INTERPLAY BETWEEN TRAUMATIC BRAIN INJURY AND
INTIMATE PARTNER VIOLENCE: A DATA-DRIVEN
APPROACH UTILIZING ELECTRONIC HEALTH RECORDS

by

LARRY YOUNG LIU

Submitted in partial fulfillment of the requirements

For the degree of Master of Science

Systems Biology & Bioinformatics

CASE WESTERN RESERVE UNIVERSITY

August, 2017
Case Western Reserve University

School of Graduate Studies

We hereby approve the thesis\(^1\) of

LARRY YOUNG LIU

for the degree of

Master of Science

Dr. Mehmet Koyutürk

Committee Chair, Adviser
Department of Electrical Engineering & Computer Science

Dr. Günnur Karakurt

Committee Member
Department of Psychiatry - Case Western Reserve University

Dr. William S. Bush

Committee Member
Department of Epidemiology & Biostatistics

Date of Defense:
April 26th, 2017

\(^1\)We certify that written approval has been obtained for any proprietary material contained therein.
To my loving and supportive parents, relatives, instructors, and dearest friends.
# Table of Contents

List of Figures iv  
List of Tables v  
Legality vi  
Acknowledgement vii  
Abstract ix  
1 Introduction 1  
2 Objective 5  
3 Methodology 6  
  3.1 Framework 6  
    3.1.1 Database and Queries 6  
    3.1.2 Control Cohorts and Associations 7  
    3.1.3 Querying Process and Subsequent Data Sets 10  
  3.2 Statistical Assessments 13  
    3.2.1 Determining the Number of Shared Terms 13  
    3.2.2 Determining the Commonly Prevalent Terms 14  
    3.2.3 Determining the Synergistically Prevalent Terms 18  
4 Results 24  
  4.1 Shared Terms between IPV and TBI 24  
  4.2 Commonly Prevalent Terms between IPV and TBI 28  
  4.3 Synergistically Prevalent Terms between IPV and TBI 35  
  4.4 Significant Terms found in the Interplay between IPV and TBI 45  
5 Discussion 47  
  5.1 Commonly Prevalent and Synergistically Prevalent Terms 47  
  5.2 Limitations 50  
  5.3 Next Steps and Suggested Future Research 52  
6 Conclusion 55  
Complete References 56
List of Figures

1  Analytical Thesis Framework .................................................. 9
2  2x2 Contingency Table to Assess for the Commonly Prevalent Terms between Two Cohorts ......................................................... 15
3  2x2 Contingency Table to Assess for the Synergistically Prevalent Terms between Two Cohorts ............................................... 19
4  Example for Visualizing and Identifying the Conditional Prevalence of Terms between Two Cohorts of Interest ........................................ 23
5  Comparing the Number of Terms Associated with TBI to the Control Conditions ................................................................. 27
6  Distribution of the Common Prevalence of Terms between DA and TBI ................................................................. 30
7  Comparative Histograms for Commonly Prevalent Terms between the TBI Cohort and the Control Cohorts ........................................ 34
8  Synergistically Prevalent Terms Found in both DA|TBI and TBI|DA Comparisons ................................................................. 37
9  Synergistically Prevalent Terms Found in all of the Control Condition Comparisons ................................................................. 38
10 Overlap Flowchart between the Synergistically Prevalent Terms between IPV and each of the Seven Conditions ................................. 42
11 3-Way Venn Diagram of all Significant Terms found in the Interplay between IPV and TBI ......................................................... 46
12 Analytical Thesis Framework with an Alternative Analytical Approach 54
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summary Table of all Queried Cohorts</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Summary Table of the Number of Synergistically Prevalent Terms from</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>all Conditional Prevalence Comparisons between an Outcome Condition</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and IPV</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>List of Synergistically Prevalent Terms in IPV and each of the Seven</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Conditions</td>
<td></td>
</tr>
</tbody>
</table>
Legality

Most of the contents therein this thesis is currently in preparation for submission to a peer-review journal for publication. As such, the release of this thesis for public viewer ship would be from a year of the date of submission to OhioLink on June 9th, 2017.
Acknowledgement

This thesis has been one of the biggest challenges I have ever undertaken and in turn, it has been one of the biggest accomplishments I have ever achieved. The path getting here however has not been a solo endeavor, as the many below has all contributed a part to help me get to where I am today, and in turn, led to this thesis. Henceforth, here is my sincerest dedication of thanks:

Firstly, I would like to thank my parents Lisa and Robert, my aunt Amy, and my dearest grandmother for their constant support and encouragement. I would have never considered pursuing studies beyond my undergraduate career if it was not for their strong words of telling me to continue challenging myself and to strive higher. Furthermore, the idea of even attending graduate school would have never been possible if it was not for the financial support of my aunt Amy and my grandmother. So I thank my entire family for making graduate school (and this thesis) a reality.

Secondly, I would like to thank my Master’s thesis committee, Dr. Günnur Karakurt, Dr. Mehmet Koyutürk, and Dr. William S. Bush. Without their insight, advice, and support, the finish line of my Master’s (and this thesis) would have never been possible. Their patience, dedication, persistence, advice, and expertise was invaluable and helped me to push myself harder and greatly improve my academic and professional skills and expertise. In addition, I would like to thank Ms. Joan Schenkel, Dr. David Lodowski, and Dr. Masaru Miyagi for helping me out with logistics. Without them, I would keep forgetting deadlines, missed out on opportunities, and other minor things throughout the Systems Biology & Bioinformatics program.
Thirdly, I would like to thank my dearest friends from over the years. Each of them has given me support and advice that collectively helped me to make the push of making this thesis happen. I would like to thank Alfredo Luque, Alexander Wang, and Mario Yepez-Olvera for being patient with me when I asked a dumb coding question and still taught me how to resolve coding issues. For that, I have my eternal thanks. I would also want to thank (in no particular order) Jeffrey Hsieh, Sean Dennehy, Steven LaDelfa, Daniyal Syed, Lloyd McCarthy, Joseph Grijalva, Armand Domalewski, Zachary Wesner, August Yadon, Maxwell Lloyd, and many others in some form or manner for encouraging to pursue (and getting through) graduate school, being there to let me rant or mumble when things got tough, providing advice and encouragement on things, or even just helping me out in a bind. For that you too have my eternal gratitude and thanks. A shout-out of thanks as well to my lab mates, Arda Durmaz and Andrew Burr, as well for helping me to brainstorm on ideas when I got stuck.

Lastly, I would like to thank Dr. Ian Korf and Dr. Patrice Koehl from the University of California, Davis. If it was not for taking the bioinformatics courses you both taught in my senior year of undergrad, I would have never considered pursuing bioinformatics or data science as a whole. Those courses alone opened the doors for motivating me to learn how to code, and pushed me to pursue things that I would never even have touched. As such, both of your courses and your works has motivated me five years down the road to pursue this fun, exciting, and pioneering field. In addition, I would like to Dr. Lynn Epstein from the University of California, Davis and Dr. Tony Chern from Diablo Valley College. Without their full recommendation and support for my pursuit of graduate school, this thesis, and the subsequent journey behind it, would have never happened. As such, I thank both of you for your support for me.
Interplay between Traumatic Brain Injury and Intimate Partner Violence: A Data-Driven Approach Utilizing Electronic Health Records

Abstract

by

LARRY YOUNG LIU

Intimate partner violence (IPV) is a prevalent issue that results in overwhelming physical and mental health consequences. It is also known that majority of victims suffer from blunt force in the head, neck and the face area. Injuries to head and neck are among the causes for traumatic brain injury (TBI). TBI often linked to neurological conditions and permanent behavioral disorders. In this study, we aim to characterize the key associations between IPV and TBI by mining de-identified electronic health records (EHR) data from the Explorys platform. We formulate a novel, data-driven, three-step analytical method to find key health associations by comparing prevalent health conditions among IPV, TBI, and six control cohorts. Our analysis suggests that health effects attributed to substance and alcohol abused livers are highly significant in contributing IPV and TBI interplay. Our results would greatly assist in improving existing screening, diagnostic, and treatment procedures of IPV-induced TBI victims, especially with increasing risk correlated with substance and alcohol abuse.
1 Introduction

Intimate Partner Violence (IPV) is an escalating global public health issue that damagingly upsets health-related, societal, and economic development [1]. According to the 2010 CDC National Intimate Partner and Sexual Violence Survey, about 30% of women within their lifetimes has experienced some form of physical violence from an intimate partner [2]. Among them, around 25% (or 1 in 4) of women experience severe physical violence, with 11.2% of women being physically beaten by an intimate partner, 14.2% of women being struck with a fist or a physical hard object, and 17.2% of women being slammed into an object [2]. Furthermore, women are more prone to IPV than men in intimate relations and sustains more violent assaults [2, 1, 3].

Complications from IPV could drastically distress the health of a victim. Most IPV victims would suffer minor injuries but for some, IPV complications would lead to permanent disabilities or death [3, 4]. IPV victims are also susceptible to psychological distresses [4], sexually transmitted diseases, and infections [3, 5], gynecological complications [6], and unexpected pregnancies with subsequent risks to mother and newborn health [3, 7]. Prior studies has also shown that IPV complications can be chronic. Victims with chronic conditions often suffers from various neurological symptoms, mental health and substance disorders, gastrointestinal problems, and chronic pain and physical ailments [4, 8].

However, there has been a prevalent and disturbing rise of IPV victims suffering from Traumatic Brain Injury (TBI) complications [9]. In a 2016 article regarding the association of IPV victims to TBI, it is estimated that at least 60% to 92% of surviving female IPV victims has received facial, head, or neck strangulation injuries [10]. Emergency room records also supports the claim, showing 38% of IPV
victims who received emergency medical treatment exhibits significant head, neck, and facial injuries [11]. A review of limited prior direct literature also finds that most IPV abusers targets their assault on the head, with neck strangulations being also prevalent enough to be classified as a symptom for TBI trauma [10, 12].

IPV victims suffering from TBI complications exhibit unique additional adverse health effects. A systematic review notes that IPV victims who exhibit anxiety, depression, dizziness, and headaches matches victims with postconcussive syndrome or lingering mild TBI [10, 11]. This suggests that in addition to PTSD and stress, these additional symptoms may hint signs of brain injury [10, 11]. A 2003 study on assessing brain damage on female victims reveals that an increase severity of abuse led to a negative correlation to cognitive function and a positive correlation of brain damage [12]. Neuroanatomical studies reveals that women who experienced IPV exhibits symptoms of injuries to four neural structures that affects behavior and decision-making skills: the amygdala, the prefrontal cortex, the hypothalamus, and the hippocampus [10, 13, 14, 15]. Damage specific to the amygdala could severely affect memory, emotional processing, and social learning adaptability [13, 16]. Damage to the amygdala also results in weakened hypothalamic-pituitary-adrenal responses [17]. A damaged hippocampus also impairs a victim’s behavior, as a healthy hippocampus would properly interpret new sequences of events that would assist a person to reach better with its environment; something a damaged hippocampus hinders [10, 18, 19]. Lastly, neuroplasticity repairs on these structures would not result full correct functionality of the pre-traumatic state [10, 20]. This is due to the fact that neuroplasticity repairs does not often take into consideration of restoring optimal cognitive functionality [20]. Therefore, permanent changes of a victim’s behavior and decision making skills is extremely likely [20]. In extreme cases, some victims developed mild temporal lobe epilepsy and permanent hearing loss in the left ear from being hit on the head with solid objects [10].
Moreover, IPV victims suffering from TBI complications are more susceptible to some commonly-attributed IPV health effects when compared to non-TBI IPV victims. Prior screening studies reveal that IPV victims with TBI are further susceptible to having impaired immune functionality, increased asthma attacks, depression and anxiety, higher risk of stroke, hypertension, heart disease, and sexually transmitted diseases [10, 11, 21].

Female IPV victims are often reported to be in poorer health overall than female victims of other crimes [9, 22]. For instance, sexual assault victims are more likely to get treatment from their assault around the initial six months of the attack, and 15-24% increase of service usage for the first year [9]. Additionally, only 27.6% of victims who has had more than two incidents of abuse report they received their medical conditions due to abuse [10, 23]. When compared to medical treatment costs of non-IPV victims, IPV victims tend to pay more for their medical treatments [9]. Clinicians often tend to overlook adverse health effects of IPV victims, attributing these health effects to incorrect causes [1, 9]. Currently established IPV screening processes are limited and not fully reliable as it often relies on the victim’s self-reporting or survey responses for assessment [24]. Most victims would not disclose their abuse to clinicians due to the negative stigma associated behind IPV abuse [10, 25, 26]. To screen IPV victims for TBI complications is even more difficult. Most TBI screening tests are designed for victims of various environmental accidents, making it difficult to reveal any signs of abuse [11]. Interestingly enough, these screening methods seems to be utilized better with the male population as opposed to the female population [11]. Existing screen methods that attempts to vet out IPV abuse from TBI victims has been underutilized and dated [10]. Furthermore, current screening techniques has led clinicians to mis-attribute diagnoses of adverse health effects, including TBI complications, to IPV [1, 9].
Since the current knowledge of specific adverse health effects related to IPV is still limited, subsequent knowledge of IPV victims suffering from TBI complications is even lower. We propose at utilizing a more data-driven study method better investigate the interplay between IPV and TBI and by using Electronic Health Record databases as a source. EHR databases contain a clinician’s diagnosis on a victim’s record. Therefore, it eliminates any inaccurate or incomplete data that self-reporting or survey-based studies provide [1]. Clinicians and healthcare providers could find EHR database records easier to understand and assess from than survey-based or self-reported data. EHR databases utilizes clinical terminology (SNOMED-CT) and coding (ICD-9) as well[1], making it more accurate to find and associate certain diagnoses [1]. Lastly, EHR databases exhibit a huge sample size to work with, often containing over tens of millions of patient records [9].

There has been previous studies utilizing EHR-sourced data-driven studies. Among them, a particularly notable study in 2017 by Whiting et al. shows promise of identifying and categorizing key adverse health effects exhibited by IPV victims by utilizing an EHR database [1]. Using the *Explorys* EHR Platform by IBM Watson Health, they discovered 2,430 significant adverse health effects among 5,870 records of female patients aged among 18-65 who has experienced IPV [1, 9]. After categorizing these health effects into 28 distinct categories, four of these categories stood out: chronic symptoms and disorders, acute injuries, mental and behavioral issues, and gynecological problems [1, 9]. These categories confirm many of the observed adverse health effects discovered in prior IPV observational studies [1].
2 Objective

For this thesis, we incorporate a data driven method that utilizes EHR data. Our approach is based on the premise that, by mining datasets extracted from an EHR database, we can find the key adverse health effects that are associated with (i.e., either arise from or contribute to) the interplay between IPV and TBI. The results from this analysis can be useful in verifying and interpreting the health effects discovered by the observational studies. To systematically investigate the adverse health correlates of IPV and TBI, we explore three questions:

1. How many terms are shared between patients who are exposed to IPV and patients who are diagnosed with TBI? What are these terms?

2. How many terms are commonly prevalent in both patient populations with IPV and patient populations with TBI? What are these terms?

3. How many terms exhibit synergistically prevalent in patients who are exposed to IPV and diagnosed with TBI? What are these terms?

Here, a term refers to a finding or diagnosis that is uniquely defined by Explorys using SNOMED-CT. The first question seeks to identify all terms that are seen alongside IPV and TBI. The second question seeks to identify terms that are associated with both IPV and TBI, in terms of its frequency in patients with IPV and TBI as compared to the overall population. The third question seeks to identify terms that are not necessarily associated with either IPV or TBI, but are associated with the existence of both IPV and TBI in a patient.
3 Methodology

3.1 Framework

3.1.1 Database and Queries

We use data mining to generate answers to the three questions that are posed in the previous section. Data mining is a commonly used technique for extracting nontrivial and inherent patterns from large data sets [27]. EHR databases serve as a good data source since they include a multitude of patient records throughout the nation’s clinical networks and each of these records contain various diagnoses [1, 9]. These databases are also large enough to represent a study population by focusing on a specific diagnosis or finding [9]. For this study, we decided to use the Explorys EHR Platform by IBM Watson Health [28]. The Explorys platform provides over 54 million unique patient records from 344 thousand unique providers nationally [28].

We query Explorys in order to identify the terms that are associated with IPV and/or TBI. Querying is the process of requesting specific information from a database by selecting particular parameters that lead one to the specific information [29]. In our case, each of our queries executed in Explorys will extract health conditions (in the form of diagnostic terms) relevant for each of our cohorts of study. A cohort in this case, represents all of the records that satisfy a set of constraints on their attributes. In our queries, all queries contain the constraint of being female (in the gender field) and being between the ages of 18 and 65 (in the age field) [9]. In addition to these constraints, we define a “cohort” specific to a condition (e.g., IPV or TBI) by adding a constraint that requires the existence of a term (e.g., “Domestic Abuse” for IPV, “Traumatic Brain Injury” for TBI) that
describes respective condition as a Finding or Diagnosis. Once a query is defined in terms of these constraints, Explorys constructs the respective cohort as the set of records that satisfy all the constraints. For each cohort, it is possible to download a table from Explorys, which includes the descriptive statistics for demographics, as well as all the frequencies of the terms that exist in the cohort. The frequency associated with a given diagnostic term with respect to a cohort notates the number of records in the cohort that reference the respective term.

3.1.2 Control Cohorts and Associations

To reduce any confounding bias or collusion, we establish control cohorts for our study [30]. The use of control cohorts also enables assessing the statistical significance of the terms that are identified to be associated with IPV and TBI. For this study, we utilize two sets of control groups, with each control group having three conditions with similar attributes: acute conditions and accident-related incidents.

The acute conditions control group represents conditions that are commonly encountered in the overall population, are not reported to have an association with IPV or TBI, are not chronic, and are not likely complications from other underlying causes. The idea behind using control cohorts that represent acute conditions is that, these conditions likely do not have any association with IPV, therefore, the patterns identified on these cohorts are likely representative of noise and any source of bias in the data. As such, we select appendicitis (App), tonsillitis (Ton), and gallstones (Gall) as our three acute condition cohort controls. These three acute conditions do not often proliferate due to other underlying causes and are not specifically attributed to IPV victims.

The second control group used in this study comprises of accident-related
incidents. Specifically, we select motor vehicle accidents (MVA), sports-related accidents (SA), and falling off stairs (FoS) as the conditions to represent the accident-related control group. These conditions are chosen as likely correlates of TBI, i.e., TBI may stem from any of these three types of incidents. Furthermore, while these conditions are not specifically related to IPV, it is possible that they are used as a decoy for IPV, because of possible legal consequence of IPV or the stigma associated with IPV. Clearly, the association with motor vehicle accidents and sports-related accidents are much weaker than that of falling of stairs, since violence against a woman at home can indeed cause falling of stairs. For these reasons, the accident-related control group serves as a way to distinguish the interplay between TBI and its accident-related causes, from its interplay with IPV, which is more systematic and has physical and psychological components.
Figure 1: **The Thesis Analytical Framework.** The blue squares represent the cohorts of the two main conditions of interest. The red square represents the cohorts of the accident-related control group. The green square represents the cohorts of the acute conditions control group. The blue arrow are diagnostic terms that are commonly exhibited in patients with both IPV & TBI. The red and green arrows depict diagnostic terms that are commonly exhibited in patients with both IPV & the conditions related to both control groups.
Figure 1 represents the overall analytical framework to this thesis. IPV & TBI represents our two key cohorts of interest in addition to all common terms associated with them. In this thesis, we focus on IPV being the key predictor variable for our analysis. Subsequently, TBI and the conditions from both of our control groups become our outcome variables. This is due to the fact that we want to find common terms between IPV and the other conditions that are IPV-induced or complications to IPV and not by other factors. Therefore, TBI becomes our primary outcome variable of interest in determining common terms between the interplay. Common terms that are associated between the acute conditions control group are queried against IPV in addition to the common terms in the accident-related conditions control group against IPV.

3.1.3 Querying Process and Subsequent Data Sets

We utilize the “Explorys Cohort Discovery” tool to create our cohort populations for IPV (DA), TBI, IPV (DA) \(\cap\) TBI, and the “Background” overall patient populations by running separate queries on Explorys. The respective queries are as follows (recall that all queries are restricted to patients who are female and are between 18 and 65 years of age):

1. All records containing the term “domestic violence” in the “Diagnosis or Findings” field (henceforth referred to as the DA cohort).

2. All records containing the term “traumatic brain injury” in the “Diagnosis or Findings” field (henceforth referred to as the TBI group).

3. All records containing the term “traumatic brain injury” \& “domestic violence” in the “Diagnosis or Findings” field (henceforth referred to as the TBI \(\cap\) DA group).

4. All records that satisfy the demographic constraints noted above (henceforth referred to as the “Background” group).
The first and second queries we execute from this list create cohorts that represent all records within the Explorys database that include a prior incidence of IPV or TBI. Concurrently, this query also generates a table that contains the frequency of all possible diagnostic terms in these cohorts. The third query on the list generates a cohort that represents all records within the Explorys dataset that note a prior incidence of IPV and a prior incidence of TBI. The last query represents the “Background” patients, or every record within the Explorys EHR database for which the patient is a woman with age between 18 and 65. We repeated the first three queries for each of our control condition cohorts as well. All queries were run on April 28th, 2017 with Explorys marking the last revision of the database at April 24th, 2017. The entire time frame of records included in our analysis starts from 1999 to April 24th, 2017. Table 1 below summarizes the cohort sizes and the number of terms attributed to all cohorts that we queried from the Explorys platform.
Table 1: **Summary of the cohort size and the number of terms exhibited by each queried cohort from the Explorys EHR Platform.** This includes all intersecting cohorts in addition to the overall population. Cohort demographics remains constant for all queried cohorts, by observing records only to females between the ages of 18 to 65. The demographics are also applied to the overall population cohort as well. The “Overall Population” cohort serves as our “Background.”

<table>
<thead>
<tr>
<th>Query Type</th>
<th>Cohort Size</th>
<th>Number of Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>8140</td>
<td>5460</td>
</tr>
<tr>
<td>TBI</td>
<td>116660</td>
<td>11185</td>
</tr>
<tr>
<td>DA ∩ TBI</td>
<td>610</td>
<td>2248</td>
</tr>
<tr>
<td>App</td>
<td>75600</td>
<td>9755</td>
</tr>
<tr>
<td>DA ∩ App</td>
<td>90</td>
<td>714</td>
</tr>
<tr>
<td>Ton</td>
<td>238220</td>
<td>11185</td>
</tr>
<tr>
<td>DA ∩ Ton</td>
<td>280</td>
<td>1446</td>
</tr>
<tr>
<td>Gall</td>
<td>319320</td>
<td>11185</td>
</tr>
<tr>
<td>DA ∩ Gall</td>
<td>450</td>
<td>2107</td>
</tr>
<tr>
<td>MVA</td>
<td>516040</td>
<td>11185</td>
</tr>
<tr>
<td>DA ∩ MVA</td>
<td>1220</td>
<td>3191</td>
</tr>
<tr>
<td>SA</td>
<td>101570</td>
<td>9547</td>
</tr>
<tr>
<td>DA ∩ SA</td>
<td>190</td>
<td>1223</td>
</tr>
<tr>
<td>FoS</td>
<td>121520</td>
<td>11185</td>
</tr>
<tr>
<td>DA ∩ FoS</td>
<td>690</td>
<td>2655</td>
</tr>
<tr>
<td>Overall Population</td>
<td>12684250</td>
<td>11185</td>
</tr>
</tbody>
</table>
It is important to note that a record noted in Explorys does not uniquely represent one patient. Hence, it is possible for a patient to have more than one record stored and noted in Explorys. This is due to the de-identified nature of the data provided by Explorys, leading to a possibility that there are more than one record attributed to a patient. Without having identified data, we do not have the ability to remove or merge possible duplicate records that exists from clerical or import errors from a healthcare provider [31]. Therefore, our cohort is based on the number of records, not by the number of patients. Consequently, terms attributed to a record can be interrelated for a single event or can represent isolated events. Without access to a chronological time frame for each entry in a record, we assume that records containing both an IPV finding and TBI diagnosis are caused by a single traumatic event from the victim [5]. Lastly, Explorys sets a limit on how much of the queried data set could be exported out of the “Explorys Cohort Discovery” web tool. Therefore, we can only export the top 50% of each queried data set. This means we only have access to the top 11,184 terms that are attributed to each queried cohort and potentially have missing terms that would be applicable for our assessment. However, the terms that are omitted are terms that are less frequent in a cohort, therefore their effect on our results is expected to be minimal.

3.2 Statistical Assessments

3.2.1 Determining the Number of Shared Terms

To determine the number of shared terms between IPV and TBI cohorts, we compare the number of terms found in the TBI ∩ DA to the cohort size of the TBI cohort. This represents the proportion of the number of shared terms found in the TBI cohort relative to the overall number of records attributing to TBI. We use the
same metric of determining the number of shared terms between IPV and each of our control cohorts for both control groups. The proportions are compared via line plots between IPV and TBI against the control cohorts in each control group, with one line plot representing one of the control groups. We use R 3.4.0 via RStudio with the ‘ggplot2’ package to graph the line plots [32, 33, 34].

3.2.2 Determining the Commonly Prevalent Terms

For a term to be considered as commonly prevalent in both the IPV and TBI cohorts, it has to be significantly prevalent in both cohorts when compared to the overall population. As such, we generate 2x2 contingency tables to find terms that are significantly more prevalent in each cohort as compared to the background population. The 2x2 contingency table is calculated based on the frequency of a term in a given cohort and the frequency of the same term in the background population. We created 2x2 contingency tables for all possible terms in each of our eight cohorts (DA, TBI, three acute conditions, three accident-related incidents). The construction of the contingency table for the computation of common prevalence is shown in Figure 2.
Figure 2: Construction of the 2x2 contingency table to assess the significance of the frequency \((f)\) of a diagnostic term \((d)\) in a cohort \((X)\) as compared to the overall population \((BG)\). The numbers \(N_{BG}, N_X,\) and \(f_X(d)\) are obtained from Explorys query results, all other cells in the contingency table are calculated using these numbers.
The 2x2 contingency tables are used to assess the statistical significance of the frequency of each term in each cohort. As shown in Figure 2, for each term, the three numbers that are available in the Explorys query results can be used to construct the entire contingency table for that term. These numbers are $N_{BG}$ (the number of records in the overall population), $N_X$ (the number of records in cohort X), and $f_X(d)$ (the frequency of term $d$ in cohort X). Once the contingency table is constructed, the significance of the frequency of a term in a cohort can be calculated using either $\chi^2$ Independence test with Yate’s and Bonferroni’s correction or Fisher’s Exact Test. However, Fisher’s Exact Test and the $\chi^2$ Independence test are not well-suited to very large sample sizes and the p-values they calculate may not be precise and informative at this level of significance [35]. An alternate measure for scoring and comparing the prevalences of the terms is the odds-ratio, which can be calculated as follows:

$$OR_{(X|BG)} = \frac{(f_X(d))(N_X - f_{BG}(d) - N_{BG} + f_X(d))}{(f_{BG}(d) - f_X(d))(N_X - f_X(d))}$$

We compute the p-values and odds ratios for all terms in Python 3.5.2 using the Anaconda 3 4.2.0 library package from Continuum Analytics, which contains the ‘NumPy’ and ‘SciPy’ packages [36], and executed under the JetBrains PyCharms 2017.1.3 IDE [37]. We specifically use the ‘NumPy’ package to form the matrix consisting the contingency table for each term. We use the ‘SciPy’ package for running the $\chi^2$ Independence test to calculate the p-value with a Yate’s correction on each table. We perform Bonferroni’s correction to the $\chi^2$ p-value, where the number of hypotheses is considered to be the number of terms that have non-zero frequency in the respective cohort. Fisher’s Exact Test and subsequent p-value calculation also utilizes the ‘SciPy’ package. Log odds-ratio calculation uses the
natural logarithm transformation of the odds-ratio results returned by Fisher’s Exact Test.

The p-values and the log-odds ratio values from our three inference tests are then used to assess the common prevalence score of each term between the IPV cohort and the other cohorts (TBI and controls). Our main objective is to identify the terms that are significantly prevalent in both the IPV cohort and TBI cohort. For this reason, we define the common prevalence score (CPS) of a term with respect to the IPV and TBI cohorts as the minimum of its log-odds ratio (LOR) in the IPV and TBI cohorts. As a control, we also compute a common prevalence score for all terms on all of the control cohorts. Namely, for a given term $d$ in cohort $X$, we define the common prevalence score of $d$ in IPV and $X$ as:

$$CPS = \min[LOR(d_{X|BG}), LOR(d_{DA|BG})]$$

Defined this way, the common prevalence score will be large for terms that have high prevalence in both IPV and TBI cohorts.

After all adjustments, we visualize the distribution of common prevalence scores of terms for each cohort using the ‘ggplot2’ package in R 3.4.0. For this purpose, we create a comparative histogram comparing the distribution of common prevalence of terms within the TBI and IPV cohorts versus the distribution of common prevalence of terms in the IPV and control cohorts.
3.2.3 Determining the Synergistically Prevalent Terms

Besides conditions that are significantly prevalent in both the IPV and TBI cohorts, we are also interested in identifying the conditions that are likely to be seen in patients who are exposed to IPV and were diagnosed with TBI. We call such terms “synergistic” terms, since their prevalence in the presence of both IPV and TBI reflects the synergy between these two conditions. To this end, conditions that are prevalent in one or both of the IPV and TBI cohorts may not be synergistic, since the prevalence of these terms may not be enhanced by the existence of both conditions.

To identify synergistic terms, we first define the conditional prevalence of terms. Conditional prevalence refers to the relative prevalence of a term in a cohort when the background population is restricted to be a specific cohort (as opposed to the overall population). Namely, for given cohorts X and Y, we compare the frequency of a term in cohort Y to its frequency in cohort X \( \cap Y \). We call the significance of this comparison the conditional prevalence of the term in cohort X with respect to Y, since it quantifies the increase in the frequency of the term when condition X is considered and condition Y is a given. We utilize 2x2 contingency tables again to assess the conditional prevalence of all terms between our IPV predictor cohort against the eight outcome cohorts. The construction of the contingency table for the computation of conditional prevalence is shown in Figure 3.
Figure 3: Construction of the contingency table for assessing conditional prevalence of a term in the DA cohort given a background cohort X (denoted DA|X). The frequency \( f \) of a diagnostic term \( d \) in cohort \((X \cap DA)\), the size of the restricted cohort \( N_{X\cap DA} \), and the size of the background population \( N_X \) are extracted from Explorys query results; all other cells in the table are calculated based on these numbers. The conditional prevalence of a term in cohort X given background cohort DA (denoted X|DA) is computed similarly, by swapping X and DA throughout.
The difference between the contingency table for conditional prevalence as compared to the table used in Section 3.2.2 for common prevalence with respect to the overall population is that the background population for this analysis is a given cohort X, while the frequency of terms being pooled is from the intersection cohort between IPV and the given cohort X (X ∩ DA). The 2x2 contingency table for conditional prevalence is also calculated twice, one for X|DA (conditional prevalence in cohort X given cohort DA) and the other for DA|X (conditional prevalence in cohort DA given cohort X), since conditional prevalence is directional. The strength of a term’s conditional prevalence found in one direction (e.g., DA|X) can be different from the other (e.g., X|DA), since observation of DA for a patient with condition X can increase the likelihood of a third condition, while observation of X for a patient with condition DA may not have any effect on the likelihood of that condition, or vice versa. Therefore, we construct these contingency tables for DA|X and X|DA for each term present in the X ∩ DA cohort. Similar to the assessment of prevalence with respect to the overall population, the p-values and LOR values are calculated for both X|DA and DA|X simultaneously using the ‘NumPy’ and ‘SciPy’ packages from the Anaconda 3 plug-in [36]. We denote the log-odds ratio of a term d in cohort X with respect to cohort DA as $LOR(d, X|DA)$ and the log-odds ratio of a term d in cohort DA with respect to cohort X as $LOR(d, DA|X)$.

To assess a term’s synergistic association with IPV and cohort X, we compare the term’s conditional prevalence in cohort X with respect to IPV ($LOR(d, X|DA)$) to its prevalence in cohort X with respect to the general population ($LOR(d, X|G)$). Similarly, we compare the term’s conditional prevalence in the IPV cohort with respect to cohort X ($LOR(d, DA|X)$) to its prevalence in the IPV cohort with respect to the general population ($LOR(d, DA|G)$). The motivation behind this approach is that the cohort X and DA are subsamples of the general population (G) and thus the frequency of a term in each of these cohorts is a function of its
frequency in the general population. To be more precise, if the cohort X was drawn uniformly at random from the general population, then the expected value of the relative frequency of a term in cohort X would be equal to its relative frequency in the general population and its variance would be proportional to size of cohort X. While it is possible to analytically characterize these expected values and variances, we rather use an empirical approach to identify terms that have unusually high conditional prevalence with respect to a specific cohort, given their prevalence in the general population.

To empirically identify terms that exhibit unusually high conditional prevalence (X|DA) given their prevalence in the general population (X|G), we group terms based on their prevalence in cohort X with respect to the general population (LOR(d,X|G)). We then assess the distribution of the conditional prevalence of the terms in each group (LOR(d,X|DA)) and construct a 95% one-sided confidence interval based on the observed mean and variance of the terms in the group. To generate the confidence intervals, we first associate the log-odds ratio values for the conditional prevalence and the common prevalence of a term in the cohort. The mean for each binned group is calculated based on the conditional prevalence values in addition to the confidence interval boundaries by 1.65 times the standard deviation of the bin, which represents the 95% one-sided upper boundary of a normal distribution. We identify all terms that fall outside this confidence interval as terms with potential synergistic association with the two cohorts of interest (X and DA), since these terms exhibit very high conditional prevalence than would be expected based on their prevalence in cohort X with respect to the general population. Using R 3.4.0 with the ‘ggplot2’ package,[32, 34] we plot the bins as well as the mean and the 95% confidence interval as a scatterplot (for the sake of visualization we also plot the 95% one-sided confidence interval below the mean). In our visualization, to increase readability, we also normalize conditional prevalence
by background prevalence, so that the mean of the y axis will be equal to one regardless of the value on the x-axis. A sample visualization that is used to identify terms with synergistic association is shown in Figure 4.
Figure 4: **Example of visualizing the method used to identify terms with synergistic association between two cohorts of interest.** In this example, we assess the synergy between the DA and TBI cohorts based on the conditional prevalence of the terms in the DA cohort with respect to the TBI cohort. For this purpose, on the x-axis, we have the log-odds ratio values $LOR(d, DA|G)$ for all terms that have non-zero frequency in the cohort $(X \cap DA)$. We retain terms that are $LOR(d, DA|G) > 0.5$ (the likelihood of observing the term does not go down more than two-fold in the DA population as compared to the general population). On the y-axis, we have the ratio of the conditional prevalence of each term in the DA cohort with respect to the TBI cohort and its prevalence in the DA cohort with respect to the general population $(LOR(d, DA|TBI)/LOR(d, DA|G))$. We divide the terms into 10 bins based on their value on the x-axis, and visualize the mean and the standard deviation of the values on the y-axis for each bin. The green curve shows the mean ratio, the red curve show the 95% one-sided confidence interval for the ratio based on the assumption of normality. The terms that lie above the upper red curve have unusually high conditional prevalence in the DA cohort with respect to the TBI cohort, as compared to their prevalence in the DA cohort with respect to the general population, suggesting a potential synergistic association between these terms and the DA and TBI cohorts.
4 Results

4.1 Shared Terms between IPV and TBI

The number of terms that are common to the IPV and TBI cohorts is 2,248. To assess the significance of this number, we compare the number of terms that are common to IPV and TBI to the number of that are common to the control conditions. These comparisons are shown in Figure 5. As seen in Figure 5(a), as cohort size grows for acute conditions, the number of terms shared with IPV also grows. However, when compared to appendicitis, tonsillitis, and gallstones, TBI has a higher number of shared terms with IPV despite its relatively small cohort size. While the gallstones cohort has nearly three times as many records (319,320) as the TBI cohort (116,660), only 2,107 terms that are present in the IPV cohort are also present in the gallstone cohort. This suggests that it is more likely for one TBI record to contain a term that is also associated with a record that contains IPV as a finding, as compared to a record that contains gallstone as a finding (same for appendicitis and tonsillitis). This observation confirms our expectations in that it indicates there are more shared diagnostic conditions between TBI and IPV than those that are shared between acute conditions and IPV.

When we compare the number of shared terms between TBI and IPV to the number of shared terms between IPV and the cohorts representing the accident-related control group, we observe a different pattern. The results of this analysis are shown in Figure 5(b). As seen in the figure, the number of terms shared between sports-related accidents and IPV is lower than the number of terms shared between TBI and IPV, although TBI and sports-related accidents have similar cohort sizes. This is not surprising, since sports-related accidents are not usually
attributed to IPV. However, when we consider the number of records contributing to falling off stairs incidents (121,520), we observe that the number of terms shared with IPV (2,655) is more than the number of terms shared between IPV and TBI. This observation could suggest that observable conditions sustained from falling off stairs incidents could be similar to those sustained by TBI. It is also possible that some incidents reported as “falling off stairs” can be associated with IPV. When we consider motor vehicle related accidents, we observe that the number of terms shared with IPV (3,191) is in line with the large cohort size associated with motor vehicle related accidents (516,040). In other words, with the exception of falling off stairs incidents, the pattern observed in Figure 5(b) follows a similar pattern observed in Figure 5(a) for the acute conditions control group. These observations suggest that TBI and falling off stairs have more in common with IPV, as compared to acute conditions like tonsillitis, gall bladder, and appendicitis, in addition to accidents including sports-related accidents and motor vehicle accidents.
(a) Number of Terms Shared Between TBI and Acute Conditions

(b) Number of Terms Shared Between TBI and Accident-Related Incidents
Figure 5: **Number of terms associated with both IPV by and TBI, as compared to the number of terms associated with both IPV and the control conditions.** (a) The cyan line shows the number of terms common to IPV and the respective acute condition as a function of cohort size (Left to Right: Appendicitis, Tonsillitis, Gallstones) and the green dot shows the number of terms common to IPV and TBI. (b) The red line shows the number of terms common to IPV and the respective accident as a function of cohort size (Left to Right: Sports-related Accidents, Falling off Stairs Incidents, Motor Vehicle Accidents). The green dot on both shows the number of terms common to IPV and TBI.
4.2 Commonly Prevalent Terms between IPV and TBI

We calculate the common prevalence scores for all terms present in both the IPV cohort and each of the other cohorts, and create comparative histograms depicting the distribution of these scores between the IPV cohort with each of the other cohorts. The histograms comparing the distribution of common prevalence according to the $\chi^2$ Independence test, Fisher’s Exact Test, and log-odds ratios for IPV vs. TBI and IPV vs. acute conditions is shown in Figure 6. As seen in both Figure 6(a) and Figure 6(b), for both $\chi^2$ Independence test and Fisher’s Exact Test, the majority of the terms have p-values to be computed with the numerical precision that is available. Note that around two-thirds of the terms after correction for multiple hypothesis testing is applied to the common prevalence score are significant between the TBI and IPV cohort. This observation illustrates the extreme sensitivity and dependency of p-values to sample size [38]. As sample sizes increase in p-value-based inference tests, even the smallest frequency differences can yield an extremely small p-value [38]. As a result, the p-values generated for analysis are too small to be informative for comparing the prevalence of different terms. In contrast, log-odds ratios are potentially more useful for comparative analyses as log-odds ratio values are more dependent on effect sizes [39]. For this reason, for the rest of this section, we use the LOR values to facilitate comparisons between terms, as well as comparisons between cohorts.
(a) $\chi^2$ Independence test

(b) Fisher's Exact Test

(c) Log Odds-Ratio
Figure 6: **Each histogram shows the distribution of common prevalence of terms in two cohorts, computed for each term as the least significant prevalence score of the term in the two cohorts of interest.** We use three alternate measures to assess prevalence ($\chi^2$ Independence test, Fisher’s Exact test, and Log-Odds Ratio) and each panel shows the distribution of common prevalence with respect to one of these measures. For $\chi^2$ Independence test and Fisher’s Exact test, the p-values are corrected for multiple hypothesis testing and log-transformed.
The common prevalence of the shared terms between IPV and TBI cohorts is significantly higher compared to that of the shared terms between IPV and each of the three conditions representing the acute conditions control group and IPV. Figure 7(a) shows that a substantial portion of observed shared terms between IPV and TBI are prevalent in both cohorts as compared to the background population, with notable number of terms having log-odds ratio values above 1.5 to nearing 3.5. Terms shared between gallstones and IPV had the second-highest observed prevalence with some observed shared terms having a log-odds ratio values of up to 2. This observation suggests that terms associated with both IPV and TBI often appear in victims of both conditions.

The distribution of common prevalence of terms in IPV and each of the three accident-related incidents, in comparison to the distribution of common prevalence of terms in IPV and TBI, are shown in Figure 7(b). As seen in the figure, terms shared between IPV and accident-related conditions exhibit high common prevalence. We observe that TBI is ranked slightly behind from the shared terms associating between falling off stairs incidents and IPV, but slightly above motor vehicle accidents and IPV, and sports-related accidents and IPV. This suggests that shared terms between falling off stairs incidents and IPV have a stronger prevalence than the shared terms between TBI and IPV. Overall, these observations suggest that the physical injuries sustained from IPV and subsequent consequences are most similar to the injuries sustained from falling off stairs as compared to other accidents (motor vehicle and sports-related). Furthermore, IPV shares more with TBI than it does with motor vehicle and sports related accidents, suggesting that injuries to the head can be relatively more common among victims of IPV as compared to victims of such accidents.

To determine the significantly common prevalent shared terms between IPV
and TBI, we calculated 1% and 5% false discovery rates for tonsillitis, gallstones, and appendicitis based on their log-odds ratio values. We choose to use the acute conditions control group over the accident-related control group for the false term discovery rates due to TBI being better ranked within the acute conditions control cohorts than the accident-related control cohorts. In turn, this selection reduces the collusion that could affect the false term discovery rates. We marked the cutoffs in Figure 7(a) with dashed lines representing 5% and solid lines representing 1%.

Gallstones has the highest 5% and 1% false discovery rate cutoff when compared to tonsillitis and appendicitis. As such, any prevalent TBI term greater than the 1% or 5% cutoff from gallstones are the significantly prevalent shared term with IPV. At a 5% cutoff, TBI has 1,023 shared terms with IPV that are significantly prevalent. At the 1% cutoff, TBI has 510 shared terms that are significantly prevalent with IPV.
(a) TBI vs. Acute Condition Controls

(b) TBI vs. Accident-related Incidents Controls
Figure 7: **Common prevalence of terms between TBI and IPV, as compared to the common prevalence of terms between IPV and control conditions.** In both panels, the cyan histogram shows the distribution of shared prevalence (the minimum of the log-odds ratio of being observed in the TBI and IPV cohorts as compared to the overall population) of terms that are common to TBI and IPV cohorts. (a) Comparison to acute conditions. The distribution of shared prevalence of terms that are common to IPV and Appendicitis (salmon pink), Gallstones (Lime Green), and Tonsillitis (magenta) are shown. The dashed line for each acute condition mark the point for which 5% of the terms have shared prevalence larger than the marked point. Solid lines represent the top 1% for each acute condition. (b) Comparison to accident-related conditions. The distribution of shared prevalence of terms that are common to IPV and falling off stairs (salmon pink), motor vehicle accidents (lime green), and sports-related accidents (magenta) are shown.
4.3 Synergistically Prevalent Terms between IPV and TBI

Figure 8 visualizes the terms that are found to be synergistically prevalent between TBI and IPV cohorts for both directions (Fig. 8(a) DA|TBI & Fig. 8(b) TBI|DA). After taking into consideration of both conditional prevalence directions, there are 30 synergistically prevalent terms that are attributed to both IPV and TBI cohorts. When compared to the acute conditions control group cohorts (Figure 9), synergistically prevalent terms found in between appendicitis and IPV (11 terms), and tonsillitis and IPV (15 terms) are less than those found for than TBI and IPV. There are more synergistically prevalent terms between gallstones and TBI (58 terms) than in TBI and IPV. This could suggest that within the gallstone population, there are contributing or effect factors that could suggest IPV trauma to some victims and vice-versa.

Meanwhile, the number of synergistically prevalent terms found between in each of our accident-related controls and IPV (Figure 9), sports-related accidents (2 terms), motor vehicle accidents (13 terms), and falling of stairs incidents (22 terms) is lower than the number of synergistically prevalent terms found in between TBI and IPV. However, for falling off stairs incidents, the number of synergistically prevalent terms found in comparison to the number found in TBI is still considerable. These observations suggest that falling off stairs incidents might also have contributing or effect factors with IPV victims.
(a) Synergistically Prevalent Terms in the DA|TBI Comparison

(b) Synergistically Prevalent Terms in the TBI|DA Comparison
Figure 8: **Synergistically Prevalent Terms Found in DA|TBI and TBI|DA.** All synergistically significant terms are highlighted in orange dots for the respective conditional prevalence direction. Blue dots represent all synergistically prevalent terms that are significant in both conditional prevalence directions. Red lines represent the 95% confidence interval from the mean. Lime green line represents the mean.
Figure 9: Synergistically significant terms found in the acute conditions control group (left two columns) and accident-related control group (right two columns). All synergistically significant terms are noted in orange dots for that particular conditional prevalence direction. Blue dots represent all highly synergistic terms that are present in both conditional prevalence directions. Red lines represent the 95% confidence interval from the mean. Lime green line represents the mean.
If we do not take into consideration of conditional prevalent directions together, we notice a higher number of synergistically prevalent terms associated with TBI|DA or DA|TBI independently. A total of 81 synergistically prevalent terms are attributed to the TBI|DA comparison and 85 synergistically prevalent terms are attributed to DA|TBI comparison. Therefore, there are more unique synergistically prevalent terms attributing specifically to one conditional prevalent direction than terms attributing for both directions. This suggests that synergistically prevalent terms are also highly dependent of directionality, suggesting that there may be multiple mechanisms and/or underlying factors that lead to synergy in the interplay of IPV and TBI. Table 2 summarizes the number of highly synergistic terms found in each control conditions for each conditional prevalence direction.
| Outcome Condition Compared to IPV | # of Synergistic Terms (Bi-Directional) | # of Synergistic Terms (X | DA) | # of Synergistic Terms (DA | X) |
|----------------------------------|----------------------------------------|---------------------------------|---------------------------------|
| TBI                              | 30                                     | 81                              | 85                              |
| App                              | 11                                     | 19                              | 32                              |
| Ton                              | 15                                     | 40                              | 65                              |
| Gall                             | 58                                     | 101                             | 100                             |
| MVA                              | 13                                     | 135                             | 31                              |
| SA                               | 2                                      | 20                              | 47                              |
| FoS                              | 22                                     | 67                              | 85                              |

Table 2: **Summary of the number of synergistically terms found in each conditional prevalence comparison between an outcome condition and IPV.** This includes all possible number of synergistic terms found individually in each conditional prevalence direction in addition to both conditional prevalence directions.
Comparing the terms that are synergistically prevalent between IPV and TBI with those that are synergistically prevalent for IPV and each of the six control conditions, we discover that a few terms are synergistically prevalent across multiple conditions. Figure 10 shows the overlap between synergistically prevalent terms discovered in each cohort.
Figure 10: Overlap between the synergistically prevalent terms between IPV and each of the seven conditions (TBI, three acute conditions, three accident-related incidents). Red squares represents the acute conditions control group. Green squares represents the accident-related control group. Blue square represents the TBI condition. Each arrow is labeled with the number of overlapping synergistically prevalent terms.
Figure 10 shows that 18 synergistically significant terms found in both the TBI and falling off stairs cohorts. The terms mostly represent pulmonary, cardiovascular, and excretory complications in addition to burns found on the limbs and hemorrhaging. These associations further indicate that conditions exhibit from victims of falling on stairs incidents have similar conditions exhibited by TBI victims. Gallstones and tonsillitis share two highly synergistically prevalent terms (disorder of soft tissue, and trichomonal vulvovaginitis). Motor vehicle accidents and gallstones further share one highly synergistically prevalent term (abnormality of surgical wound). Leucopenia is unique that this synergistically prevalent term is found to be associated with gallstones, TBI, and falling off stairs incidents cohorts.
Table 3: **List of synergistically prevalent terms in IPV and each of the seven conditions.** Blue squares represent a synergistically prevalent term that is identified as synergistically prevalent in IPV and the respective condition. Some of these terms can be synergistically prevalent in IPV and multiple conditions.
4.4 Significant Terms found in the Interplay between IPV and TBI

To determine the significant terms found in the interplay between IPV and TBI, we construct a three-way Venn diagram to compare the terms that are commonly prevalent and those that are synergistically prevalent in IPV and TBI. For this purpose, we use a web tool called Venny [40]. To identify a set of terms that are commonly prevalent in IPV and TBI, we choose the 5% false discovery rate computed based on the acute control conditions. The results of this analysis are shown in Figure 11. As seen in the figure, there are only eight significantly prevalent and synergistic terms found in both IPV and TBI.

The final shared terms are the following: malnutrition, acquired thrombocytopenia, post-traumatic wound infection, local infection of wound, poisoning by cardiovascular drug, alcoholic cirrhosis, alcoholic fatty liver, and drug-induced cirrhosis.

Additionally, we can further interpret Figure 11 if we de-couple the conditionally prevalence directions of the DA|TBI and TBI|DA comparisons. In doing so, we find five terms that are highly significant from both the synergistic and commonly prevalent point of view of DA|TBI, and 17 terms from TBI|DA. If we also de-couple the comparisons looking at the synergistically prevalent terms, we would observe 22 highly significant terms that appears between DA|TBI and TBI|DA.
Figure 11: 3-Way Venn Diagram of all significant terms found in the interplay between IPV and TBI. Purple and yellow circles represent the synergistically prevalent terms of TBI|DA and DA|TBI respectively. Green circle represents the commonly prevalent terms in DA and TBI cohorts at the 5% false discovery rate. Upper overlap represents associated significant terms between TBI|DA and DA|TBI. Left overlap represent shared significant terms between TBI|DA and shared prevalent terms between DA|G and TBI|G. Right overlap represents associated significant terms between DA|TBI and shared prevalent terms between DA|G and TBI|G.
5 Discussion

5.1 Commonly Prevalent and Synergistically Prevalent Terms

As previously mentioned, there are eight significant terms that are attributed to the interplay of TBI and IPV that are both commonly prevalent and highly synergistic. These terms are the following: malnutrition, acquired thrombocytopenia, post-traumatic wound infection, local infection of wound, poisoning by cardiovascular drug, alcoholic cirrhosis, alcoholic fatty liver, and drug-induced cirrhosis. The overall results are surprising, as none of the terms are attributed directly towards head and facial trauma-based conditions observed by prior observational studies. However, the terms that were discovered are still relevant and possibly represent the likely of adverse health effects observed in victims of both IPV and TBI. For instance, the presence of wound infections suggests that physical injuries frequently occur in victims of both IPV and TBI, but the location of these wounds are not directly referenced. As for liver cirrhosis and observed fatty liver due to alcohol and substance abuse, they have been observed in IPV victims. Alcohol and drug consumption is often used as a coping mechanism for IPV victims suffering from post-assault [41, 42, 43, 44]. A reference to malnutrition suggests that victims of IPV also have a higher likelihood of suffering from anemia and being underweight compared to non-IPV victims [45]. We were surprised to find thrombocytopenia in our results. Thrombocytopenia interestingly has an association with pregnancy complications [46], but more importantly, is also reported as a side-effect exhibited by severe TBI and head trauma [47, 48].

When comparing the TBI|DA prevalent terms versus the commonly prevalent terms associated with both IPV and TBI in the overall population, we note 17
distinct shared terms. These conditions range from concussion, chronic post-traumatic headache, hematoma, various types of cranial hemorrhages, alcoholic and drug related abuse and poisoning, delirium, pneumonia, and caloric malnutrition. At a quick glance, we could note that victims in the overall population that could have experienced IPV trauma are likely to have also suffered TBI complications if these aforementioned conditions are diagnosed. Most of these conditions exhibited are consistent with conditions noted from prior observational studies with diagnoses of adverse health effects in the head region [10, 11, 12].

Conversely when observing the direction of DA|TBI, we noted five distinct shared terms. These conditions mostly attribute to drug abuse, and hepatitis caused due to drug or alcohol abuse. This set of conditions were perplexing as it mostly focuses on additional liver-related damages. However, it is found that victims who has experienced TBI trauma has a higher risk factor of acquiring liver cirrhosis due to intracerebral hemorrhaging [49]. Subsequently, the risk of acquiring liver cirrhosis or complications to the liver is high from increased alcoholic or drug consumption [41, 42, 43, 44, 49]. This suggests that liver cirrhosis and any subsequent complications of the liver could be a strong condition indicator for the interplay of IPV and TBI, on top of substance and alcohol abuse. This also affirms that such conditions are significant beyond just an extreme likelihood referred earlier. Hence, victims in the overall population that could have experienced TBI trauma are likely to have also suffered from an IPV event could be hinted if they have signs of any adverse liver conditions from drug or alcoholic abuse.

For the overlap between TBI|DA and DA|TBI, which signifies highly significant conditions attributing to either TBI or IPV from the victims suffering from the respective comorbid counterpart population, we observe 22 distinct terms. Among the terms, we note conditions related to the circulatory and cardiovascular systems,
hematoma-related hemorrhages, uteretic stones, leucopenia, pulmonary-related issues, spinal disorders, burns, and drug-based allergies. Most of these conditions seem to be in-line with conditions observed from prior IPV and TBI observational studies, as these conditions are usually injuries directly sustained from an IPV event, or complications after [4, 8, 11, 12]. Interestingly enough, around half of these conditions are also observed with victims who had a falling off stairs incidents, which strongly suggests that falling off stairs incidents may be associated as well to the interplay of TBI and IPV.

The results visualized by Figure 11 changed our initial expectation of our analyses in determining the significant adverse health conditions that contributes to the interplay between IPV and TBI. As stated earlier, by only observing significant health conditions that are both highly synergistic in both conditional prevalent directions and commonly prevalent in the overall population, we will only observe the most extreme cases of conditions that are exhibited by a IPV-induced TBI victim and vice-versa. Therefore, by only considering this overlap representation from Figure 11 to be the sole determinant of significance, paints an incomplete picture of determining the rue significant health effects that formulates the interplay between IPV and TBI. Therefore, one has to also consider the other overlaps observed in Figure 11.

Two of these overlaps indicates that the comparative direction of conditional prevalence between TBI and DA for synergy can be independent factors. For instance, the synergistic terms found in the DA|TBI conditional prevalence comparison is attributed to be highly prevalent for a TBI victim to have suffered an IPV event. Simultaneously, the synergistic terms found in the TBI|DA conditional prevalence comparison is attributed to be highly prevalent for an IPV victim to be suffering from a TBI complication. When each of these factors are considered with
the commonly prevalent terms found in the overall population for TBI and IPV, we note the previously mentioned overlaps, each contributing additional terms that can be considered significant to the interplay of IPV and TBI.

At the same time, an additional overlap can be observed in context of the scope of just the cohort populations. Some terms are observed to be significant that appears to be exclusive to the cohort populations of just IPV and TBI, and directly in line with the overall population. This could infer to the type of significant conditions found in that overlap between TBI|DA and DA|TBI. As referenced earlier, these conditions also are closely aligned to physiological conditions observed in victims from prior studies suggesting that cohort populations could also contribute as a factor.

Thus, there are a possible of four, distinct outcomes in determining the significant adverse health effects and each of these outcomes contributes to a part to the complete deduction of determining the significant health conditions. By focusing each element visualized in Figure 11 independently, clinicians could observe unique patterns of assuming any interplay being involved with a victim depending on the victim’s prior known conditions. This in part, circumvents the lack of a dimension of time of knowing if a victim suffering from a TBI injury was due to an IPV event or vice-versa from the EHR database. As such, our analyses has proved to be robust and will open up new angles of study that we could continue to pursue.

5.2 Limitations

With what is done so far, there are still a few limitations that constrains the potential of our analysis. The data we are able to access on the Explorys EHR platform is HIPAA de-identified data. In other words, the only raw metric for each
term for all of our data-mined cohorts are the frequency of occurrence each term had for records containing our cohort condition. As a result, we do not have the dimension of time for each condition attributed to the record. If access to more information per record is available, we can assess if a record containing TBI and IPV shows both those condition events are related or separate incidences. Therefore without knowing that aspect from the data set, the overall representation of IPV and TBI in that context is severely underrepresented within the entire EHR population. In other words, notable observed symptoms attributing IPV-induced TBI could still be attributed either to just IPV or just TBI under Explorys’ record assessment. Yet as mentioned earlier, by de-coupling the conditional prevalence comparisons of synergy, we can to some extent work around this limitation.

However, conditions attributing heavily towards TBI is still presumed to be confounded to other conditions especially with the falling off stairs incident cohort. Throughout our analyses, the shared terms found strongly commonly prevalent, highly synergistic, and comorbid between TBI and IPV are extremely close in-line with FoS incidents and IPV. This pattern consistently infers there is an association between TBI and FoS incidents, and the physiological conditions attributing to both of them appears to be similar. However it is also of note that FoS incidents could be perceived as a confounding to IPV as well, since victims can fall down stairs as a reflexive reactionary event during an assault. The resulting trauma sustained from such an incident could be TBI-like. However, a study from Crandall et al. notes that facial injuries are statistically more prevalent to occur for an IPV victim during a violent assault than from a motor vehicle accident or from a fall [50]. Hence, using falling off stairs as a control cohort would still deem appropriate in this context. Yet based on our results from our analyses, Crandall et al.’s conclusion may not sound as decisive as it seems. Further exploration regarding the association of FoS incidents (or just falls in general) to both TBI and IPV would help clear this
possible issue. Nevertheless in regards to our study objectives, we retain our assumption of the FoS incident control condition to be a possible limitation. Lastly, the assumption mode made for each medical record from our EHR database queries represent a unique patient stands.

5.3 Next Steps and Suggested Future Research

There are a few possible next steps that could help find more shared associative terms that are exhibited by victims of both IPV and TBI. One method would be to utilize a network associative map that observes for neighbor-to-neighbor terms that are associated with the each of our overlaps visualized by Figure 11 in our study. An example of this method was done by Whiting et al. in a 2017 study where they tried to find relative associations of categorized IPV symptomatic conditions that frequently could occur in an IPV victim [1]. This map not only validated many associative symptoms found in IPV victims from observational studies, but also hinted at novel associations such as stress being highly associated to compromises of the patient’s immune, circulatory, and gastrointestinal systems [1]. Therefore, a comorbid network associative map generated from the terms in each overlap being the new cohorts, could yield us more associative terms linking IPV with TBI.

Another step would be to repeat this study by using datasets queried from another EHR database system. Another large scale EHR database system like Epic would help to not only validate our existing results, but also would open us to potentially discovering new significant terms of interest and even bigger population sizes. This also would help to expand our overall diversity of possible diagnostic terms in addition to population demographics and size.

For researchers wanting to continue to dig deeper into the interplay between
IPV and TBI, focusing studies on substance and alcoholic-induced, certain circulatory, excretory, gastrointestinal, and cranial complications conditions would potentially open up new avenues, especially within the focus of substance and alcohol abuse as factors. Most of these listed conditions are also supported by prior literature from both an IPV and TBI point-of-view. Therefore, these further studies would lead to the verification of certain conditions being a definite equivalent of a biomarker for the interplay of TBI and IPV. Subsequently, these studies could also improving screening, diagnostic, and treatment for seeking out IPV-induced TBI victims or validating an IPV abuse suffered by a TBI victim.

Lastly, there is alternative approach that we could verify our proposed method. When mentioning our overall study framework with Figure 1, we assigned IPV as our predictor variable for the study, with TBI and all of the control conditions being the outcome variables. Therefore, all of our control condition cohorts were being compared to with IPV directly and not with TBI. However Figure 12 demonstrates that based on our existing study framework, one could make TBI the predictor variable instead of IPV, and IPV becomes one of the outcome variables. By making TBI the predictor variable, we could find common terms that are IPV-induced from TBI victims. This alternative method could be more difficult to analyze as our control conditions are more prone to collusion with the TBI cohort than the IPV cohort. However, this approach is possible to analyze and in theory, would yield similar results to our current approach. Theoretically, both approaches should complement each other within the overall population. As such, we would potentially get similar overall results to help validate our current method, despite our controls being more colluding with TBI. Simultaneously, we will even have the possibility of discovering additional significant shared terms. Therefore, utilizing this next step would be a wise one that could be immediately done.
Figure 12: **The analytical framework of this study but also showing the alternative analytical approach.** The blue squares represent the cohorts of the two main conditions of interest. The red square represents the cohorts of the accident-related control group. The green square represents the cohorts of the acute conditions control group. The arrow marked with (1&2) are diagnostic terms that are commonly exhibited in patients with both IPV & TBI. The arrows marked with (1) are diagnostic terms that are commonly exhibited in patients with both IPV & the conditions related to both control groups. Although not visualized, every cohort square in this flowchart is considered a subset to the overall population.
6 Conclusion

Data queried from electronic health record databases opens up a new perspective on trying to find the significant adverse health effects that constitutes the interplay between intimate partner violence and traumatic brain injury. The aim of using electronic health record data is to not only verify existing known adverse health effects referenced by existing and limited observational studies, but to hopefully discover new patterns and health effects often overlooked by those studies. In this thesis, we develop a novel data-driven analytical method that would try to determine these adverse health effects from an active electronic health record platform queries. The potential results from our analysis would assist to jumpstart new angles of research to improve the accuracy and confidence of existing clinical screening techniques on determining IPV-induced TBI diagnoses from victims. Furthermore, the method we propose could also become a starting point for developing a versatile and predictive tool that assists clinicians to better determine a diagnoses of IPV-induced TBI to a victim.

Our proposed method would help to bring a new perspective and approach in fully understanding the interplay between IPV and TBI. Our results would also be complementary to the already existing repository of results on this growing global condition. We hope that our proposed method and the subsequent results would help clinicians and other research get closer towards improving the healthcare treatments and quality of life to victims of IPV-induced TBI globally.
References


