credibility, Recency, Frequency, and Salience, an evaluation proposed by Abbasi and Adjeroh [9]. Social media channels such as Twitter and certain health forums have lower credibility as they are prone to spam. On the other hand, forums exhibit greater salience as they are capable of containing greater background and covering more context than a 140-character tweet, and far more relative to a query

encompassing a few search terms [9], despite having lower volume of content than Twitter and search queries.

ADE signal detection from social media resources incorporates methodologies which are good enough to detect potential adverse events earlier than the gold standard FDA's MedWatch. Precision, Recall and detection time has been the prominent evaluating factors for such ADE surveillance systems. Abbasi et al. [14] in their study discuss that most of the ADE signal detection approaches use "mention models" that build ADR occurrence frequency time-series at different temporal granularities (e.g., weekly, monthly, yearly), and apply temporal association rules or z-score thresholds to the time series [14].

Many works in automated ADE surveillance have relied on evaluating individual social media source channels: forum or microblogs or search query logs, rather a combination of these channels to evaluate adverse drug events detection. When applied to a large dataset, these methods have very low detection rates. In a prior work, Adjeroh et al. demonstrated correlation-based peak labeling fusion scheme on Twitter and search query logs [8]. They showed that fusing these social media channels together could generate good detection rate for adverse events. Nevertheless, as is well known, correlation does not necessarily imply causation, neither is correlation necessary for causation [27]–[29]. We address this issue by applying causality models to fuse channel-wise time series data. Our causality problem using time series from drug-ADR references is different from traditional causality analysis: rather than the usual long-range time series [28], we focus on causality over local temporal windows.

Another important aspect of this work is to determine the accuracy of suggested adverse drug events with respect to the reference ADE narrative for which we use semantic similarity. Additionally, semantic similarity can also be used to improve the strength of ADE signals from social media channels such as microblogs, chat rooms, web forums, social networks, and so on.

2.2 Semantic Similarity Measures for Biomedical Vocabularies

Semantic similarity is defined as a relatedness measure between two terms in a taxonomy having an IS-A relationship [19]. Semantic relatedness defines functional relationships, such as PART-OF, TREATS, AFFECTED BY, and other functional relations in addition to IS-A relation. Semantic similarity measures are mainly classified into knowledge-based measures and distributional-based measures [30]. Knowledge-based semantic similarity measures are taxonomy-based measures. Typical examples include random walk, path finding, and information content (IC) measures [30]. Path finding based semantic similarity measures use the distance between two concepts in a taxonomy tree as the main objective of computing semantic similarity. A drawback of path

finding based measures is that they give equal weight to all relationships between concepts. This limitation is addressed by Information content (IC) based measures by allocating different weights to different relationships based on the information content of concepts [30]. Information content is a measure of concept specificity. Intrinsic IC measures compute information content (IC) of concepts from the taxonomic structure itself. The idea of intrinsic IC is based on the assumption that the taxonomic structure of vocabulary is organized in a comprehensive way, where concepts with many children and few parents have lower IC value than the concepts which are more specific or have less children. Random walk measures on the other hand simulate walks on a concept graph and define the relatedness on overall connectivity between concepts unlike the shortest path in path finding based and IC-based measures.

Distributional-based measures deploy a domain corpus in addition to the taxonomic structure of the vocabulary [31]. A study by Pedersen et al. proposed a distributional-based measure called context vector measure for semantic relatedness and showed that this measure outperformed knowledge-based path finding measures [31]. Sánchez et al. showed that knowledge-based intrinsic IC measures outperformed distributional measures [32]. Garla and Brandt [30] observed that these studies have methodological differences preventing a direct comparison. However, they showed that for a wide range of vocabularies and benchmarks, intrinsic IC measures performed as well or better than distributional measures. In addition, they suggested the use of UMLS vocabularies for higher concordance with human judgments. Yet, no ADE specific evaluation has been done. Moreover, the performance of similarity measures heavily depends on vocabulary chosen.

2.3 Biomedical Vocabularies in the UMLS Metathesaurus

The UMLS, or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems [17]. The Metathesaurus is the biggest component of the UMLS. It is a large biomedical thesaurus that is organized by concept, or meaning, and it links similar names for the same concept from over 200 different source vocabularies. These vocabularies are electronic versions of various thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research [33]. Some of the prominent source vocabularies in UMLS Metathesaurus includes: ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification), LOINC (Logical Observation Identifiers Names and Codes), MSH (Medical Subject Headings), RxNorm, and SNOMED CT (Systemized Nomenclature of Medicine Clinical Term).

In UMLS Metathesaurus, source vocabularies are represented by the acronym SAB (Source Abbreviation) and are organized based on concepts described by Concept Unique Identifier (CUI). CUI is the basic and most general representation of a concept or terminology wherein each CUI has its own definition/meaning, and the possible relations (REL) to other concepts are defined based on CUIs. Refer Table B.1 in Appendix B, for the list of all relationships defined in UMLS [33].

Several studies have earlier evaluated semantic similarity measures. These measures have been evaluated based on a specific standard rating coded by healthcare professionals as seen in Pedersen et al. [31] and Sánchez et al. [32], where pairs were coded by physicians and experts. In addition to this very few biomedical ontologies have been addressed in testing the semantic measures. Most efforts on this issue relied only on SNOMED CT or MSH considering these vocabularies as gold standard [30]–[32], while ignoring other biomedical vocabularies.

Pesquita et al. [34] addressed some aspects of selection of semantic similarity measures, but the work is limited to Gene Ontologies and its specific applications. In our work, we consider the use of semantic similarity measures in general biomedical applications, especially where the terms are extracted from social media healthcare resources and other microblogging websites.

For a social media generated signal, we have the language as a major concern and hence the testing on selection of semantic similarity measure and source vocabulary should be based on ratings obtained from general healthcare consumers, in addition to ratings from healthcare professionals. Thus, we used human ratings as the standard to compare the performance of each combination of measure and vocabulary configuration. The human ratings obtained for this evaluation consists of ratings from people who use social media as a primary source for health-related information as well as ratings from people who are working in healthcare industry.

Our methodology involves comparison of similarity values for every combination of semantic similarity measures and selected vocabulary configurations with human ratings as a benchmark to select the best combination as discussed in Chapter 4. Therefore, the objective of this work is to have a best combination of the vocabularies from UMLS Metathesaurus and the semantic similarity algorithms to compare the narratives in the reference document (E.g. FDA black box warnings), and the target narrative identified by the automated drug surveillance system using social media.

In this thesis, we first demonstrate signal fusion using causality model to detect Adverse Drug Events and report the detection rate using this methodology. In addition to this, we evaluate the suggested ADE narrative against FDA's black box warnings to measure the efficiency of the system using biomedical semantic similarity measures. We evaluate the efficiency in terms of precision, recall and timeliness as discussed in Chapter 3.

Chapter 3

Causality Based Signal Fusion for ADE Detection

In this chapter, we propose a new methodology to fuse signals from three different social media user-generated content (UGC) channels: Twitter, Discussion Forums and FDA Adverse Event Reporting System (FAERS) based on a causality model for ADE surveillance. For causality modeling, we used graphical causal models to represent causal relations, and then used the Granger causality test to detect potential flags for ADE. This chapter is organized as follows: Section 3.1 introduces the dataset and signals we utilized for our work. Section 3.2 describes the causality model in detail including the Granger Causality tests, ADE signal detection and evaluation of detected ADE narratives against official FDA documents. Section 3.3 discusses the experimental setup for selecting possible candidates from the signals and filtering them to obtain the potential flag for ADEs. Finally, Section 3.4 provides a discussion of our experimental results and compares it with the correlation-based signal fusion methodology described in [8].

3.1 Dataset and Signals

The methodology to generate signals has been adapted from Adjeroh et al. [8], where the authors described the signal generation process as simple drug-ADR reference model, based on a predefined list of keywords for human anatomy, drug reactions, and drug administration problems. That is, for a given data source, we consider joint references to a given drug (or its various aliases) and a keyword from each of the three keyword sets. We recorded the number of such references in terms of weeks from 2008 to 2014, and then formed a time series by normalizing these counts into empirical probabilities and z-scores.

To identify potential ADR mentions, lexicons were developed for anatomy-related terms, reactions, and drug administration keywords. The lexicons, which were developed by research assistants with backgrounds in biology and medicine, were used to tag the tweets. For example, the statement “I’ve been through headaches since I started taking Actos.” would be tagged as “I’ve been through <ANATOMY><REACTION> since I started taking <DRUG>”. For word-sense disambiguation, we used the CMU part-of-speech tagger designed specifically for tweets [35], to help improve the likelihood that anatomy and side-effect tags were applied appropriately.

For an adverse event E, given a time window $t_i \in T = \{t_1 \ldots t_g\}$, where t_g is the current time period of the analysis, and t_g is less than the eventual event time period t_e. Let D(d) represent the

number of drug names associated with event E that appear in document d. Let C = {d_1 ... d_n} signify the set of documents occurring during t_i within a given channel, where each $D(d_j) \geq 1$. Further, let $A(d_j)$, $R(d_j)$, and $M(d_j)$ represent the number of anatomy, reaction, and administration terms present in d_j, respectively. The total raw score for time t_i is then computed as:

$$s(t_i) = \sum_{j=1}^{n} (D(d_j) + A(d_j) + R(d_j) + M(d_j)) \quad (1)$$

Each $s(t_i)$ is converted to a z-score $z(t_i) = (s(t_i) - \mu_g)/\sigma_g$, where μ_g and σ_g are the mean and standard deviation, respectively, across all t_i in T plus the training period where $s(t_i) > 0$. T can vary and depending on the resolution of the signals—such as daily, weekly, and monthly time models, as well as the value of the current window time period t_g.

We computed the signals for a total of 90 Drugs which had an FDA black box warning for ADEs between 2008 and 2015. Data from three user-generated content channels was collected: twitter, forums, and FDA Adverse Event Reporting System (FAERS). Approximately 12 million tweets containing drug-name keywords spanning 2008 to 2014 were gathered through Topsy's API. Over 5 million postings from 10 popular health forums were obtained using web crawlers. The forums include: AskAPatient, CafePharma, DailyStrength, DrugBuyersGuide, Drugs.com, Drugs-Forum, eHealth, MedHelp, MedsChat, and PatientsLikeMe. The postings spanned the time period 2008 onwards. In addition to the social media signals, we used FAERS data obtained for the selected drugs for the years 2008 to 2014 and processed them as signals using the above approach.

3.2 Causality Based Signal Fusion

For causality modeling, we used graphical causal models [15], [29], [36]–[38] to represent causal relations, and then used the Granger causality test [28], [39]–[42] to detect potential causal relations. Causality between two variables, say A and B is determined by checking their relationship with a third variable, say C, in particular their informational (in)dependence with C. For example, Graph (19) in Figure 3.1 (each rectangle contains equivalent structures. Figure adapted from [15]). An arrow indicates dependence between nodes. Thus, A and B are independent, while A and C have a dependence relationship. An overall network of causal relations in a large system can then be constructed by combining triples such as (A, B, C). Three key assumptions for graphical causal models are causal sufficiency, Markov condition, and stability [12]. In particular, the Markov condition states that the probability of a node can be written by conditioning on the node's parent. Thus, given the network: A→B→C←D, the joint distribution can be written as: P (A, B, C, D) = P(A).P(D).P(B|A).P(C|B, D). Since C has two parents

B and D, both are involved in its representation. By computing all the possible joint probabilities for a given network triple, Kwon and Bessler [15] identified 11 possible classes of observationally equivalent causal structures for a given network triple. Figure 3.1 shows these classes for variables A, B and C. Each block denotes an equivalent class. For example, from Bayes theorem, we see that for Graph (12) (A→C→B), P (A, B, C) = P(A).P(C|A).P(B/C) = P(A). ([P(A|C).P(C)]/P(A)).P(B|C) = P(A|C).P(C).P(B|C). Similarly, for Graph (13) (A←C→B), P (A, B, C) = P(C).P(A|C).P(B|C). Their joint probabilities are same; thus the two graphs are equivalent.

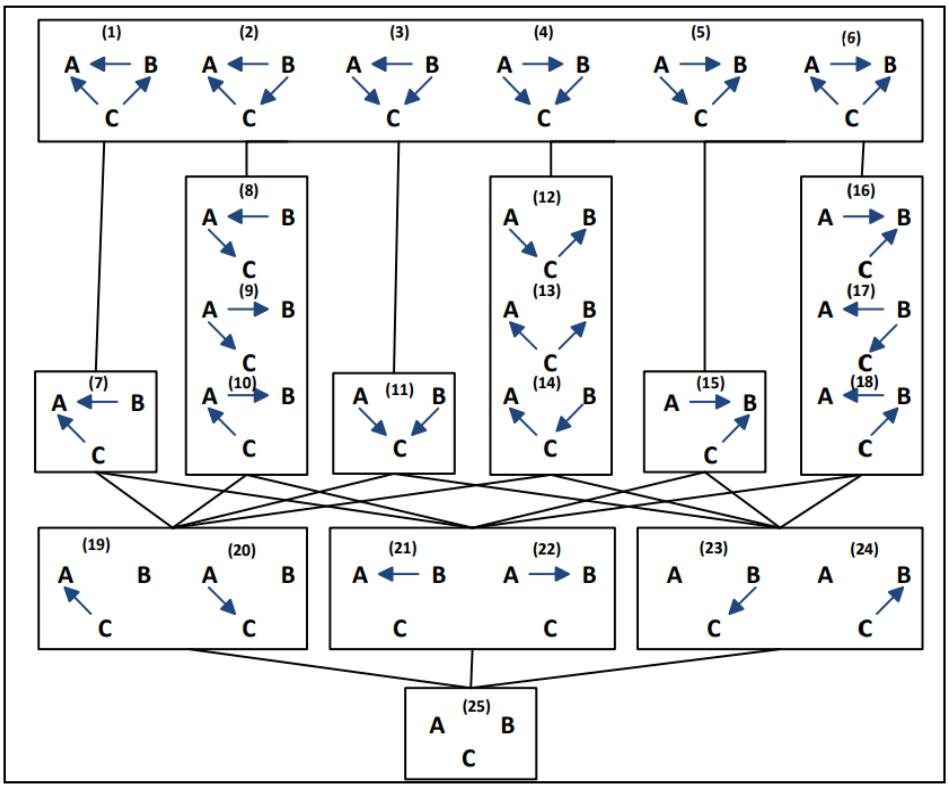

Figure 3.1: Equivalent classes in graphical causal model. (From [15])

3.2.1 Granger Causality Test

Engle and Granger [28] developed a method to check whether a given time series, say X(t) is caused by another time series, say Y(t), even when X and Y are not correlated. Here, Y(t) is said to be caused by X(t) if a series of t-tests and F-tests on lagged values of X and of Y, show that the statistically significant information about future values of Y are provided by the X values (see Figure 3.2). Basically, Y(t) causes X(t) if the future values of X(t) can be predicted more accurately using the lagged values of both Y(t) and X(t) than using only lagged values of X(t). In this work,

we model dependence based on Granger-causality. That is, A Granger-causes B (A→B) implies that B depends on A.

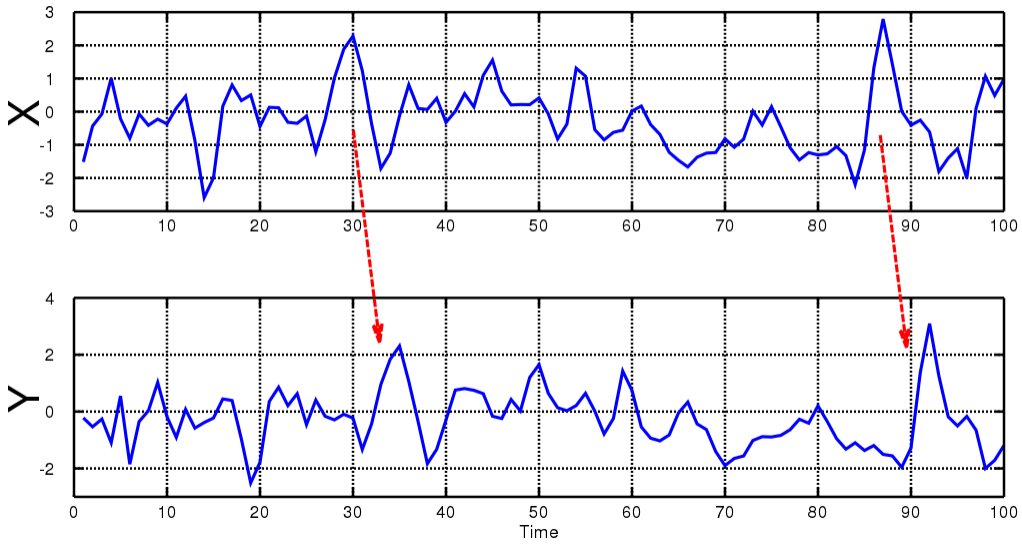

Figure 3.2: Causality between two time-series variables (X and Y).

3.2.2 Causality Based ADE Detection

From Figure 3.1, we observe three interesting classes in Graph (7) (C→A←B), Graph (11) (A→C←B), and Graph (15) (A→B←C) (see Figure 3.3). These structures are unique -- they contain unconnected colliders or v-structures. Their joint probabilities cannot be factored into other representations [15]. These three classes identify causation among the given set of variables.

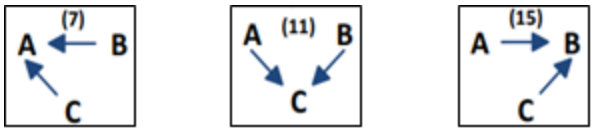

Figure 3.3: Causal v-structures.

Our problem is thus to search over the space of the causal directed acyclic graphs to identify these causal v-structures for our problem of ADRs. By specifying how the variables A, B, and C relate to our channels Forums, FAERS and Twitter, we convert our problem of signal detection to that of finding causal v-structures. Here, the nodes (variables) will correspond to

Forums, FAERS and Twitter, signals for each drug. We propose the following steps: (1) Map A, B, and C (network node triples) appropriately to our channels for each drug; (2) Identify pairwise dependencies (cointegration or causation) between variables using the Granger model based on a defined threshold for F-value (τ_f) and p-value (τ_p) for the F-tests; (3) Analyze results to determine causal v-structures. Each local region where a causal v-structure is detected becomes a candidate for an ADE. For each drug, we compute the candidates for the three v-structures. The potential flags are identified as the candidates occurring in at least one v-structure.

For a given drug, the month with the highest number of flags denotes the detection of ADE as alert signal. The time specified by the alert signal will be considered as the detection time for ADE. We search the anatomy and reaction terms based on the detection time of the flag across all the three channels: Twitter, Forums and FAERS, and accumulate unique terms for both anatomy and reaction categories. We then match the obtained anatomy and reaction terms for the ADE against the anatomy and reaction terms given in FDA's black box warning for the drug to evaluate precision and recall of the detected ADE terms. For computing precision and recall, we apply semantic similarity algorithms from the biomedical domain (discussed in chapter 4).

3.3 Experiment Setup

We obtained signals as described in Section 3.1 for a total of 335 weeks ranging from 2007-12-30 to 2014-05-25 from three channels: Forums, FAERS, and Twitter for the list of 90 drugs shown in Table A.1 in Appendix A. Of the 90 drugs there were 16 drugs which had multiple black box warnings on different dates and hence we analyzed them for each of the dates.

For each drug, we compute the Granger causality across the permutation of pairs formed by the 3 channels using the *grangercausalitytests* program from *statsmodels* package in Python [43]. We specify the input parameter maximum lag as 3 and we design our Granger test for multiple window sizes (Δ = 8, 10, 12, 14, and 16 weeks). Graphically, our Granger model for testing causality between any two social media channels (say Forums(A) and Twitter(C)) can be depicted in Figure 3.4. Our causality testing for ADE surveillance is different from traditional causality analysis which usually focuses on long-range time series [28]. The figure explains that we test Granger causality over local temporal windows defined by the window size Δ. The whole experiment is performed for both overlapping as well as non-overlapping windows for the 335-week dataset.

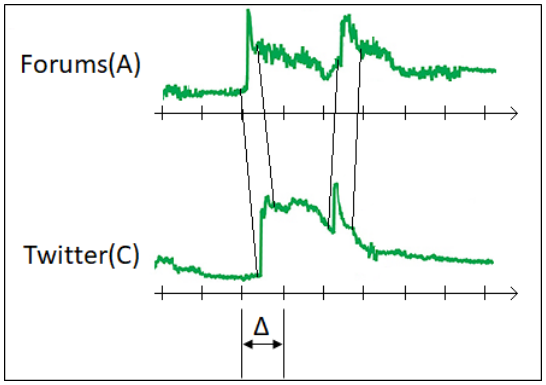

Figure 3.4: Implementation of Granger causality tests over local temporal windows.

3.3.1 Selecting Candidates for V-Structures

We now analyze the Granger test results for the pairs of channels by forming the v-structures for each window size separately. We defined our threshold for Granger results: τ_f as 2.5 and τ_p as 0.15 a slightly moderate one to get more combination of flags. The set of candidate windows in v-structure (say A→C←B) are added from both Granger tests, A→C and B→C. For each window having any lag satisfying thresholds (τ_f, τ_p) in Granger test A→C, we select the corresponding nearest window from B→C which satisfies (τ_f, τ_p), as candidates in A→C←B. Likewise, for each window having any lag satisfying thresholds (τ_f, τ_p) in Granger test B→C, we select the corresponding nearest window from A→C satisfying (τ_f, τ_p), (see Figure 3.5).

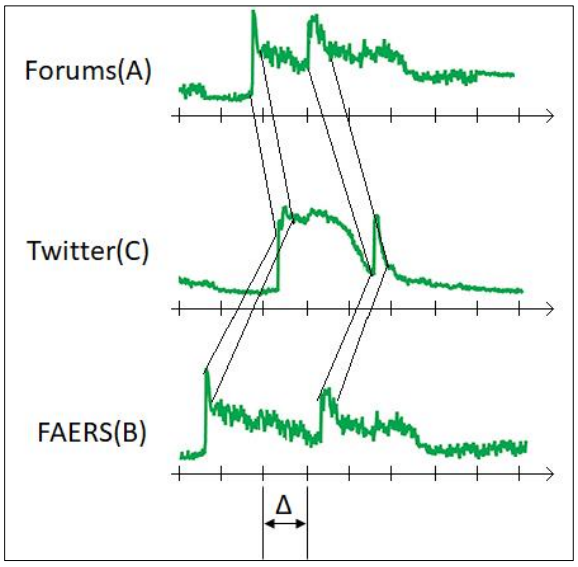

Figure 3.5: Forming candidates in causal v-structure (A→C←B).

To automate this process, we developed the procedure for candidate selection as shown in Figure 3.6 for choosing the candidate set of v-structures.

Procedure for Selecting Candidates in V-Structures

Algorithm FormStructure(A→C, B→C):
1: Consider v-structure A→C←B (say A represents Forum, B represents FAERS, and C represents Twitter) and Δ be the window size (for overlapping windows Δ =1).
2: Scan the results of A → C satisfying thresholds to get candidates, A_candidates
3: D = SelectCandidates(A_candidates, B→C)
4: Scan the results of B → C satisfying thresholds to get candidates, B_candidates
5: D1 = SelectCandidates(B_candidates, A→C)
6: Merge D & D1, store complete candidate set for A→C←B for window size Δ.

Algorithm SelectCandidates(A→C, B→C):
1: Initialize set D.
2: **for** each row A_ID in A_candidates:
// to find corresponding row in B → C, A_corres
3: set B_ID = A_ID
4: **for** B_IDArray = [B_ID – 2Δ, B_ID – Δ, B_ID, B_ID + Δ, B_ID + 2Δ]
5: rowID=B_IDArray[2]
6: **if** rowID has a candidate:
7: A_corres = rowID
8: **break**
9: rowID1= B_IDArray[1], rowID2= B_IDArray[3]
10: **if** rowID1 or rowID2 has a candidate:
11: **if** rowID1 has a candidate, A_corres = rowID1
12: **else** A_corres = rowID2
13: **break**
14: rowID1= B_IDArray[0], rowID2= B_IDArray[4]
15: **if** rowID1 or rowID2 has a candidate:
16: **if** rowID1 has a candidate, A_corres = rowID1
17: **else** A_corres = rowID2
18: **break**
19: A_corres = B_ID
// if none of the neighbors is found
20: **end if**
21: **end for**
22: merge A_candidates and A_corres, add to D.
23: **end for**
24: **return** D

Figure 3.6: Candidate selection algorithm.

Figure 3.7 shows the graphical representation of the candidate selection process for the v-structure A→C←B. As shown in figure we have n windows for Granger test results for A→C and B→C, here n varies on the number of weeks and also the window size Δ. For each window (say x) having any lag satisfying thresholds (τ_f, τ_p) in A→C we find the corresponding window in B→C such that any lag in the window y=x or its closest neighboring windows (y-2Δ, y-Δ, y+Δ, y+2Δ) satisfies the thresholds (τ_f, τ_p). The process is repeated for windows in B→C, and we form such (x,y) candidates from A→C and B→C for the v-structure A→C←B.

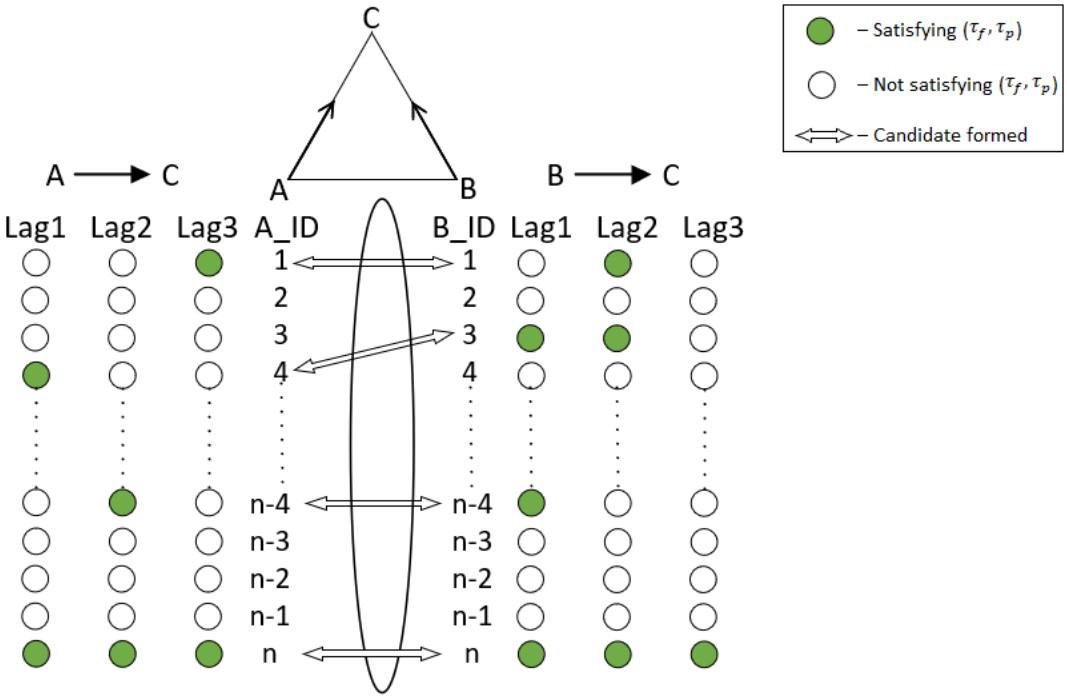

Figure 3.7: Candidate selection for A→C←B.

Using this procedure, we compute candidate sets for all the window sizes (Δ = 8, 10, 12, 14, 16 weeks), for the v-structures: A→C←B, A→B←C, and B→A←C.

3.3.2 Finding Potential Flags

We now report the potential flags for each drug representing a potential ADE. One can clearly observe that the procedure used to select candidates for forming the v-structures as discussed in Section 3.3.1 is not strict, as any window having a single lag satisfying the threshold (τ_f, τ_p) is considered to be a candidate. Thus, there is a need to filter the candidates before processing them for finding potential flags. In theory, it is desired to have all lags satisfying the threshold for a selected candidate, but this would be too strict and could miss some candidates.

In practice, we made the filtering process flexible by varying threshold (τ_f, τ_p) and defining rules based on the number of lags satisfying a given threshold (τ_f, τ_p). Given a v-structure A→C←B, we define rules Rule(α, β) as follows: For a selected candidate in A→C←B, at least α number of lags should satisfy (τ_f, τ_p) on Granger test A→C and at least β number of lags should satisfy (τ_f, τ_p) on Granger test B→C, and vice versa. We choose rules:

1. Rule(1,1),
2. Rule(2,1),
3. Rule(2,2), and
4. Rule(3,1).

where each rule would indicate the number of lags that satisfies (τ_f, τ_p) for Granger tests of a candidate in v-structure.

Figure 3.8 shows an example for Rule(2,1). Here the Rule(2,1) for v-structure A→C←B, would only select candidates satisfying thresholds for at least 2 lags on Granger test A→C and at least 1 lag on Granger test B→C, and at least 1 lag on A→C and at least 2 lags on B→C.

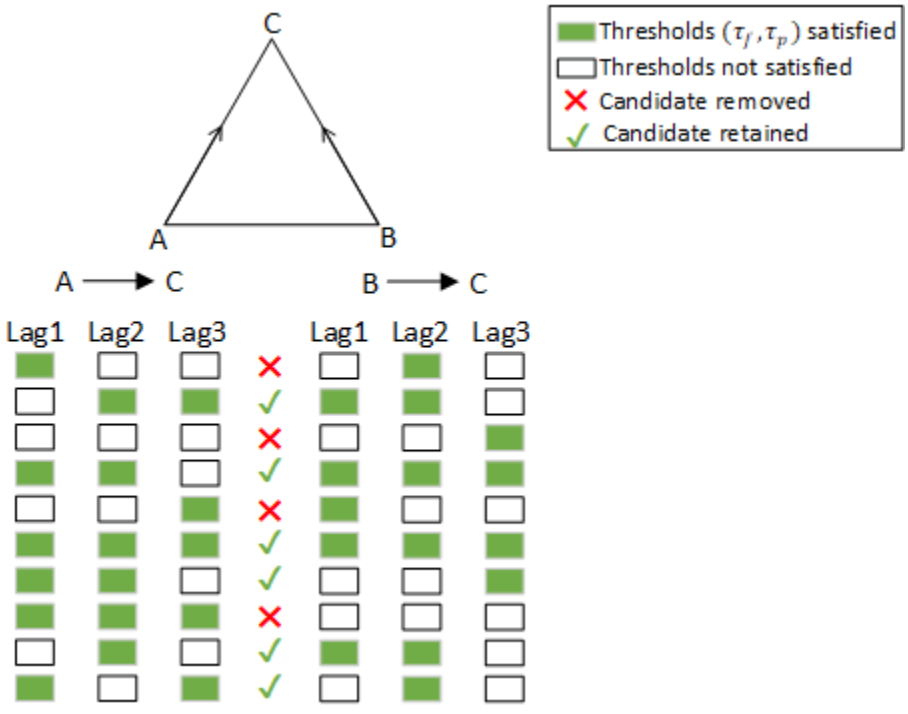

Figure 3.8: Filtering candidate selection for A→C←B using Rule(2,1).

Empirically we decided two thresholds:

1. τ_f=2.5 with τ_p=0.15 for all the 4 lag rules, and
2. τ_f=3.0 with τ_p=0.05 only for Rule(1,1) and Rule(2,1).

With the above threshold and lag rule combination we have 6 separate settings to filter candidate sets from v-structures:

1. τ_f=2.5 with τ_p=0.15 for Rule(1,1),
2. τ_f=2.5 with τ_p=0.15 for Rule(2,2),
3. τ_f=2.5 with τ_p=0.15 for Rule(2,1),
4. τ_f=2.5 with τ_p=0.15 for Rule(3,1),
5. τ_f=3 with τ_p=0.05 for Rule(1,1), and
6. τ_f=3 with τ_p=0.05 for Rule(2,1).

For each drug we set the target date as 3 months before FDA's black box warning date. The evaluation of potential flags for each setting is based on the filtered candidate set for v-structures, such that the selected candidate for a flag should be present in at least two v-structures and at least one signal in it should end before the target date.

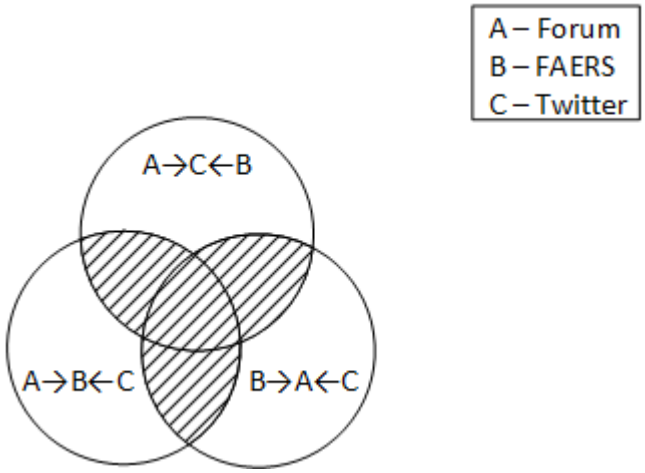

Figure 3.9: Finding potential flags from v-structures.

For a given flag if both the signals in it ends before the target date then we consider it as red flag otherwise it is considered to be a yellow flag. We follow the procedure shown in Figure 3.10 in filtering the candidate set and evaluating potential flags.

Procedure for finding Flags from Candidate Set

Algorithm FindFlags(A→C←B, A→B←C, B→A←C):

1: Consider setting (τ_f, τ_p) as threshold with lag rule Rule(n_1, n_2) and Δ as window size. Take the target, as 3 months before FDA's black box warning date. Initialize list D1.
2: d1=FilterCandidates(A→C←B, τ_f, τ_p, n_1, n_2, Δ)
3: d2=FilterCandidates(A→B←C, τ_f, τ_p, n_1, n_2, Δ)
4: d3=FilterCandidates(B→A←C, τ_f, τ_p, n_1, n_2, Δ)
5: s=MultipleOccurrence(d1, d2, d3)
6: flags=SearchFlags(d1, d2, d3, s)
7: <red_flags, yellow_flags> = RedYellow(flags, target)

Algorithm FilterCandidates(X, τ_f, τ_p, n_1, n_2, Δ):

1: Let A_ID, A_f_1, A_p_1, A_f_2, A_p_2, A_f_3, A_p_3 denote attributes for left-hand side of X. Let B_ID, B_f_1, B_p_1, B_f_2, B_p_2, B_f_3, B_p_3 denote attributes for right-hand side of X.
2: Initialize D as empty dataframe.
3: **for** candidate rows, C in X:
4: Initialize $count_1$ = 0, $count_2$ = 0.
5: **for** i=1 to 3:
6: **if** A_f_i >= τ_f and A_p_i <= τ_p, increment $count_1$, **end if**
7: **if** B_f_i >= τ_f and B_p_i <= τ_p, increment $count_2$, **end if**
8: **end for**
9: **if** ($count_1$ >=n_1 and $count_2$ >=n_2) or ($count_1$ >=n_2 and $count_2$ >=n_1) add C to dataframe D, **end if**
// D is filtered candidate set for v-structure X.
10: **end for**
11: **return** D.

Algorithm MultipleOccurrence(d1, d2, d3):

1: Initialize s as empty set
// Finding potential flags.
2: **for** dataframe, X in [d1, d2, d3]:
3: initialize s_1 as empty set
4: **for** candidate row, C in X:
// to add unique signals from the v-structure X
5: **if** A_ID is not in s_1, append A_ID to s_1, **end if**
6: **if** B_ID is not in s_1, append B_ID to s_1, **end if**
7: **end for**
8: append s_1 to s.
9: **end for**
10: **for** j in s:
11: **if** count(j) < 2, remove j from s, **end if**

```
        // candidates from multiple v-structures
12: end for
13: return s

Algorithm SearchFlags(d1, d2, d3, s):
1: Initialize K as empty dataframe.
2: for each dataframe X, in [d1, d2, d3]:
3:      for candidate row, C in X:
4:          if A_ID in s, append C to K, end if
5:          if B_ID in s and A_ID not in s, append C to K, end if
6:      end for
7: end for
8: return K

Algorithm RedYellow(flags, target):
1: for each row, C in flags:
2:      if A_endDate < target and B_endDate < target, append C to red_flags.
3:      else if A_endDate < target or B_endDate < target, append C to yellow_flags.
4:      end if
5: end for
6: return <red_flags, yellow_flags>
```

Figure 3.10: Algorithm for finding potential flags.

Following the above algorithm, we computed red and yellow flags for the combinations of window sizes 10, 12, 14 and 16 weeks with the 6 rule settings defined above.

3.3.3 Evaluating ADE Narratives

Now that we obtained potential flags for all drugs using different settings and window configurations, we compute the month which has the highest number of potential red flags to be considered as the time for the alert signal for the drug. We also specify that in the absence of red flags, we consider the month which has the highest number of yellow flags as the alert signal. The complete methodology is summarized in Figure 3.11.

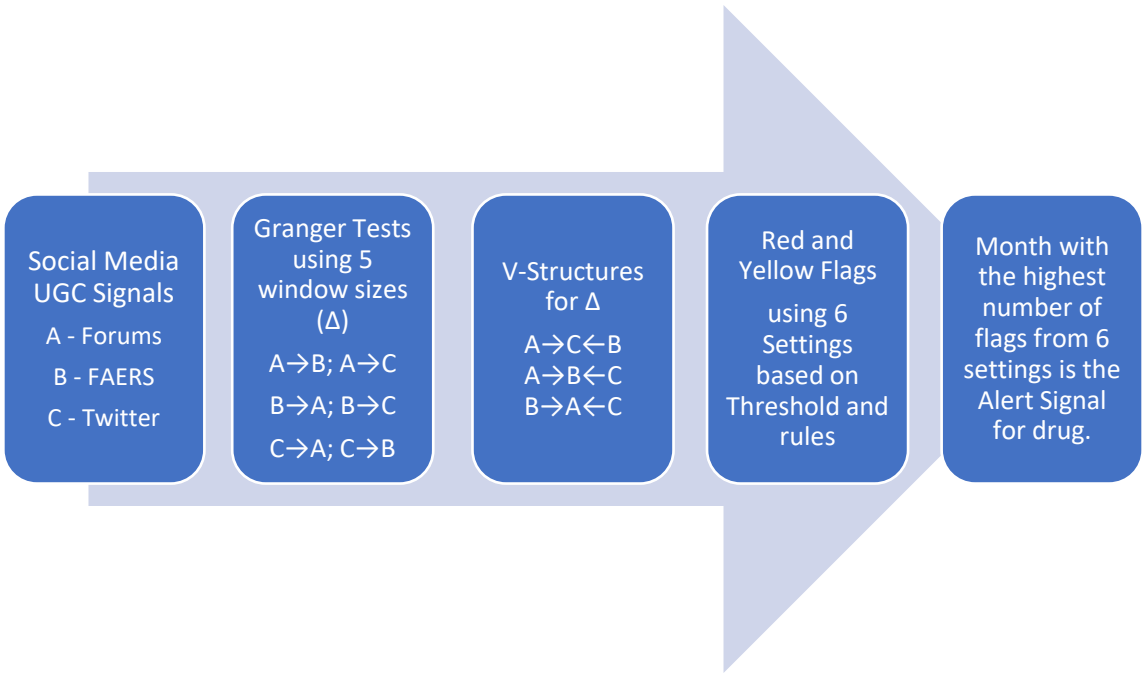

Figure 3.11: Methodology for causality-based signal fusion.

Based on the alert signal month for each drug, we extract anatomy and reaction terms from all the 3 channels: Twitter, Forums and FAERS. The extracted anatomy and reaction terms are again processed to remove redundancy. Finally, we use Semantic Similarity measure *sanchez* with CHV-SNOMEDCT vocabulary configuration (refer Chapter 4) for evaluating our social media-based ADE narratives against the official FDA documents. The performance is evaluated in terms of detection rate, precision and recall for both overlapping windows and non-overlapping windows setup.

3.4 Experiment Results and Discussion

3.4.1 Experiment Results

As mentioned in the methodology section, we evaluated our approach using overlapping as well as non-overlapping window configurations for the time series data of the 90 drugs having a total of 107 FDA black box warnings. Table 3.1 and Table 3.2 represent a detailed result for detecting red and yellow flags for the complete dataset.

Table 3.1: Detection using non-overlapping windows.

Setting		Total BBW Detected	Detected as Red	Detected as Yellow	Maximum Flags for a Drug
(τ_f, τ_p)	Rule				
(2.5, 0.15)	(1, 1)	65	55	10	96
(2.5, 0.15)	(2, 1)	54	47	7	41
(2.5, 0.15)	(2, 2)	20	18	2	8
(2.5, 0.15)	(3, 1)	23	19	4	12
(3.0, 0.05)	(1, 1)	37	31	6	15
(3.0, 0.05)	(2, 1)	15	12	3	5

Table 3.2: Detection using overlapping windows.

Setting		Total BBW Detected	Detected as Red	Detected as Yellow	Maximum Flags for a Drug
(τ_f, τ_p)	Rule				
(2.5, 0.15)	(1, 1)	62	62	0	661
(2.5, 0.15)	(2, 1)	52	52	0	203
(2.5, 0.15)	(2, 2)	20	20	0	17
(2.5, 0.15)	(3, 1)	32	32	0	36
(3.0, 0.05)	(1, 1)	46	45	1	60
(3.0, 0.05)	(2, 1)	14	14	0	24

For each configuration we computed the month which has the highest number of potential red flags of all the settings as alert signal for the drug. We also specify that in the absence of red flags, we consider the month which has the highest number of yellow flags as the alert signal. Thus, for each drug we obtain alert signal. We evaluate the performance of the

system in terms of Precision and Recall using the semantic similarity techniques. We evaluated our results for the drugs dataset discussed in Section 3.1. Table 3.3 shows these results summarized in terms of mean and median statistics over all the drugs for both overlapping and non-overlapping window configurations. Here we present the performance factors in terms of detection rate: the proportion of drugs identified as having a potential Adverse Drug Event. We also evaluate the suggested ADE narrative against the gold standard FDA black box warnings by computing the precision and recall for anatomy and reaction terms.

Table 3.3: Performance statistics for causality configurations.

Window Configuration	Detection Rate	Statistic	Anatomy		Reaction	
			Precision	Recall	Precision	Recall
Overlapping	0.62	Mean	0.1883	0.5637	0.1778	0.5339
		Median	0.1667	0.6667	0.1667	0.5000
Non-Overlapping	0.63	Mean	0.1972	0.4524	0.1755	0.4649
		Median	0.1434	0.3765	0.1429	0.5500

Timeliness or detection time is another perspective to evaluate an automated drug surveillance system. Detection rate tells us whether the considered approach is able to detect the adverse drug event or not, however one would also like to see how early the adverse drug events gets detected; so that it helps the concerned authorities like FDA to respond. Thus, we computed the detection time for each black box warning in terms of number of months prior to the FDA. Figure 3.12 shows the list of drugs detected with their earliest detection time prior to the FDA's black box warning for overlapping as well as non-overlapping windows.

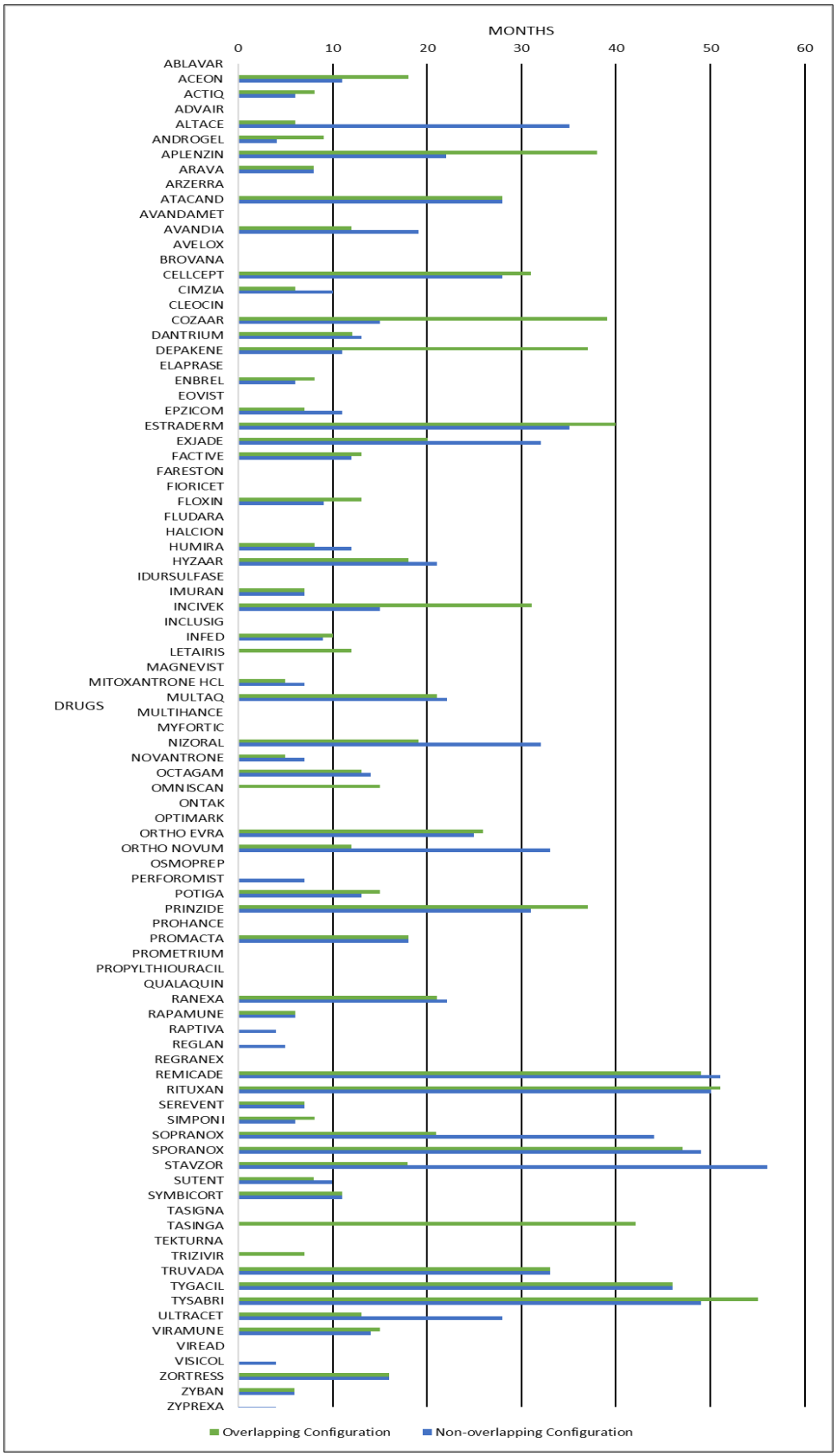

Figure 3.12: Detection time for drugs.

3.4.2 Comparison with Prior Work

We have also compared our results against the correlation-based peak labeling fusion scheme for search engine query log terms and Twitter data used by Adjeroh et al. [8]. One key observation here is that Adjeroh et al. [8] used the dataset which had only 46 drugs experimented for FDA alerts. On the other hand, the drug dataset for this work is based on FDA blackbox warnings (BBWs) for a total of 90 different drugs. Furthermore, 16 drugs had more than one black box warning (the drug Aceon, had three black box warnings) making it a total of 107, which is more than double the size of dataset used in [8]. Thus, a direct comparison of all the performance factors cannot be done; however, the detection rate and the timeliness of detection provide a fair measure to compare how well the ADEs have been detected. Figure 3.13 and Table 3.4 show the comparison in terms of detection rate. (first 5 rows are described in [8]).

Table 3.4: Comparing detection rate for fusion techniques.

Fusion Technique	Detection Rate
fuse([Q, T], [52,104])	0.6522
fuse([Q, T], [n,52,104])	0.5435
fuse[Q, T], [n])	0.3478
fuse([Q], [n])	0.4783
fuse([T], [n])	0.3261
Causality-based (overlapping)	0.6222
Causality-based (non-overlapping)	0.6333
Causality-combined	0.6777

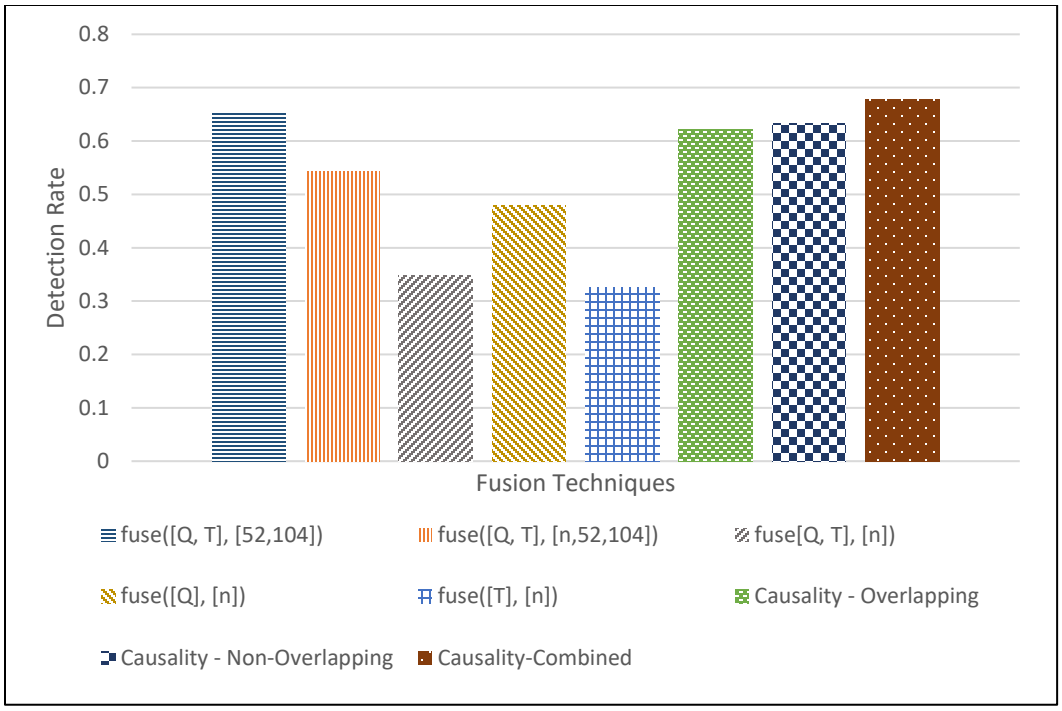

Figure 3.13: Detection rate comparison.

The detection rate for both overlapping and non-overlapping configurations performed well better than 4 of the fusion techniques proposed by Adjeroh et al. [8]. Also, we achieved a detection rate of 68 percent when we take into account the total detections from both overlapping as well as non-overlapping configurations which is more than the highest detection rate: 65 percent described in [8].

We represent the timeliness in terms of maximum, mean and median statistic of detection time before FDA action for overlapping as well as non-overlapping window configurations; and then compare against the prior work [8] (see Table 3.5).

Table 3.5: Timeliness comparison.

Fusion Technique	**Detection time before FDA action (in months)**		
	Mean	**Median**	**Maximum**
Causality-based (overlapping)	19.75	14.99	55.07
Causality-based (non-overlapping)	19.46	14.00	56.02
fuse([Q, T], [52,104]) [8]	23.58	23.5	36

For both the causality-based window configurations we get almost the same result in terms of detection time. Our causality-based fusion techniques had significantly good maximum

detection time in comparison with prior work. However, prior work's detection time for mean statistic was slightly better than our results, whereas the median statistic shows more difference in the timeliness of causality-based fusion techniques against the correlation-based fusion scheme described in [8].

As mentioned earlier, we used three different user-generated content (UGC) channels in this study for 107 blackbox FDA warnings. On the other hand, prior work used only Twitter and search query logs for the 46 FDA alerts drug dataset [8]. Thus, considering the dataset and the social media channels being used, our comparative analysis tells us that, the causality-based technique discussed in this work has relatively better performance.

3.4.3 Discussion

Clearly, we can see that Non-overlapping window configuration has slightly better detection rate when compared to overlapping window configuration (see Table 3.3). This could be due to the fact that the non-overlapping window covers a greater range of weeks for candidates in a v-structure allowing it to get a longer cross channel detection, thus having more potential candidates for drug signals. On the other hand, the overlapping window has a focused and shorter range of weeks for a given flag. This observation also answers the point that overlapping configuration has relatively greater precision and recall values than non-overlapping configuration, as the focused signals are the ones that have a high signal strength for a flag.

From Table 3.2, it is evident that no yellow flags are detected for overlapping configuration except for the setting: ($\tau_f = 3.0, \tau_p = 0.05$) with Rule(1,1), essentially forming the flags for shorter range and having more red flags than non-overlapping window configuration. Additionally, the maximum number of flags detected for any drug is far more than non-overlapping window configuration for all settings. This observation could indicate that overlapping configuration is better capable of capturing more number of closer hidden connections between channels, but one needs to be cautious with the false alarms. A false alarm is defined as the potential flag representing an ADE falsely, i.e. an alert signal which does not correspond to the FDA action for a drug. Differentiating false alarms and improving the social media alert signal for the drugs could be an interesting future aspect to the study.

In this work, we propose causality-based methodology to identify ADEs from the associations between social media UGC channels. Identifying the false alarms and detecting ADEs for unknown FDA blackbox warnings could be some of the directions for the future work.

Chapter 4

Evaluating Semantic Similarity for ADE Narratives

In this chapter, we discuss a new approach to evaluate semantic similarity measures in biomedical domain for comparing ADE narratives. Our objective is to evaluate all the possible similarity measurement algorithms (SMAs) listed in UMLS-Similarity program along with the vocabulary configurations (VCs) from UMLS Metathesaurus database to determine the best combination of measures and respective vocabulary configurations in computing semantic similarity in the domain of adverse drug event surveillance. This chapter is organized as follows: Section 4.1 introduces the materials and methods focusing on the dataset we used and the methodology involved in this work. Section 4.2 describes the experiment and results in detail including the experimental setup, results computed in each phase, and finally results showing evaluation of ADE narratives. The last Section 4.3 presents a brief discussion on our methodology in the light of the results we obtained.

4.1 Materials and Methods

Our methodology follows the procedure: 1) Identify the best vocabulary configurations (VCs) to use, 2) Determine the best combination of VCs and similarity measurement algorithms (SMAs) via joint optimization, and 3) Perform semantic similarity measurement using VC and SMA on narratives from social media against FDA narratives.

4.1.1 Datasets

Problem Domain Terms: In order to evaluate vocabulary configurations and similarity measures, we used anatomy and reaction terms extracted from social media channels for the 90 drugs described in Chapter 3. The dataset was formed after the extracted terms were processed for removing redundancy. The dataset had 105 initial anatomy terms and 202 initial reaction terms (referred as clusters), which was expanded with words with similar meanings, resulting in a new list with 178 anatomy terms, and 417 reaction terms. Refer Appendix B for the list of problem domain terms and clusters.

Human Ratings: Language is a major concern in evaluating the signals generated from social media, hence, the testing on SMAs and VCs should be based on the ratings obtained from general healthcare consumers along with healthcare professionals. Thus, we used human ratings as the standard to compare the performance of each combination of SMA and VC. Initially, we had 178 anatomy terms and 417 reaction terms, and forming pairs with all these terms would lead to over 100,000 pairs and that would have been impossible for the respondents to rate the similarity. Thus, we randomly selected 30 anatomy terms forming a set of 435 [(30*29)/2] anatomy pairs and 40 reaction pairs forming a set of 780 [(40*39)/2] reaction pairs. Further, to rate these 1215 pairs we contacted 6 computer science graduate researchers having appreciable knowledge of biomedical vocabulary usage over social media. Finally, based on their ratings a template with a set of 100 pairs was designed comprising 50 anatomy pairs and 50 reaction pairs. This template had rating options 0, 0.25, 0.5, 0.75 and 1 indicating levels from non-similar to very similar. We obtained 130 user ratings across the United States. This consists of 54 individuals coming from 5 different universities with health sciences and engineering background, and 76 from Amazon Mechanical Turk users having at least US Bachelor's degree. Further, we selected 117 ratings by excluding the outliers that had a negative correlation with the mean. We also analyzed the inter-rater agreement in terms of average correlation between raters. We filtered the ratings to achieve the benchmark of 80% average correlation and this resulted in a total of 107 ratings.

FDA BBW: To evaluate our work, we used FDA black box warning (BBW) labels as gold standard references and extracted ADE terms from the labels. We used FDA BBW data from January 2008 to April 2015 (http://www.fda.gov/safety/medwatch/safetyinformation/). This included 107 BBWs, on 90 drugs over the seven-year period.

4.1.2 Selection of Vocabulary Configurations (VCs)

Since the biomedical terms are found in multiple vocabularies it becomes a challenging question to decide which vocabulary to be used. The harder part is to find how good a given vocabulary is, in terms of covering all terms in a given problem domain.

Initial Selection: As stated earlier, UMLS has a huge collection of biomedical vocabularies which serves as a good resource for our work. However, we cannot use all the vocabularies in UMLS-Similarity due to performance and computational issues (see [44] for example). For our domain-specific social media extracted ADE terms, we followed the discussions in Park et. al [23], and selected vocabularies represented by source abbreviation (SAB): SNOMEDCT_US, CHV, MSH, LCH_NW, LNC, RXNORM, NCI_FDA, VANDF, and MTHSPL from UMLS [33]. We note that the work in [23] was based on terms extracted from social media using queries for terms related to

diabetes which is one of the most common groups of diseases and with a high degree of co-morbidity. For a more comprehensive treatment, we have considered some additional vocabularies where the content is closely related to ADE terms; namely, FMA, MDR, UWDA, WHO, NCI_NICHD, NCI_CTCAE, NDFRT_FDASPL, ICD10CM, MTHHH, and GS. Thus, given our specific problem domain of analyzing adverse drug events over social media channels, we had a total of 19 vocabularies to start our study as shown in Table 4.1.

Table 4.1: Selected vocabularies from UMLS.

Source Name	**Source Abbreviation (SAB)**
US Edition of SNOMEDCT	SNOMEDCT_US
Consumer Health Vocabulary	CHV
Medical Subject Headings	MSH
Library of Congress Subject Headings, Northwestern University subset	LCH_NW
Logical Observation Identifiers Names and Codes (LOINC)	LNC
RxNorm Vocabulary	RXNORM
U.S. Food and Drug Administration	NCI_FDA
Veterans Health Administration National Drug File	VANDF
Metathesaurus FDA Structured Product Labels	MTHSPL
Foundational Model of Anatomy Ontology	FMA
Medical Dictionary for Regulatory Activities Terminology (MedDRA)	MDR
University of Washington Digital Anatomist	UWDA
WHO Adverse Reaction Terminology	WHO
National Institute of Child Health and Human Development	NCI_NICHD
Common Terminology Criteria for Adverse Events	NCI_CTCAE
National Drug File – FDASPL	NDFRT_FDASPL
International Classification of Diseases, 10th Edition, Clinical Modification	ICD10CM
Metathesaurus HCPCS Hierarchical Terms	MTHHH
Gold Standard Drug Database	GS

Refining the VC selection: Now that we have the vocabularies chosen from UMLS, our next task is to reduce the list to get the best possible vocabularies based on the concepts defined in each VC, and the coverage of problem domain terms. To filter the vocabularies, we consider the following five features:

1. Total CUI's: Total number of concept unique identifiers (CUIs) listed for the vocabulary;
2. Terms Detected: number of problem domain terms detected in the vocabulary;
3. Concept coverage: number of concepts (CUI's) listed for problem domain terms;
4. Unique concepts: number of unique CUIs listed for each vocabulary; and
5. Clusters Detected: number of clusters which had at least one term detected as CUI.

For our purpose, good vocabularies are expected to have higher values for each of these features.

4.1.3 Similarity Measurement Algorithms (SMAs)

For automated evaluation of semantic similarity, the vocabulary is just one piece of the puzzle. Another key piece is the specific algorithm to be used to perform the similarity evaluation using the identified vocabulary. Thus, having narrowed down the vocabulary list as described above we now turn to the problem of selecting the SMAs. Interestingly, the match performance can also be influenced by the vocabulary used. Thus, the final choice of vocabulary cannot be made in isolation, but must consider the specific semantic similarity measurement algorithm being used. We used all the similarity measurement algorithms listed in UMLS-Similarity program except the *vector* measure which is meant to compute semantic relatedness (see Table 4.2). Each algorithm could have a range different for the similarity values, but for most, the range is from 0 to 1. However, a value of -1 would indicate there is no similarity between the pair of terms based on the vocabulary configuration. A similarity value could be -1 for two reasons: either one or both terms in a pair is (are) not found in the given configuration, or there is no path in the configuration connecting the term pairs.

Table 4.2: Similarity Measurement Algorithms in UMLS-Similarity.

S. No.	UMLS-Similarity Notation	Type	Reference
1	*lch*	path finding	Leacock and Chodorow(1998) [45]
2	*wup*	path finding	Wu and Palmer (1994) [20]
3	*zhong*	path finding	Zhong et al. (2002) [46]
4	*path*	path finding	path measure [18]
5	*upath*	path finding	Undirected path [18]
6	*cdist*	path finding	Rada et al. (1989) [19]
7	*nam*	path finding	Nguyen and Al-Mubaid (2006) [47]
8	*res*	IC-based	Resnik (1995) [48]
9	*lin*	IC-based	Lin (1988) [49]
10	*jcn*	IC-based	Jiang and Conrath (1997) [21]
11	*vector*	context vector	Pedersen et al. (2007) [31]
12	*pks*	path finding	Pekar and Staab (2002) [50]
13	*faith*	IC-based	Pirro and Euzenat (2010) [51]
14	*cmatch*	feature-based	Maedche and Staab (2001) [52]
15	*batet*	feature-based	Batet et al. (2011) [53]
16	*sanchez*	IC-based	Sánchez et al. (2012) [22]

4.1.4 Joint Selection of VC and SMA

We computed similarity values for the problem domain terms using each combination of selected VCs and the SMAs. To select the best SMA and VC, we compared their results with those from human observers. Comparison of the computed similarity values against the human ratings is performed in two steps: (1) using Pearson correlation against the mean rating from human observers, and (2) using information retrieval measures.

Correlation Analysis: For the mean representation of human ratings, we calculated the Pearson correlation coefficient against the corresponding computed similarity values. We used SciPy package in Python [54] to compute correlations. The syntax for correlation coefficient is given as:

$$r_{12} = \left[\sum (Y_{i1} - \overline{Y_1}) * (Y_{i2} - \overline{Y_2})\right] / \left[\sum (Y_{i1} - \overline{Y_1})^2 * \sum (Y_{i2} - \overline{Y_2})^2\right]^{\frac{1}{2}} \quad (2)$$

Correlation results helped us in reducing the number of combination of vocabulary configurations and similarity measurement algorithms. The combined results suggested favorable vocabularies as well as SMAs. We use these results for two key purposes: (1) Filtering the similarity measurement algorithms given all VCs; and (2) Analyzing the influence of SMAs on selection of vocabulary configurations.

Information Retrieval Factors: To further evaluate which combination of measurement algorithms and vocabulary configurations produces computed ratings that best mirror the human ratings, we grouped the problem domain term pairs into three classes: similar pairs, unknown pairs, and non-similar pairs. Let S(x,y) be the semantic similarity value between term pair (x, y), as returned by a given algorithm. We then used two thresholds τ_1 and τ_2 ($\tau_1 \geq \tau_2$) to classify a word pair (v_1, v_2) as follows:

$$Class\big(S(v_1, v_2)\big) = \begin{cases} similar, & S(v_1, v_2) > \tau_1 \\ unknown, & \tau_1 \geq S(v_1, v_2) \geq \tau_2 \\ not\ similar, & S(v_1, v_2) < \tau_2 \end{cases} \quad (3)$$

We used traditional information retrieval measures, namely, Precision (Pr), Recall (Rc), and F-measure (Fm) to evaluate the performance of selected combinations of vocabulary configurations and similarity measurement algorithms across the three classes. The formula to compute each of these factors for a given class is given as:

$$Precision\ (Pr) = \frac{N_C \cap N_H}{N_C} \quad (4)$$

$$Recall\ (Rc)\ =\ \frac{N_C\ \cap\ N_H}{N_H} \tag{5}$$

where,
N_C = Number of computed pairs in a given class
N_H = Number of human pairs in a given class

$$F-measure\ (Fm)\ =\ \frac{2 * Pr *\ Rc}{\mathrm{Pr}\ +\ Rc} \tag{6}$$

For the final selection of best combination of vocabulary configuration and similarity measurement algorithm for the given problem domain terms, we combine the information from the correlation analysis, and from the information retrieval measures.

4.2 Experiments and Results

4.2.1 Filtering Vocabularies

Using programs from the UMLS-Interface [18], we listed the Concept Unique Identifiers (CUIs) for vocabularies configured with combinations of relations (see Table 4.3). In Table 4.3, SAB refers to vocabularies and PAR (parent), CHD (child), RB (broader), and RN (narrower) are the relations defined in UMLS [17]. We observed that most vocabularies contain only PAR, CHD relations. While some have RB or RN as a primary way of representing hierarchy as seen for Medical Subject Headings (MSH). Interestingly, some vocabularies have concepts but are not connected by any relations. Thus, we chose to use relation categories: PAR, CHD; RB, RN; Similarity relations; and all relations. Similarity relations include all relations except XR (Not related), Empty relations and DEL (Deleted concept). For the complete list of all relations defined in UMLS, refer Table B.1 in Appendix B.

Using the UMLS-Interface package, we obtain all the concepts (CUIs) for the problem domain terms for each vocabulary configuration. Thus, we can evaluate the vocabularies based on various features discussed in Section 4.1.2. Figures 4.1 – 4.4 show some of the features used to filter the vocabularies.

Combination with CHV: Based on the results obtained (see Figure 4.1 – 4.4), we observed that the top 5 vocabularies for anatomy category are SNOMEDCT_US, CHV, LNC, MSH, and FMA. The top 5 vocabularies for reaction category are SNOMEDCT_US, CHV, MDR, MSH, and LNC. However, in Table 4.3, we see that CHV has only 2 CUIs for all the various types of relations specified. Clearly, this doesn't mean that CHV has only 2 concepts defined in it (see Figure 4.1). In fact, CHV has no relations defined between CUIs which restricts its use independently. On the

other hand, we see that there are other vocabularies where we have concepts obtained for different relation configurations like PAR/CHD, RB/RN, similar relations and all relations. Interestingly, it is noted that when we provide more number of relations, the UMLS-Similarity program raises an error and more relations would have a huge computational impact. Thus, we decided to include significant relations based on the number of concepts retrieved in Table 4.3 as shown in Table 4.4. The sources listed in Table 4.4 were used in combination with CHV as it has more coverage of terms and improves results as seen in previous work [11], [23], [30].

Table 4.3: Number of concepts in UMLS vocabularies using the SAB/REL configurations.

SAB/REL	PAR, CHD	RB, RN	Similar Relationships	All Relationships
CHV	2	2	2	2
FMA	97817	2	97830	97830
GS	2	2	2	2
ICD10CM	91673	2	101407	101407
LCH_NW	2	2	14578	14578
LNC	113526	24157	166393	166393
MDR	23439	2	53175	53175
MSH	28575	346054	366174	366174
MTHHH	7142	2	7142	7142
MTHSPL	2	2	50635	50635
NCI_CTCAE	2	2	2	2
NCI_FDA	2	2	2	2
NCI_NICHD	2	2	2	2
NDFRT_FDASPL	2	2	9137	9137
RxNorm	2	173552	202077	202077
SNOMEDCT_US	321004	43208	357226	357997
UWDA	61087	61087	61087	61087
VANDF	2	25072	31727	31727
WHO	1737	3176	3178	3178

Table 4.4: Relations used for selected source vocabularies.

Source vocabulary(SAB)	Relations(REL)
SNOMEDCT_US	PAR/CHD
MSH	RB/RN
LNC	PAR/CHD
MDR	PAR/CHD
FMA	PAR/CHD

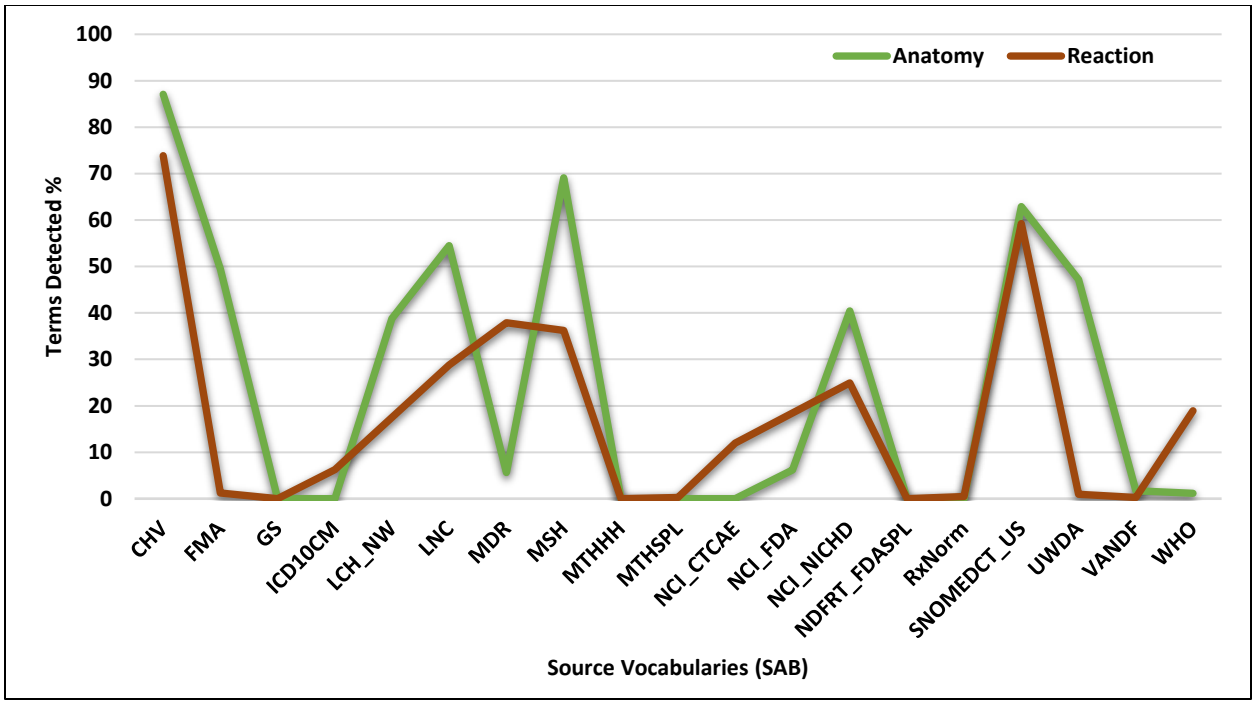

Figure 4.1: Percentage of terms detected.

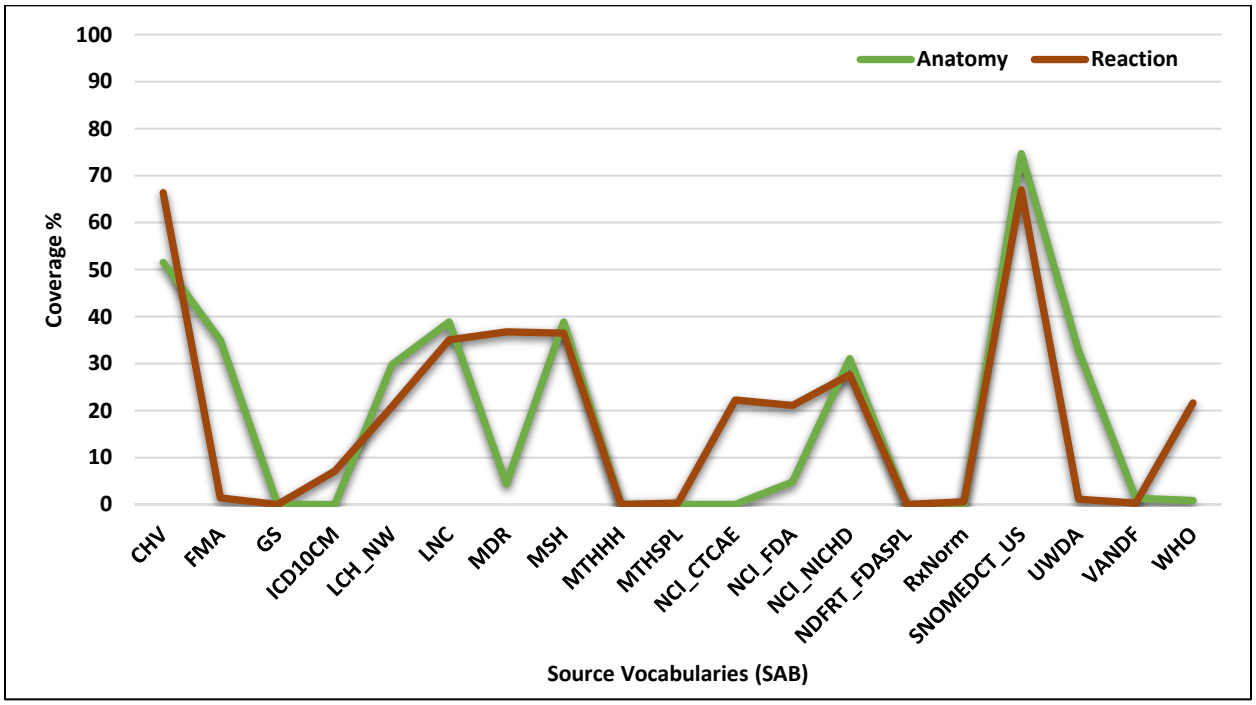

Figure 4.2: Percentage of concepts(CUIs) covered for terms.

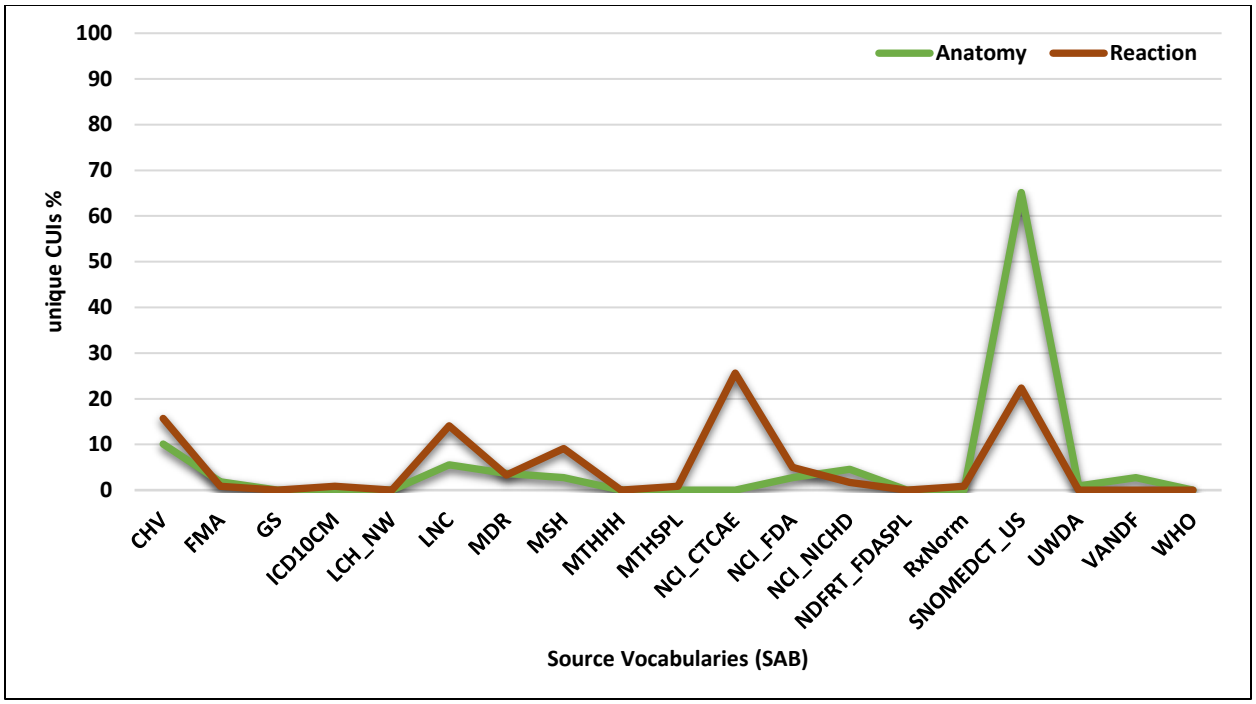

Figure 4.3: Percentage of unique concepts(CUIs) obtained.

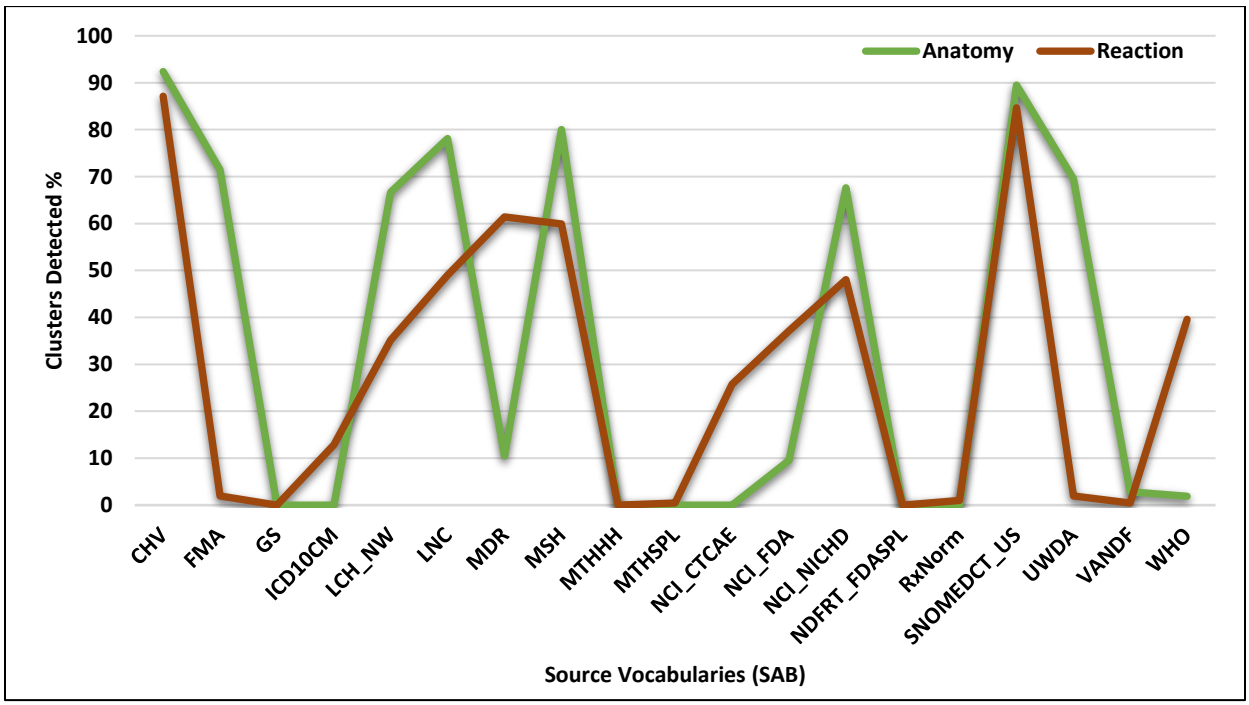

Figure 4.4: Percentage of clusters detected.

4.2.2 Joint Selection of VC and SMA

Correlation Analysis: If the significance level is ≤ 5% (i.e., P-value ≤ 0.05) and the corresponding correlation coefficient is positively high for any vocabulary configuration and similarity measurement algorithm, then we say that SMA or VC is favored. From Table 4.5, and Figures 4.5 and 4.6, we can see that for anatomy category the similarity measurement algorithms which frequently appear to be good are *cmatch*, *jcn* and *sanchez* with vocabulary configurations CHV-SNOMEDCT_US and CHV-LNC.

For reaction category, we did not get significant p-value to favor any of the algorithms. However, it has been observed that *nam* has very high correlation coefficient with vocabularies CHV-MDR and CHV-MSH, and undefined value for CHV-LNC. This behavior is because of the similarity values being -1.0 for most term pairs, resulting in less variability. Overall, the correlation analysis suggests that CHV-SNOMEDCT_US and CHV-MDR are the best VCs for working on reaction category terms (see Figure 4.6). Detailed results showing correlation coefficient and p-value for each SMA and selected VC are given in Appendix B, refer Table B.4 and Table B.5 for anatomy and reaction categories respectively.

Table 4.5: Outcomes of Pearson correlation.

Category	SMA favored	VC favored
Anatomy	*cmatch, jcn, sanchez*	CHV-SNOMEDCT_US, CHV-LNC
Reaction	*nam*	CHV-SNOMEDCT_US, CHV-MDR

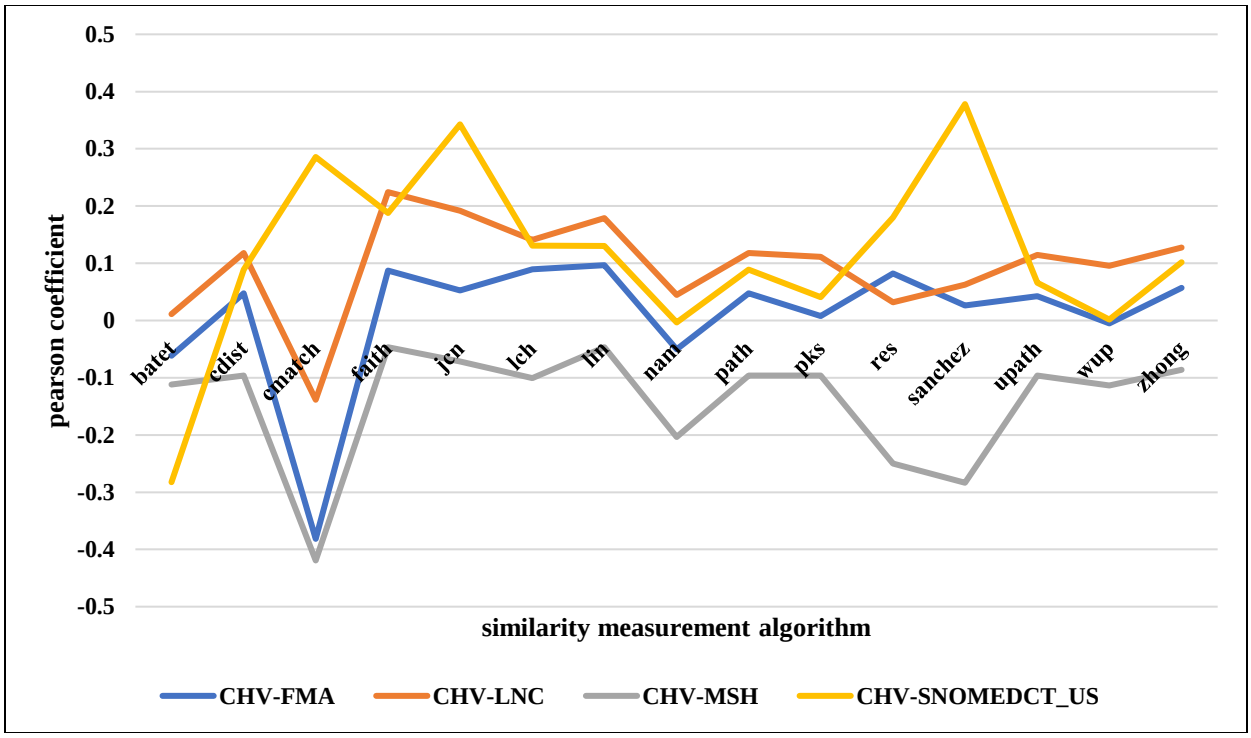

Figure 4.5: Correlation of computed similarity with human ratings – Anatomy pairs.

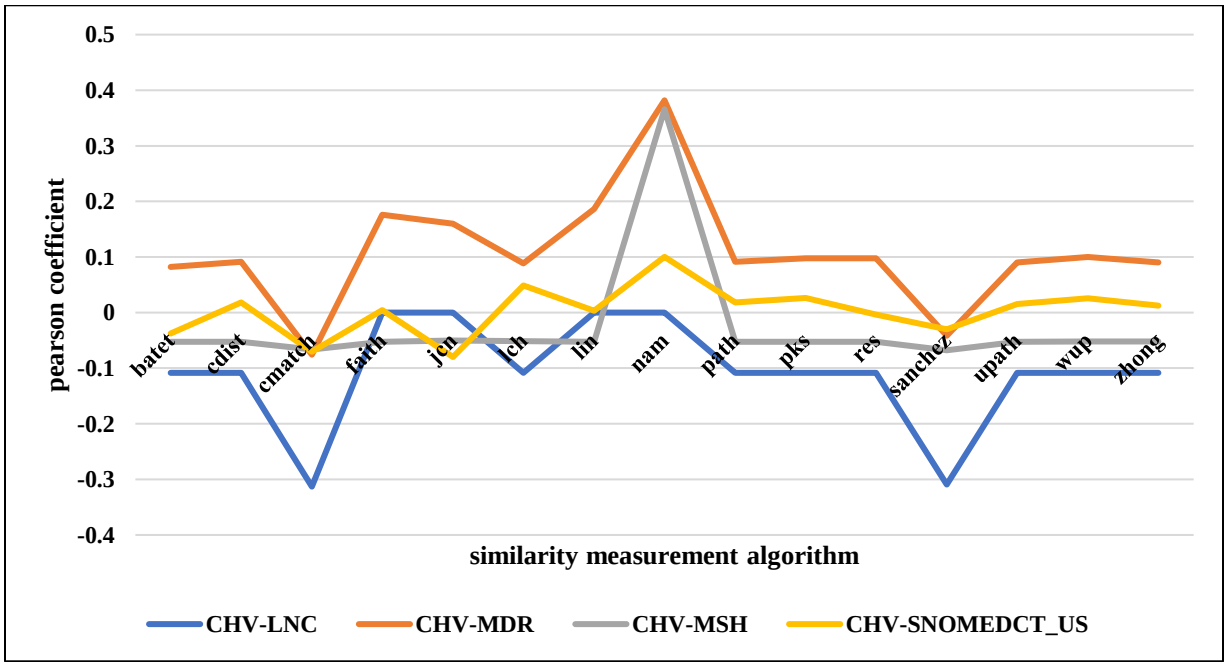

Figure 4.6: Correlation of computed similarity with human ratings – Reaction pairs.

Information Retrieval Factors: For the median of human ratings, we chose thresholds τ_1 as 0.75 and τ_2 as 0.3 to classify them into similar pairs, unknown pairs, and non-similar pairs. Similar to human ratings, for the SMA-VC obtained similarity values we chose τ_1 ranging from 0.5 to 0.95 and τ_2 ranging from 0.05 to 0.45 with a step size of 0.05. We selected the top 5 SMA-VCs based on F-measure against human rating statistic. For anatomy terms (Table 4.6), we found that the algorithms *jcn, faith, lin, cmatch* and *sanchez* with CHV-SNOMEDCT_US vocabulary are having high F-measure values with respect to human ratings. For reaction category (Table 4.7), the algorithms *wup, lin, pks, cmatch* with CHV-SNOMEDCT_US vocabulary configuration, and *res* with CHV-MDR vocabulary configuration performed well. Interestingly, we observe that *sanchez* has good F-measure for both CHV-SNOMEDCT_US and CHV-MDR.

Table 4.6 and 4.7 shows results only for similar pairs class. Detailed results for each class including similar pairs, non-similar pairs, unknown pairs are described in Table B.6 and Table B.7 given in Appendix B for anatomy and reaction categories respectively.

Table 4.6: Top 5 Similarity Algorithm/Vocabulary Configurations (Similar Pairs – Anatomy).

Measure	τ_1	τ_2	τ_{diff}	Configuration	Pr	Rc	Fm
jcn	0.8	0.5	0.3	CHV-SNOMEDCT_US	0.89	0.62	0.73
faith	0.7	0.5	0.2	CHV-SNOMEDCT_US	0.89	0.62	0.73
lin	0.8	0.45	0.35	CHV-SNOMEDCT_US	0.89	0.62	0.73
cmatch	0.5	0.45	0.05	CHV-SNOMEDCT_US	0.73	0.62	0.67
sanchez	0.8	0.5	0.3	CHV-SNOMEDCT_US	0.73	0.62	0.67

Table 4.7: Top 5 Similarity Algorithm/Vocabulary Configurations (Similar Pairs – Reaction).

Measure	τ_1	τ_2	τ_{diff}	Configuration	Pr	Rc	Fm
pks	0.55	0.35	0.2	CHV-SNOMEDCT_US	1	0.3	0.46
res	0.8	0.3	0.5	CHV-MDR	1	0.3	0.46
sanchez	0.5	0.4	0.1	CHV-MDR	1	0.3	0.46
wup	0.75	0.3	0.45	CHV-SNOMEDCT_US	1	0.3	0.46
sanchez	0.85	0.3	0.55	CHV-SNOMEDCT_US	0.75	0.3	0.43

4.2.3 Application to Evaluating ADE Surveillance Systems

Considering both the information retrieval factors the correlation analysis, our results suggest the following: for anatomy term pairs, we should use *jcn*, *cmatch*, or *sanchez* similarity measurement algorithm with CHV-SNOMEDCT_US vocabulary configuration. For reaction term pairs, we should use *sanchez*, *res*, or *wup* similarity measurement algorithm, with CHV-

SNOMEDCT_US or CHV-MDR vocabulary configuration. A key observation is the need for a combination of vocabularies (typically, CHV with some others), rather than one single vocabulary as has been used in prior work, such as [8]. Prior work also did not consider the impact of the similarity measurement algorithm on the results. We evaluated suggested ADE narratives from social media based on the method described in Chapter 3 for non-overlapping windows using the BBW data discussed in Section 4.1.1. We considered four cases (see Table 4.8): exact match i.e., not using semantic similarity; and the other 3 cases with similarity measure algorithm *sanchez* along with vocabulary configurations CHV, SNOMEDCT_US and combination of CHV-SNOMEDCT_US. Exact match is a string matching technique as used in Adjeroh et al. [8] where the authors used this methodology to compare the ADE narratives by expanding terms having similar meanings. The obtained results indicate that using semantic similarity has significantly greater improvement, especially, our suggested approach using a vocabulary configuration combining CHV-SNOMEDCT_US outperformed others.

Table 4.8: Evaluating social media ADE narratives for BBW data.

Approach	Anatomy			Reaction		
	Pr	Rc	Fm	Pr	Rc	Fm
exact match	0.048	0.176	0.076	0.022	0.140	0.038
CHV	0.048	0.176	0.076	0.024	0.141	0.041
SNOMEDCT_US	0.181	0.395	0.249	0.155	0.402	0.224
CHV-SNOMEDCT_US	0.197	0.452	0.275	0.175	0.465	0.255

4.3 Discussion

In our implementation, we chose UMLS-Similarity as it is built on UMLS which provides access to multiple vocabularies unlike other alternatives which require configuring vocabularies individually. In addition to this advantage, it has been observed in prior studies that using UMLS vocabularies would generate good results having higher agreement with human judgements [11], [23], [30].

The human ratings we used had a good representation of doctors, health professionals, health science students, engineering graduates and general graduate students. We even collected responses from Amazon's Mechanical Turk users having at least a US Bachelor's Degree [55]. Overall we achieved an interrater agreement of 80% average correlation for over a hundred human observer ratings. As the participants were familiar with social media as a significant source of healthcare information and considering the interrater agreement, we believe our dataset best fits the testing.

We followed a-step-by-step approach testing all the vocabulary configurations and similarity measure algorithms exhaustively, to get the best suitable VC and SMA combination for the adverse drug event terms. Our results showed that the configuration of CHV-SNOMEDCT_US is the best for anatomy terms using the IC-based similarity measure algorithms *sanchez* and *jcn*. It is also observed that CHV-MDR and CHV-SNOMEDCT_US configurations work well for reaction category terms with *sanchez* similarity measure algorithm. However, our results also indicate that using biomedical ontologies and the similarity measures is not sufficient for reaction category terms. The major reason is that reaction terms are more general and are not as specific when compared to anatomy category terms. Thus, we believe that using general English vocabularies such as WordNet [56] along with UMLS would improve the semantic similarity for reaction category terms.

Our findings also show that the vocabulary MedDRA -- Medical Dictionary for Regulatory Activities (abbreviated as MDR in UMLS) has a good representation of reaction category problem domain terms. This can be considered in the light of the fact that SIDER, a well-known dataset for representing side effects uses MedDRA to generate side effect names [57].

Chapter 5

Conclusion and Future Work

We study the problem of Adverse Drug Events and the postmarketing drug surveillance involved to detect such harmful events. The study included examining prior works in adverse drug events (ADE) detection using social media as a prime resource. Primarily our objective was focused on fusing social media UGC channels for ADE detection, as the signal fusion technique had seen to be generating promising results in terms of early detection of ADE. During this study we introduced a novel approach of using graphical causal model for social media signal fusion. Using the proposed Causality-based technique, we were able to investigate ADE detection on 90 drugs having a total of 107 FDA black box warnings. Further, we presented a methodology to evaluate precision and recall of detected ADE narratives against the gold standard FDA using semantic similarity algorithms published in biomedical domain.

We experimented different similarity measure algorithms designed for biomedical ontologies. We showed that choosing a measure alone is not enough for computing semantic similarity for terms in a problem domain. Likewise, having known of a vocabulary which is related to a particular problem domain does not solve the problem of computing semantic similarity for the terms in that problem domain. We defined a way of choosing the vocabulary first, we also showed that combining a vocabulary with CHV improves the concept coverage and thereby covering more terms from the problem domain and later we experimented each configuration of vocabulary with different measures. The results shown in this work are based on the existing measures published in UMLS-Similarity program version 1.47 and the source vocabularies extracted from UMLS version 2017AA. For future releases of UMLS and the UMLS-Similarity program the methodology we developed can still be used to find the best measure and vocabulary configuration combination for a given problem domain terms.

Unlike most of the prior studies which focused only on ADE detection or some of them just representing the recall of detected ADE narrative, we evaluated the detected ADEs in terms of timeliness, recall and precision. Our results had a good detection rate, precision and recall considering the dataset we have used representing over 100 FDA black box warnings. In future we would like to further examine causality on fusing additional social media channels including search query logs. Identifying the false alarms and detecting ADEs for unknown FDA blackbox warnings could be some of the prospective studies. Another direction for future work could be to implement semantic similarity algorithms in capturing signals from social media channels. Also utilizing general English vocabularies like WordNet [56] in addition to UMLS could be one more interesting aspect to consider.

References

[1] "Guideline for Industry Clinical Safety Data Management: Definitions and Standards for Expedited Reporting," 1995. [Online]. Available: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073087.pdf. [Accessed: 19-Mar-2018].

[2] Y. Ji, H. Ying, P. Dews, M. S. Farber, *et al.*, "A fuzzy recognition-primed decision model-based causal association mining algorithm for detecting adverse drug reactions in postmarketing surveillance," in *IEEE World Congress on Computational Intelligence, WCCI*, 2010.

[3] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, *et al.*, "Utilizing social media data for pharmacovigilance: A review," *J. Biomed. Inform.*, vol. 54, pp. 202–212, 2015.

[4] K. Wester, A. K. Jönsson, O. Spigset, H. Druid, *et al.*, "Incidence of fatal adverse drug reactions: A population based study," *Br. J. Clin. Pharmacol.*, vol. 65, no. 4, pp. 573–579, 2008.

[5] "MedWatch Voluntary Report." [Online]. Available: https://www.accessdata.fda.gov/scripts/medwatch/index.cfm?action=reporting.home. [Accessed: 04-Apr-2018].

[6] "Yellow Card Scheme - MHRA." [Online]. Available: https://yellowcard.mhra.gov.uk/. [Accessed: 04-Apr-2018].

[7] "UMC | VigiBase." [Online]. Available: https://www.who-umc.org/vigibase/vigibase/. [Accessed: 04-Apr-2018].

[8] D. Adjeroh, R. Beal, A. Abbasi, W. Zheng, *et al.*, "Signal Fusion for Social Media Analysis of Adverse Drug Events," *IEEE Intell. Syst.*, vol. 29, no. 2, pp. 74–80, 2014.

[9] A. Abbasi, D. Adjeroh, M. Dredze, M. J. Paul, *et al.*, "Social media analytics for smart health," *IEEE Intell. Syst.*, vol. 29, no. 2, pp. 60–80, 2014.

[10] A. Abbasi, T. Fu, D. Zeng, and D. Adjeroh, "Crawling credible online medical sentiments for social intelligence," *Soc. Comput. Int. Conf. IEEE*, pp. 254–263, 2013.

[11] C. C. Yang, H. Yang, and L. Jiang, "Postmarketing Drug Safety Surveillance Using Publicly Available Health-Consumer-Contributed Content in Social Media," *ACM Trans. Manag. Inf. Syst.*, vol. 5, no. 1, pp. 1–21, 2014.

[12] R. B. Correia, L. Li, and L. M. Rocha, "Monitoring potential drug interactions and reactions via network analysis of Instagram user timelines.," *Biocomput. Proc. Pacific Symp. (pp. 492-503).*, vol. 21, pp. 492–503, 2016.

[13] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, *et al.*, "Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *J. Am. Med. Informatics Assoc.*, vol. 22, no. 3, pp. 671–681, 2015.

[14] A. Abbasi, J. Li, S. Abbasi, D. Adjeroh, *et al.*, "Don't Mention It? Analyzing User-Generated Content Signals for Early Adverse Drug Event Warnings," *Proceedings, Wkshp. Inf. Technol. Syst. (WITS), Dallas, TX.*, pp. 1–16, 2015.

[15] D. H. Kwon and D. A. Bessler, "Graphical methods, inductive causal inference, and

econometrics: a literature review," *Computational Economics*, vol. 38, no. 1. pp. 85–106, 2011.
[16] Q. T. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," *Journal of the American Medical Informatics Association*, vol. 13, no. 1. pp. 24–29, 2006.
[17] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. 90001, pp. D267–D270, 2004.
[18] B. T. McInnes, T. Pedersen, and S. V. S. Pakhomov, "UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity.," *AMIA Annu. Symp. Proc.*, pp. 431–5, 2009.
[19] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 1, pp. 17–30, 1989.
[20] Z. Wu and M. Palmer, "Verbs semantics and lexical selection.," *Proc. 32nd Annu. Meet. Assoc. Comput. Linguist. -*, pp. 133–138, 1994.
[21] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *Proc. Int. Conf. Res. Comput. Linguist. Taiwan*, 1997.
[22] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7718–7728, 2012.
[23] M. S. Park, Z. He, Z. Chen, S. Oh, *et al.*, "Consumers' Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites.," *JMIR Med. Informatics*, vol. 4, no. 4, p. e41, Nov. 2016.
[24] R. J. W. Stephanie N. Schatz, "Adverse Drug Reactions," *Pharmacother. self Assess. Progr.*, 2015.
[25] M. Hashiguchi, S. Imai, K. Uehara, J. Maruyama, *et al.*, "Factors Affecting the Timing of Signal Detection of Adverse Drug Reactions," *PLoS One*, vol. 10, no. 12, p. e0144263, Dec. 2015.
[26] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, *et al.*, "Text and data mining techniques in adverse drug reaction detection," *ACM Comput. Surv.*, vol. 47, no. 4, pp. 1–39, 2015.
[27] G. Sugihara, R. May, H. Ye, C. Hsieh, *et al.*, "Detecting causality in complex ecosystems.," *Science*, vol. 338, no. 6106, pp. 496–500, 2012.
[28] R. Engle and C. W. J. Granger, "Co-integration and error-correction representation, estimation and testing," *Econometrica*, vol. 55, no. 2, pp. 251–276, 1987.
[29] R. Sangüesa and U. Cortés, "Learning causal networks from data: a survey and a new algorithm for recovering possibilistic causal networks *," *AI Commun.*, vol. 10, no. 1, pp. 31–61, 1997.
[30] V. N. Garla and C. Brandt, "Semantic similarity in the biomedical domain: An evaluation across knowledge sources," *BMC Bioinformatics*, vol. 13, no. 1, 2012.
[31] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *J. Biomed. Inform.*, vol. 40, no. 3, pp. 288–299, 2007.
[32] D. Sánchez and M. Batet, "Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective," *J. Biomed. Inform.*, vol. 44, no. 5, pp. 749–759, 2011.
[33] National Library of Medicine (US), *UMLS® Reference Manual*. National Library of Medicine (US), 2009.

[34] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, *et al.*, "Semantic similarity in biomedical ontologies," *PLoS Comput. Biol.*, vol. 5, no. 7, p. e1000443, 2009.
[35] K. Gimpel, N. Schneider, B. O'Connor, D. Das, *et al.*, "Part-of-Speech tagging for Twitter: Annotation, Features, and Experiments," *Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Shortpapers*, no. 2, pp. 42–47, 2011.
[36] J. Pearl, "Causal inference in statistics: An overview," *Stat. Surv.*, vol. 3, no. 0, pp. 96–146, 2009.
[37] W. Buntine, "A guide to the literature on learning probabilistic networks from data," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 2, pp. 195–210, 1996.
[38] C. N. Glymour and G. F. Cooper, *Computation, causation, and discovery*. Aaai Press, 1999.
[39] C. W. J. Granger, "Some properties of time series data and their use in econometric model specification," *J. Econom.*, vol. 16, no. 1, pp. 121–130, 1981.
[40] C. W. J. Granger and A. A. Weiss, "Time series analysis of error-correction models," in *Studies in Econometrics, Time Series, and Multivariate Statistics*, Elsevier, 1983, pp. 255–278.
[41] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *J. Am. Stat. Assoc.*, vol. 77, no. 378, pp. 304–313, 1982.
[42] J. F. Geweke, "Measures of conditional linear dependence and feedback between time series," *J. Am. Stat. Assoc.*, vol. 79, no. 388, pp. 907–915, 1984.
[43] S. Seabold and J. Perktold, "Statsmodels: econometric and statistical modeling with Python," in *9th Python in Science Conference*, 2010, pp. 57–61.
[44] "Re: [umls-similarity] Practical large coverage configuration." [Online]. Available: https://www.mail-archive.com/umls-similarity@yahoogroups.com/msg00334.html. [Accessed: 15-Mar-2018].
[45] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification.," *WordNet An Electron. Lex. database.*, pp. 265–283, 1998.
[46] J. Zhong, H. Zhu, J. Li, and Y. Yu, "Conceptual graph matching for semantic search.," *Int. Conf. Concept. Struct.* , no. Springer, Berlin, Heidelberg., pp. 92–106, 2002.
[47] H. Al-Mubaid and H. A. Nguyen, "A cluster-based approach for semantic similarity in the biomedical domain," in *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 2006, pp. 2713–2717.
[48] P. Resnik, "Using information content to evaluate seantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.
[49] D. Lin, "An information-theoretic definition of similarity," *Proc. ICML*, pp. 296–304, 1998.
[50] V. Pekar and S. Staab, "Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision," *Proc. 19th Int. Conf. Comput. Linguist. - Vol. 1*, pp. 1–7, 2002.
[51] G. Pirró and J. Euzenat, "A feature and information theoretic framework for semantic similarity and relatedness," in *International Semantic Web Conference*, 2010, pp. 615–630.
[52] A. Maedche and S. Staab, "Comparing ontologies-similarity measures and a comparison study," 2001.
[53] M. Batet, D. Sánchez, and A. Valls, "An ontology-based measure to compute semantic

similarity in biomedicine," *J. Biomed. Inform.*, vol. 44, no. 1, pp. 118–125, 2011.
[54] T. E. Oliphant, "SciPy: Open source scientific tools for Python," *Comput. Sci. Eng.*, vol. 9, pp. 10–20, 2007.
[55] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data?," *Perspect. Psychol. Sci.*, vol. 6, no. 1, pp. 3–5, 2011.
[56] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[57] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, 2016.

Appendix A

Causality Based Signal Fusion

Table A.1: List of 90 drugs used in Causality Based Signal Fusion.

ABLAVAR	EXJADE	AVELOX	LETAIRIS	RITUXAN
FIORICET	ONTAK	PERFOROMIST	COZAAR	SEREVENT
ACTIQ	MULTAQ	MULTIHANCE	MAGNEVIST	SIMPONI
TEKTURNA	RANEXA	OMNISCAN	CELLCEPT	RAPAMUNE
ALTACE	DEPAKENE	PROHANCE	MYFORTIC	OSMOPREP
ARZERRA	HALCION	OPTIMARK	VIRAMUNE	VISICOL
ATACAND	ADVAIR	EOVIST	NIZORAL	SUTENT
AVANDAMET	OCTAGAM	HUMIRA	ORTHO EVRA	TASIGNA
AVANDIA	ZORTRESS	PRINZIDE	MITOXANTRONE HCL	TASINGA
IMURAN	ELAPRASE	HYZAAR	NOVANTRONE	INCIVEK
REGRANEX	RAPTIVA	IDURSULFASE	ZYPREXA	ANDROGEL
BROVANA	TRUVADA	INCLUSIG	ORTHO NOVUM	FARESTON
SYMBICORT	ENBREL	INFED	ACEON	TYGACIL
ZYBAN	ESTRADERM	SOPRANOX	PROMACTA	TYSABRI
APLENZIN	POTIGA	SPORANOX	PROPYLTHIOURACIL	ULTRACET
CIMZIA	FACTIVE	EPZICOM	QUALAQUIN	PROMETRIUM
CLEOCIN	FLOXIN	TRIZIVIR	REGLAN	STAVZOR
DANTRIUM	FLUDARA	ARAVA	REMICADE	VIREAD

Appendix B

Evaluation of Semantic Similarity for ADE Narratives

Table B.1: Relationships defined in UMLS.

REL (Relationship)	Description
AQ	Allowed qualifier
CHD	has child relationship in a Metathesaurus source vocabulary
DEL	Deleted concept
PAR	has parent relationship in a Metathesaurus source vocabulary
QB	can be qualified by.
RB	has a broader relationship
RL	the relationship is similar or "alike". the two concepts are similar or "alike". In the current edition of the Metathesaurus, most relationships with this attribute are mappings provided by a source, named in SAB and SL; hence concepts linked by this relationship may be synonymous, i.e. self-referential: CUI1 = CUI2. In previous releases, some MeSH Supplementary Concept relationships were represented in this way.
RN	has a narrower relationship
RO	has relationship other than synonymous, narrower, or broader
RQ	related and possibly synonymous.
RU	Related, unspecified
SIB	has sibling relationship in a Metathesaurus source vocabulary.
SY	source asserted synonymy.
XR	Not related, no mapping
	Empty relationship

Problem Domain Terms used in Evaluation of Semantic Similarity

This appendix lists all the biomedical terms used in this research. It is organized as follows:

1. Clusters - The problem domain terms for this research are represented in terms of clusters having one or more terms for each cluster. The clusters are organized into:
    a. Anatomy Clusters
    b. Reaction Clusters

2. Terms – The multiple terms representing each cluster are expanded to get total terms in each category:
    a. Anatomy Terms
    b. Reaction Terms

1.a Anatomy Clusters

We have 105 clusters in anatomy category listed as follows:

1. {abdomen}
2. {achilles}
3. {anus, anal}
4. {appendix}
5. {arm, arms}
6. {artery, arteries, arterial}
7. {back}
8. {bladder}
9. {blood}
10. {bone marrow}
11. {bone, bones}
12. {brain}
13. {breast, breasts, boob, boobs}
14. {buttocks, butt, ass}
15. {canal}
16. {cardiovascular, cardio}
17. {cervex, cervical}
18. {cheek, cheeks, cheekbones}
19. {chest}
20. {child, children, childrens, children's}
21. {chin}
22. {clavical}
23. {cognitive, cognition}

24. {colon}
25. {ear, ears, earlobe, earlobes}
26. {elbow, elbows}
27. {erectile}
28. {eye, eyes}
29. {face}
30. {female, females}
31. {foot, feet}
32. {forearm, forearms}
33. {forehead}
34. {gastric}
35. {genital, genitals}
36. {gland, glands}
37. {hair}
38. {hand, palm}
39. {head}
40. {heart, heartbeat}
41. {heel}
42. {hip, hips}
43. {hive, hives}
44. {immune system}
45. {impair, impaired}
46. {infant, infants}
47. {intestinal, intestine, intestines}
48. {joint, joints}
49. {kidney}
50. {knee, knees}
51. {leg, legs}
52. {ligament, ligaments}
53. {lip, lips}
54. {liver}
55. {lobe, lobes}
56. {lumbar}
57. {lung, lungs}
58. {lymph node, lymph nodes, lymph gland, lymph glands}
59. {lymph}
60. {macular}
61. {male, males}
62. {man, men}
63. {mental}
64. {mouth}
65. {muscle, muscles, muscular}
66. {nail, nails}
67. {neck}

68. {nerve, nerves}
69. {newborn, newborns}
70. {nipple, nipples}
71. {nose}
72. {ovarian, ovary, ovaries}
73. {pancreas}
74. {pectoral}
75. {pelvis}
76. {peptic}
77. {plasma cell, plasma cells}
78. {pregnant, pregnancy}
79. {pulmonary}
80. {pulse}
81. {rectum, rectal}
82. {respiratory}
83. {retina, retinal}
84. {rheumatic}
85. {shoulder, shoulders}
86. {sinus}
87. {skin}
88. {spine, spinal cord}
89. {spleen}
90. {sternum}
91. {stomach}
92. {tendon}
93. {testicle, testicular, testes}
94. {thigh, thighs}
95. {thoracic}
96. {throat}
97. {tongue}
98. {tonsil, tonsils}
99. {tooth, teeth}
100. {urinary}
101. {vagina, vaginal}
102. {vein, venous, veins}
103. {white blood cell, white blood cells}
104. {women, woman}
105. {wrist, wrists}

1.b Reaction Clusters

We have 202 clusters in reaction category listed as follows:

1. {abnormality, abnormalities, abnormal}
2. {ache, aching, aches, ached}
3. {acne}
4. {acute}
5. {addiction, addictive}
6. {adverse}
7. {aggression, aggressive}
8. {agitate, agitated, agitates, agitation}
9. {akathisia}
10. {allergic, allergy, allergen}
11. {amnesia}
12. {anemia}
13. {angina}
14. {anorexia, anorexic}
15. {anxiety, anxious}
16. {appendicitis, appendectomy}
17. {arrhythmia}
18. {asthenia}
19. {atrocious}
20. {attack}
21. {awful}
22. {bad}
23. {benign}
24. {bleed, bleeding, bleedings, bleeds, blood, bloody}
25. {blind, blindness}
26. {blister, blisters}
27. {blur, blurred, blurry, blurs}
28. {bradycardia}
29. {breakdown}
30. {breath, breathe, breathing}
31. {burn, burning, burns, burned}
32. {cancer, cancerous}
33. {cause, causes, caused}
34. {chill, chills}
35. {chronic}
36. {clot, clots, clotting}
37. {colitis}
38. {confusion}
39. {constipation}
40. {convulsion, convulsions}

41. {cramp, cramps, cramping}
42. {crohns disease, crohn syndrome, regional enteritis}
43. {crystalization, crystal, crystals}
44. {damage, damaged, damages}
45. {danger, dangers, dangerous, dangerously}
46. {deaf, deafness}
47. {death, dead, died}
48. {decrease, decreasing, decreased}
49. {depressed, depression}
50. {destruction, destroy, destroys, destroyed}
51. {diabetes, diabetic}
52. {diarrhea}
53. {difficult}
54. {dire}
55. {disorder}
56. {diverticulitis}
57. {diverticulosis}
58. {dizziness, dizzy}
59. {drowsiness, drowsy}
60. {dysfunction}
61. {dyskinesia}
62. {dyspepsia}
63. {dyspnea}
64. {eczema}
65. {edema}
66. {effect, effects}
67. {epilepsy}
68. {excess, excessive, overly}
69. {exhaustion, exhausted, exhausting}
70. {explode, exploding, explosive, explosion}
71. {failure, failures}
72. {faint, fainting}
73. {fatigue, fatigued, fatiguing, fatigues}
74. {fetal circulation}
75. {fever, fevers}
76. {flush, flushed, flushes, flushing}
77. {fracture, fractures}
78. {gas, gaseous, gassy, gastritis}
79. {hallucinating, hallucinations}
80. {headache, headaches}
81. {heartburn}
82. {hepatitis c, hcv}
83. {hive, hives}
84. {horrible, horrific, horrifying}

85. {hostile}
86. {human immunodeficiency virus, hiv, acquired immunodeficiency syndrome, aids}
87. {hurt, hurts, hurting}
88. {hyperactive}
89. {hyperglycemia}
90. {hyperkalemia}
91. {hypertension}
92. {hypoglycemia}
93. {hypotension}
94. {ill, illness}
95. {impair, impaired, impairs, impairing}
96. {impotence, impotent}
97. {impulsive}
98. {inability, unable}
99. {increase, increased, increasing}
100. {indigestion}
101. {infect, infected, infection, infects, infections}
102. {inflamation, inflamed, inflame}
103. {injure, injury, injuries, injured}
104. {insomnia}
105. {interaction, interactions}
106. {interval, intervals}
107. {irreparable}
108. {irreversible}
109. {irritable, irritate, irritated, irritability, irritates}
110. {itch, itching, itchy, itches}
111. {jaundice}
112. {ketoacidosis}
113. {leukemia}
114. {leukoencephalopathy, leukodystrophy}
115. {loss, losses}
116. {lymphoma}
117. {malignancies, malignancy, malignant}
118. {melanoma}
119. {mellitus}
120. {miscarriage}
121. {mortified}
122. {murmur, murmurs}
123. {myopathy, myopathic}
124. {nausea, nauseous}
125. {nervousness}
126. {neuropathy}
127. {numb, numbness, numbing}
128. {obstructive sleep apnea, osa}

129. {pain, painful, pains}
130. {palpitations}
131. {pancreatitis}
132. {panicky}
133. {paresthesia}
134. {parkinsonism}
135. {persistent}
136. {pneumonia}
137. {pounding}
138. {pressure}
139. {priapism}
140. {problem, problems}
141. {progressive multifocal leukoencephalopathy, pml}
142. {psychiatric, psychotic, psychosis, psycho}
143. {pulmonary arterial hypertension, pah, pulmonary hypertension}
144. {pulsate, pulsating}
145. {rapid, rapidly}
146. {rash, rashes}
147. {react, reaction, reactions}
148. {reduce, reducing, reduced, reduction, reductions}
149. {regulatory, regulation, regulate}
150. {rhythm, rhythms}
151. {ringing}
152. {runny}
153. {rupture, ruptures, ruptured, rupturing}
154. {sad, sadness}
155. {sclerosis}
156. {seizure, seizures, seizing}
157. {sensation}
158. {sensitivity, sensitive}
159. {serious, seriousness, seriously}
160. {serotonin syndrome, serotonin toxicity, serotonin sickness, serotonin poisoning}
161. {severe, severely}
162. {shakiness, shaky}
163. {sharp}
164. {short, shortness, shortening, shorter}
165. {sleep, sleeping, sleepiness, slept}
166. {sore, soreness}
167. {spasm, spasms}
168. {stroke, strokes}
169. {stuffy, stuffiness, congest, congested, congestion}
170. {sudden}
171. {sugar, glycosylate, glycosylated, hemoglobin, hba1c}
172. {suicidal, suicide}

173. {sweat, sweats, sweating}
174. {swell, swelling, swollen, swells}
175. {syncope}
176. {tachycardia}
177. {temper, tempers}
178. {tenderness}
179. {tendonitis}
180. {terrible}
181. {terrified, terrifying}
182. {thrombosis, thromboembolism}
183. {tingle, tingling}
184. {tinnitus}
185. {tired, tiredness, tire}
186. {torsades de pointes, ventricular tachycardia}
187. {toxicity}
188. {trauma, traumatic}
189. {tremor, tremors}
190. {tumor, tumors, tumorous}
191. {ulcer, ulcers, ulcerative}
192. {unexplained}
193. {upset, irritated, irritable, irritate, irritation, upsetting}
194. {vaginitis}
195. {vertigo}
196. {virus, viral}
197. {vomit, vomits, vomiting, vomited}
198. {wart, warts}
199. {watery}
200. {weak, weaken, weakening, weakness, weaknesses}
201. {weight}
202. {worse, worsen, worsening}

2.a Anatomy Terms

For the 105 anatomy clusters we have 178 anatomy terms as shown in Table B.2.

Table B.2: Anatomy terms.

abdomen	achilles	anus
anal	appendix	arm
arms	artery	arteries
arterial	back	bladder
blood	bone marrow	bone
bones	brain	breast
breasts	boob	boobs
buttocks	butt	ass
canal	cardiovascular	cardio
cervex	cervical	cheek
cheeks	cheekbones	chest
child	children	childrens
children's	chin	clavical
cognitive	cognition	colon
ear	ears	earlobe
earlobes	elbow	elbows
erectile	eye	eyes
face	female	females
foot	feet	forearm
forearms	forehead	gastric
genital	genitals	gland
glands	hair	hand
palm	head	heart
heartbeat	heel	hip
hips	hive	hives
immune system	impair	impaired
infant	infants	intestinal
intestine	intestines	joint
joints	kidney	knee
knees	leg	legs
ligament	ligaments	lip
lips	liver	lobe
lobes	lumbar	lung
lungs	lymph node	lymph nodes
lymph gland	lymph glands	lymph
macular	male	males

man	men	mental
mouth	muscle	muscles
muscular	nail	nails
neck	nerve	nerves
newborn	newborns	nipple
nipples	nose	ovarian
ovary	ovaries	pancreas
pectoral	pelvis	peptic
plasma cell	plasma cells	pregnant
pregnancy	pulmonary	pulse
rectum	rectal	respiratory
retina	retinal	rheumatic
shoulder	shoulders	sinus
skin	spine	spinal cord
spleen	sternum	stomach
tendon	testicle	testicular
testes	thigh	thighs
thoracic	throat	tongue
tonsil	tonsils	tooth
teeth	urinary	vagina
vaginal	vein	venous
veins	white blood cell	white blood cells
women	woman	wrist
wrists		

2.b Reaction Terms

For the 202 reaction clusters we have 417 reaction terms as shown in Table B.3.

Table B.3: Reaction terms.

abnormality	abnormalities	abnormal
ache	aching	aches
ached	acne	acute
addiction	addictive	adverse
aggression	aggressive	agitate
agitated	agitates	agitation
akathisia	allergic	allergy
allergen	amnesia	anemia
angina	anorexia	anorexic
anxiety	anxious	appendicitis
appendectomy	arrhythmia	asthenia
atrocious	attack	awful
bad	benign	bleed
bleeding	bleedings	bleeds
blood	bloody	blind
blindness	blister	blisters
blur	blurred	blurry
blurs	bradycardia	breakdown
breath	breathe	breathing
burn	burning	burns
burned	cancer	cancerous
cause	causes	caused
chill	chills	chronic
clot	clots	clotting
colitis	confusion	constipation
convulsion	convulsions	cramp
cramps	cramping	crohns disease
crohn syndrome	regional enteritis	crystalization
crystal	crystals	damage
damaged	damages	danger
dangers	dangerous	dangerously
deaf	deafness	death
dead	died	decrease
decreasing	decreased	depressed
depression	destruction	destroy
destroys	destroyed	diabetes

diabetic	diarrhea	difficult
dire	disorder	diverticulitis
diverticulosis	dizziness	dizzy
drowsiness	drowsy	dysfunction
dyskinesia	dyspepsia	dyspnea
eczema	edema	effect
effects	epilepsy	excess
excessive	overly	exhaustion
exhausted	exhausting	explode
exploding	explosive	explosion
failure	failures	faint
fainting	fatigue	fatigued
fatiguing	fatigues	fetal circulation
fever	fevers	flush
flushed	flushes	flushing
fracture	fractures	gas
gaseous	gassy	gastritis
hallucinating	hallucinations	headache
headaches	heartburn	hepatitis c
hcv	hive	hives
horrible	horrific	horrifying
hostile	human immunodeficiency virus	hiv
acquired immunodeficiency syndrome	aids	hurt
hurts	hurting	hyperactive
hyperglycemia	hyperkalemia	hypertension
hypoglycemia	hypotension	ill
illness	impair	impaired
impairs	impairing	impotence
impotent	impulsive	inability
unable	increase	increased
increasing	indigestion	infect
infected	infection	infects
infections	inflamation	inflamed
inflame	injure	injury
injuries	injured	insomnia
interaction	interactions	interval
intervals	irreparable	irreversible
irritable	irritate	irritated
irritability	irritates	itch
itching	itchy	itches

jaundice	ketoacidosis	leukemia
leukoencephalopathy	leukodystrophy	loss
losses	lymphoma	malignancies
malignancy	malignant	melanoma
mellitus	miscarriage	mortified
murmur	murmurs	myopathy
myopathic	nausea	nauseous
nervousness	neuropathy	numb
numbness	numbing	obstructive sleep apnea
osa	pain	painful
pains	palpitations	pancreatitis
panicky	paresthesia	parkinsonism
persistent	pneumonia	pounding
pressure	priapism	problem
problems	progressive multifocal leukoencephalopathy	pml
psychiatric	psychotic	psychosis
psycho	pulmonary arterial hypertension	pah
pulmonary hypertension	pulsate	pulsating
rapid	rapidly	rash
rashes	react	reaction
reactions	reduce	reducing
reduced	reduction	reductions
regulatory	regulation	regulate
rhythm	rhythms	ringing
runny	rupture	ruptures
ruptured	rupturing	sad
sadness	sclerosis	seizure
seizures	seizing	sensation
sensitivity	sensitive	serious
seriousness	seriously	serotonin syndrome
serotonin toxicity	serotonin sickness	serotonin poisoning
severe	severely	shakiness
shaky	sharp	short
shortness	shortening	shorter
sleep	sleeping	sleepiness
slept	sore	soreness
spasm	spasms	stroke
strokes	stuffy	stuffiness
congest	congested	congestion
sudden	sugar	glycosylate

glycosylated	hemoglobin	hba1c
suicidal	suicide	sweat
sweats	sweating	swell
swelling	swollen	swells
syncope	tachycardia	temper
tempers	tenderness	tendonitis
terrible	terrified	terrifying
thrombosis	thromboembolism	tingle
tingling	tinnitus	tired
tiredness	tire	torsades de pointes
ventricular tachycardia	toxicity	trauma
traumatic	tremor	tremors
tumor	tumors	tumorous
ulcer	ulcers	ulcerative
unexplained	upset	irritated
irritable	irritate	irritation
upsetting	vaginitis	vertigo
virus	viral	vomit
vomits	vomiting	vomited
wart	warts	watery
weak	weaken	weakening
weakness	weaknesses	weight
worse	worsen	worsening

Correlation of Computed Similarity against Human Ratings

Table B.4: Anatomy – Pearson correlation results.

Configuration	CHV-FMA		CHV-LNC		CHV-MSH		CHV-SNOMEDCT	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
batet	-0.0950	0.5116	0.0404	0.7805	-0.1362	0.3455	-0.2833	0.0462
cdist	0.0141	0.9225	0.1449	0.3153	-0.1224	0.3971	0.0945	0.5138
cmatch	-0.4208	0.0023	-0.1452	0.3144	-0.4549	0.0009	0.3032	0.0323
faith	0.0563	0.6978	0.2314	0.1059	-0.0694	0.6318	0.1821	0.2057
jcn	0.0226	0.8760	0.2004	0.1628	-0.0953	0.5103	0.3001	0.0342
lch	0.0572	0.6930	0.1670	0.2465	-0.1284	0.3741	0.1336	0.3550
lin	0.0651	0.6534	0.1860	0.1958	-0.0694	0.6318	0.1219	0.3991
nam	-0.0816	0.5733	0.0451	0.7556	-0.2189	0.1267	0.0183	0.8995
path	0.0141	0.9225	0.1449	0.3153	-0.1224	0.3971	0.0945	0.5138
pks	-0.0257	0.8597	0.1374	0.3412	-0.1216	0.4002	0.0370	0.7987
res	0.0495	0.7330	0.0498	0.7311	-0.2792	0.0496	0.1675	0.2451
sanchez	-0.0179	0.9017	0.0654	0.6517	-0.3198	0.0236	0.3988	0.0041
upath	0.0090	0.9506	0.1414	0.3273	-0.1224	0.3971	0.0711	0.6237
wup	-0.0397	0.7845	0.1214	0.4009	-0.1397	0.3332	-0.0056	0.9691
zhong	0.0250	0.8631	0.1540	0.2855	-0.1114	0.4413	0.1064	0.4620

Table B.5: Reaction – Pearson correlation results.

Configuration	CHV-LNC		CHV-MDR		CHV-MSH		CHV-SNOMEDCT	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
batet	-0.0831	0.5661	0.1034	0.4747	-0.0371	0.7981	-0.0326	0.8222
cdist	-0.0831	0.5661	0.1138	0.4311	-0.0357	0.8056	0.0215	0.8820
cmatch	-0.3233	0.0220	-0.0630	0.6640	-0.0594	0.6820	-0.0873	0.5467
faith	nan	1.0000	0.1856	0.1969	-0.0363	0.8021	0.0129	0.9291
jcn	nan	1.0000	0.1721	0.2321	-0.0351	0.8087	-0.0384	0.7914
lch	-0.0831	0.5661	0.1130	0.4345	-0.0340	0.8147	0.0506	0.7270
lin	nan	1.0000	0.1942	0.1767	-0.0363	0.8021	0.0105	0.9422
nam	nan	1.0000	0.3745	0.0074	0.3454	0.0140	0.0818	0.5722
path	-0.0831	0.5661	0.1138	0.4311	-0.0357	0.8056	0.0215	0.8820
pks	-0.0831	0.5661	0.1204	0.4048	-0.0358	0.8053	0.0302	0.8350
res	-0.0831	0.5661	0.1210	0.4026	-0.0363	0.8021	0.0172	0.9056
sanchez	-0.3156	0.0256	-0.0283	0.8452	-0.0577	0.6908	-0.0488	0.7363
upath	-0.0831	0.5661	0.1133	0.4332	-0.0357	0.8056	0.0203	0.8887
wup	-0.0831	0.5661	0.1225	0.3968	-0.0357	0.8059	0.0296	0.8385
zhong	-0.0831	0.5661	0.1132	0.4339	-0.0349	0.8097	0.0167	0.9082

Information Retrieval Factors

Table B.6: Anatomy – Top 20 SMAs/VCs. (Ranked by Fm_similar)

measure	τ_1	τ_2	τ_{diff}	configuration	Pr_similar	Rc_similar	Fm_similar	Pr_unknown	Rc_unknown	Fm_unknown	Pr_nonsimila	Rc_nonsimila	Fm_nonsimila	Pr_total	Rc_total	Fm_total
jcn	0.8	0.5	0.3	CHV,SNOMEDCT_US	0.78	0.64	0.70	0.00	0.00	0.00	0.79	0.94	0.86	0.76	0.76	0.76
faith	0.7	0.5	0.2	CHV,SNOMEDCT_US	0.78	0.64	0.70	0.00	0.00	0.00	0.79	0.91	0.85	0.74	0.74	0.74
lin	0.8	0.5	0.4	CHV,SNOMEDCT_US	0.78	0.64	0.70	0.00	0.00	0.00	0.73	0.67	0.70	0.58	0.58	0.58
cmatch	0.5	0.5	0.1	CHV,SNOMEDCT_US	0.64	0.64	0.64	0.00	0.00	0.00	0.79	0.94	0.86	0.76	0.76	0.76
sanchez	0.8	0.5	0.3	CHV,SNOMEDCT_US	0.64	0.64	0.64	0.00	0.00	0.00	0.77	0.82	0.79	0.68	0.68	0.68
pks	0.6	0.5	0.1	CHV,SNOMEDCT_US	0.64	0.64	0.64	0.00	0.00	0.00	0.74	0.70	0.72	0.60	0.60	0.60
wup	0.8	0.3	0.5	CHV,SNOMEDCT_US	0.47	0.64	0.54	0.00	0.00	0.00	0.67	0.48	0.56	0.46	0.46	0.46
faith	0.5	0.5	0.1	CHV,FMA	1.00	0.36	0.53	0.00	0.00	0.00	0.71	0.97	0.82	0.72	0.72	0.72
pks	0.6	0.5	0.1	CHV,FMA	0.80	0.36	0.50	0.50	0.33	0.40	0.73	0.91	0.81	0.72	0.72	0.72
lin	0.7	0.3	0.4	CHV,FMA	0.80	0.36	0.50	0.00	0.00	0.00	0.70	0.94	0.81	0.70	0.70	0.70
pks	0.6	0.5	0.2	CHV,FMA	0.80	0.36	0.50	0.33	0.33	0.33	0.72	0.85	0.78	0.68	0.68	0.68
wup	0.8	0.5	0.3	CHV,FMA	0.80	0.36	0.50	0.21	0.67	0.32	0.69	0.55	0.61	0.52	0.52	0.52
upath	0.5	0.4	0.2	CHV,SNOMEDCT_US	1.00	0.27	0.43	0.00	0.00	0.00	0.70	1.00	0.83	0.72	0.72	0.72
path	0.5	0.4	0.2	CHV,SNOMEDCT_US	1.00	0.27	0.43	0.00	0.00	0.00	0.70	1.00	0.83	0.72	0.72	0.72
cdist	0.5	0.4	0.2	CHV,SNOMEDCT_US	1.00	0.27	0.43	0.00	0.00	0.00	0.70	1.00	0.83	0.72	0.72	0.72
pks	0.6	0.5	0.1	CHV,LNC	1.00	0.27	0.43	0.00	0.00	0.00	0.70	1.00	0.83	0.72	0.72	0.72
jcn	0.8	0.3	0.5	CHV,FMA	1.00	0.27	0.43	0.25	0.17	0.20	0.72	0.94	0.82	0.70	0.70	0.70
faith	0.6	0.4	0.2	CHV,LNC	1.00	0.27	0.43	0.00	0.00	0.00	0.70	0.97	0.81	0.70	0.70	0.70
wup	0.7	0.3	0.4	CHV,LNC	1.00	0.27	0.43	0.00	0.00	0.00	0.68	0.91	0.78	0.66	0.66	0.66
sanchez	1	0.3	0.7	CHV,LNC	0.75	0.27	0.40	0.00	0.00	0.00	0.68	0.91	0.78	0.66	0.66	0.66

Table B.7: Reaction – Top 20 SMAs/VCs. (Ranked by Fm_similar)

measure	τ_1	τ_2	τ_{diff}	configuration	Pr_similar	Rc_similar	Fm_similar	Pr_unknown	Rc_unknown	Fm_unknown	Pr_nonsimilar	Rc_nonsimilar	Fm_nonsimilar	Pr_total	Rc_total	Fm_total
pks	0.6	0.4	0.2	CHV,SNOMEDCT_US	1.00	0.30	0.46	1.00	0.09	0.17	0.63	1.00	0.77	0.66	0.66	0.66
res	0.8	0.3	0.5	CHV,MDR	1.00	0.30	0.46	0.00	0.00	0.00	0.62	1.00	0.76	0.64	0.64	0.64
sanchez	0.5	0.4	0.1	CHV,MDR	1.00	0.30	0.46	0.00	0.00	0.00	0.62	1.00	0.76	0.64	0.64	0.64
wup	0.8	0.3	0.5	CHV,SNOMEDCT_US	1.00	0.30	0.46	0.36	0.82	0.50	0.77	0.59	0.67	0.58	0.58	0.58
sanchez	0.9	0.3	0.6	CHV,SNOMEDCT_US	0.75	0.30	0.43	0.29	0.18	0.22	0.69	0.93	0.79	0.64	0.64	0.64
lin	0.6	0.5	0.1	CHV,SNOMEDCT_US	0.75	0.30	0.43	0.00	0.00	0.00	0.63	1.00	0.77	0.64	0.64	0.64
lch	0.8	0.3	0.5	CHV,MDR	0.75	0.30	0.43	0.00	0.00	0.00	0.55	0.76	0.64	0.50	0.50	0.50
res	0.9	0.4	0.5	CHV,SNOMEDCT_US	0.31	0.40	0.35	0.33	0.27	0.30	0.64	0.62	0.63	0.50	0.50	0.50
jcn	0.8	0.3	0.6	CHV,SNOMEDCT_US	1.00	0.20	0.33	0.50	0.09	0.15	0.63	1.00	0.77	0.64	0.64	0.64
faith	0.7	0.3	0.4	CHV,SNOMEDCT_US	1.00	0.20	0.33	0.50	0.09	0.15	0.63	1.00	0.77	0.64	0.64	0.64
path	0.5	0.2	0.4	CHV,SNOMEDCT_US	1.00	0.20	0.33	0.33	0.09	0.14	0.62	0.97	0.76	0.62	0.62	0.62
cdist	0.5	0.2	0.4	CHV,SNOMEDCT_US	1.00	0.20	0.33	0.33	0.09	0.14	0.62	0.97	0.76	0.62	0.62	0.62
upath	0.5	0.3	0.2	CHV,SNOMEDCT_US	1.00	0.20	0.33	0.00	0.00	0.00	0.60	1.00	0.75	0.62	0.62	0.62
wup	0.6	0.4	0.2	CHV,MDR	1.00	0.20	0.33	0.00	0.00	0.00	0.62	1.00	0.76	0.62	0.62	0.62
cmatch	0.6	0.2	0.4	CHV,SNOMEDCT_US	0.67	0.20	0.31	0.25	0.18	0.21	0.69	0.93	0.79	0.62	0.62	0.62
batet	0.5	0.3	0.2	CHV,SNOMEDCT_US	0.19	0.70	0.30	0.00	0.00	0.00	0.75	0.31	0.44	0.32	0.32	0.32
batet	0.5	0.1	0.5	CHV,MDR	0.30	0.30	0.30	0.00	0.00	0.00	0.55	0.76	0.64	0.50	0.50	0.50
batet	0.6	0.1	0.6	CHV,MSH	0.50	0.20	0.29	0.50	0.18	0.27	0.60	0.86	0.70	0.58	0.58	0.58
lch	1	0.1	0.9	CHV,MSH	0.29	0.20	0.24	0.00	0.00	0.00	0.60	0.86	0.70	0.54	0.54	0.54
sanchez	0.7	0.1	0.7	CHV,MSH	0.29	0.20	0.24	0.00	0.00	0.00	0.60	0.86	0.70	0.54	0.54	0.54