Noname manuscript No.
(will be inserted by the editor)

Identifying health correlates of intimate partner violence against pregnant women

Serhan Yılmaz · Bushra Alghamdi · Srinidhi Singuri · Ahmet Mert Hacıaliefendioğlu · Tülin Özcan · Mehmet Koyutürk · Günnur Karakurt

Received: date / Accepted: date

Abstract
Purpose: Violence against women during pregnancy is a serious public health concern due to its significant adverse health consequences for both the mother and the baby. This study aims to systematically identify common health problems and synergistic health correlates of intimate partner violence (IPV) that specifically affect pregnant women.
Methods: We mine large-scale electronic health record (EHR) data from the IBM Explorys database to identify health problems that are prevalent in both IPV and pregnancy populations, as well those that are synergistically associated with the presence of IPV during pregnancy. For this purpose, we develop methods that enhance the statistical reliability of identified patterns by constructing confidence intervals that take into account systematic bias and measurement errors in addition to the variance in estimation.
Results: We identify with high confidence 668 and 2750 terms that are respectively prevalent in respectively IPV and pregnancy populations. Of these terms, 279 are common. We also identify 16 synergistic health correlates with high confidence. Our results suggest that mental health, substance abuse, and genitourinary complications are prevalent among pregnant women exposed to IPV. The synergistic terms we identify reveal potential conditions that can be direct consequences of trauma (e.g., tibial fracture), long-term health consequences (e.g., chronic rhinitis), markers associated with the demograhics of affected populations (e.g., acne), and risk factors that potentially increase vulnerability during pregnancy (e.g., disorders of attention and motor control).
Conclusions: Our results indicate that IPV significantly affects the well-being of pregnant women in multiple ways. The findings of this study can be useful for screening of IPV in pregnant women. Finally, the methodology presented here can also be useful for investigating the synergy between other medical conditions using EHR databases with privacy constraints.

Keywords IPV · Pregnant Women · EHR · Data Mining · Synergistic Health Correlates

Serhan Yılmaz
Department of Computer and Data Sciences
Case Western Reserve University, Cleveland, OH, USA
E-mail: serhan.yilmaz@case.edu

Bushra Alghamdi
Systems Biology and Bioinformatics Program
Case Western Reserve University, Cleveland, OH, USA

Srinidhi Singuri
Cleveland Clinic Lerner College of Medicine
Case Western Reserve University, Cleveland, OH, USA

Ahmet Mert Hacıaliefendioğlu
Department of Computer and Data Sciences
Case Western Reserve University, Cleveland, OH, USA

Tülin Özcan
Department of Reproductive Biology
Case Western Reserve University, Cleveland, OH, USA
University Hospitals, Cleveland, OH, USA

Mehmet Koyutürk
Department of Computer and Data Sciences
Center for Proteomics and Bioinformatics
Case Western Reserve University, Cleveland, OH, USA

Günnur Karakurt
Department of Psychiatry
Case Western Reserve University, Cleveland, OH, USA
University Hospitals, Cleveland, OH, USA

1 Introduction

Background. Intimate partner violence (IPV), sometimes referred to as domestic abuse (DA), is defined by

the Centers for Disease Control and Prevention (CDC) as physical, sexual, or emotional abuse by a current or former intimate partner [3]. According to The Bureau of Justice Statistics, IPV accounts for 21% of all violent crime, with 1 in 4 women experiencing severe forms of IPV [23]. IPV has significant adverse effects on the physical and mental well-being of affected women. Harms to physical health experienced by IPV survivors include acute and often visible injuries such as bruises, lacerations, fractures, sight and hearing damage, in addition to chronic conditions such as hypertension, irritable bowel syndrome, fibromyalgia, asthma exacerbation, and chronic pain syndromes [8, 17]. Mental health consequences of IPV include acute responses, such as emotional distress, suicidality, as well as chronic conditions and patterns of behavior including depression, anxiety, eating-disorders, sleep-disorders, post-traumatic stress disorders, and substance use [17].

Significance. Many of the adverse health effects of IPV are amplified during pregnancy [5]. When IPV is experienced by pregnant women, the consequences are extended to both mother and child. The likelihood of suffering from miscarriage, stillbirth, fetal injury, preterm birth, and low birth weight increases with the presence of IPV [5, 17]. Indeed, pregnancy outcomes such as preterm birth and low birth weight are suggested to be mediated by mechanisms involving the prolonged stress experienced by survivors of IPV [5]. Non-fatal injuries incurred to the child during gestation are particularly harmful because poor pregnancy outcomes are associated with a host of physical and developmental problems that persist into adulthood [14]. This potential harm is corroborated by findings that confirm the high prevalence of mental health consequences these mothers face during and after pregnancy [25]. Furthermore, pregnant women subjected to IPV delay time of prenatal care, exhibit poor maternal nutrition, and are more likely to use tobacco, alcohol, and illicit drugs during their pregnancy [25]. All of these social, mental, and behavioral factors predispose women to poor mental and behavioral health, and pose serious additional harms to the development of the child.

Objectives. In this study, we utilize electronic health records (EHRs) to systematically investigate the relationship between IPV and pregnancy by identifying specific health correlates of IPV against pregnant women. Specifically, we aim to answer the following research questions:

1. What are the diagnostic terms that are frequently observed in the presence of pregnancy and IPV? We call these *prevalent* terms. Such terms can be potentially indicative of IPV in pregnant women.
2. Does IPV exhibit significant *co-occurrence* with pregnancy i.e., is IPV more frequently observed among pregnant women compared to the general population? If so, to what extent?
3. Are there diagnostic terms whose presence lead to stronger co-occurrence of IPV and pregnancy as compared to the general population? We refer to such terms as *synergistic* terms. Note that, synergistic terms are not necessarily strongly associated with IPV and pregnancy individually. Such terms can potentially indicate adverse health effects of pregnancy that are further aggravated by IPV. Alternatively, they can be health consequences of IPV that affect pregnant women more dominantly.

Approach. To help answer these questions, we utilize the IBM Explorys database, which provides access to millions of patient records across hundreds of hospitals in the United States. The abundant data in Explorys provides opportunities for data mining tools to uncover health correlates of specific target conditions like "IPV against pregnant women". However, the privacy measures employed by this database pose significant challenges: 1) Individual patients records are not available. The data is provided only in the form of frequencies, i.e., the number of records that contain each term in a given cohort. 2) The frequencies reported by the database are obfuscated by rounding to the nearest ten. The additional uncertainty due to rounding poses additional challenges for investigating rare conditions.

To tackle these challenges, we define statistical criteria to rigorously assess the prevalence and co-occurrence of diagnostic terms by taking into account the sample variation, selection bias, and uncertainty due to rounding. To ensure the robustness of our results, we repeat our analysis for several control conditions that are either: (i) clearly related to pregnancy such as miscarriage, or (ii) acute conditions that are not expected to be strongly related to IPV or pregnancy, such as chronic kidney disease. By systematically applying our computational and statistical framework to large scale EHR data, we identify several significantly prevalent and/or synergistic terms that may provide insights into the interplay of IPV and pregnancy.

Contributions. Overall, the contributions of this study can be summarized as follows:

– We develop a computational and statistical framework to systematically investigate the relationship between multiple conditions using electronic health

records (EHR) data presented in the form of frequencies. Using this framework, we investigate the relationship between intimate partner violence (IPV) and pregnancy.
- We suspect that IPV can be susceptible to a positive reporting bias due to more meticulous record keeping. We develop methods to estimate the effect of such systematic bias in available EHR data, document the extent of bias, and compute confidence intervals for the prevalence and co-occurrence of diagnostic terms by accounting for this bias.
- We introduce the notions of shared prevalence and synergy as distinct notions in characterizing the relationship between a diagnostic term and two conditions of interest. In the context of IPV and pregnancy, we show that diagnostic terms that are significantly prevalent in both IPV and pregnancy cohorts (separately) are distinct from terms that increase the likelihood of observing IPV and pregnancy together.
- By utilizing control cohorts that represent conditions that are potentially associated (either negatively or positively - miscarriage, intra-uterine device, endometriosis) or not associated (chronic kidney disease) with pregnancy, we further investigate the ability of proposed methods in discerning signal from noise.

2 Materials and Methods

2.1 Data collection

IBM Explorys Cohort Discovery is a browser-based search engine that allows querying for frequencies of diagnostic terms in patient cohorts constructed based on the existence of specific diagnoses, findings, or demographics [1, 18]. A query requires specifying (i) ICD-9/ICD-10 codes and (ii) demographic attributes as criteria for inclusion and/or exclusion, which are used to form a cohort of records. For consistency throughout this paper, we use "term" to refer to a clinical diagnosis we obtain from this search tool.

2.2 Querying and Cohort Formation

In order to investigate the relationship of IPV and pregnancy, we first generate our cohorts of interest by performing separate queries (Fig. 1a). These queries/cohorts are specified as follows:

- *Background (BG) Cohort*: All records of women 18-65 years of age with a diagnosis of a disease. All other cohorts are a subset of this cohort.
- *IPV Cohort*: All records that contain "Domestic Abuse" in the findings field.
- *Preg Cohort*: All records that contain "Pregnant" in the findings field.
- *IPV&Preg Cohort*: All records that contain both "Domestic Abuse" and "Pregnant" in the findings field.

We also generate cohorts for each of the control conditions we consider. Letting Z denote one of Chronic Kidney Disease (CKD), Endometriosis (Endo), Miscarriage (MC) and Intrauterine Device (IUD), we construct the following cohorts for each control condition Z:

- *Z Cohort*: All records that that contain "Z" in the findings or diagnosis field.
- *IPV&Z Cohort*: All records that contain both "Domestic Abuse" and "Z" in the findings or diagnosis field.
- *Preg&Z Cohort*: All records that contain both "Pregnant" and "Z" in the findings or diagnosis field.

We run all queries in October 2018. In total, we obtain 16 cohorts representing: the background population, IPV, Pregnancy, four control groups, and the combinations of IPV and Pregnancy with each other and the control groups (Supplementary Data 1). For each cohort X, we obtain the following information: (1) Cohort size N_X i.e., the number of records in X and (2) A frequency table f_X that contains, for each term t, the frequency $f_X(t)$ of t in cohort X (i.e., the number of records in X having diagnosed with t, Fig. 1b). Table 1 shows the cohort sizes and the number of terms with non-zero frequencies for each of these 16 cohorts.

For two conditions of interest X and Y (IPV, Preg, or one of the control groups), we utilize the cohort sizes and term frequency tables to compute contingency tables for each term. Namely, we compute the number of records $N(t, X, Y)$ for all combinations of X, Y, and t (Fig. 1c). We use the following notation:

- N: Cohort size (number of records) of background (BG) cohort.
- $N(t)$: Number of records diagnosed with t. It is directly obtained from term frequency table f_{BG} of BG cohort.
- $N(X)$, $N(Y)$: Cohort sizes of cohort X and cohort Y.
- $N(t, X)$, $N(t, Y)$: Number of records diagnosed with t in cohort X, in cohort Y. Directly obtained from term frequency tables f_X and f_Y.
- $N(X, Y)$: Cohort size of $X\&Y$ cohort.
- $N(t, X, Y)$: Number of records diagnosed with t in $X\&Y$ cohort. Directly obtained from term frequency table $f_{X\&Y}$.

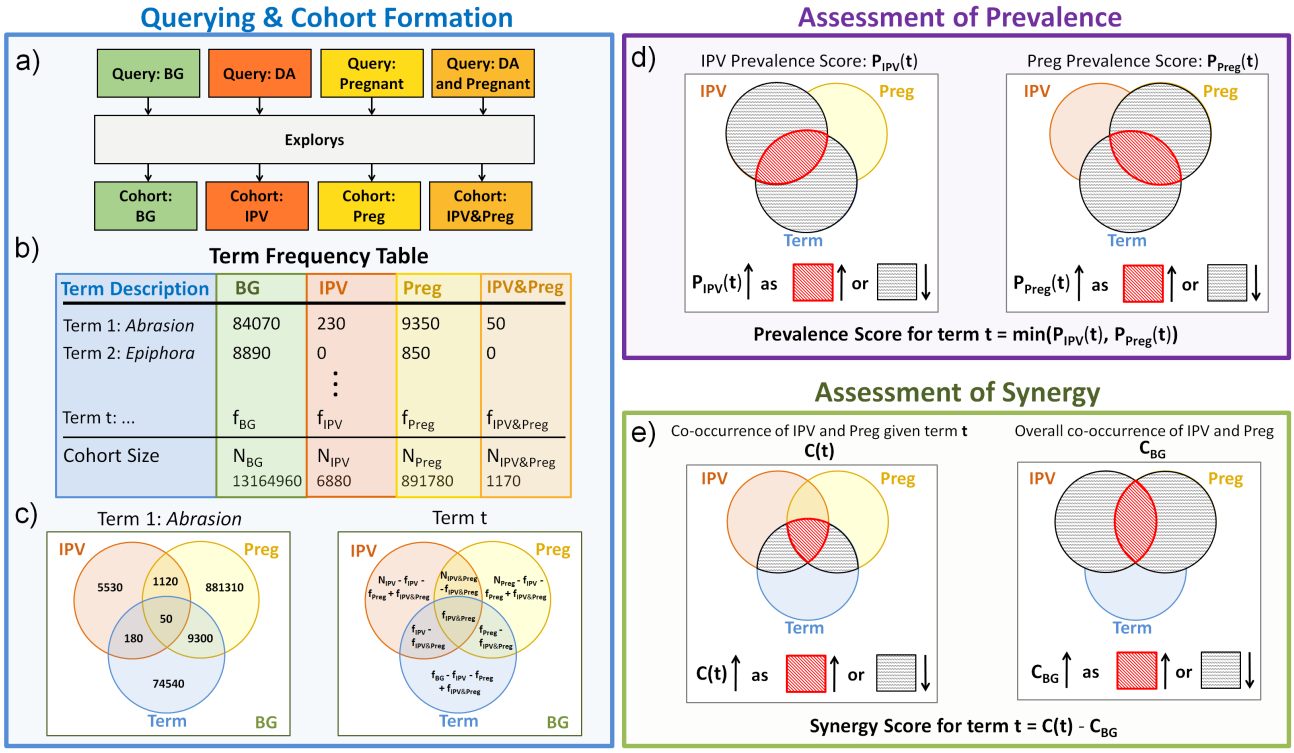

Fig. 1 The flow chart of our pipeline for mining electronic health records to investigate the interplay between IPV and pregnancy. (a) The generated cohorts for the analysis of IPV and pregnancy. (b) Each generated cohort contains a frequency table indicating the number of records for each term. Cohort size is the total number of records in a cohort. (c) Contingency tables (shown as Venn diagrams) are constructed for each term using the term frequencies and cohort sizes. (d, e) Assessment of prevalence and synergy. Shaded red fields indicate the portion of contingency table that increases the corresponding prevalence or synergy score. BG: Background, DA: Domestic Abuse, IPV: Intimate Partner Violence.

2.3 Data Analysis

2.3.1 Computing prevalence score

For a condition of interest X (e.g., IPV, pregnancy), we consider a term to be *prevalent* if it is significantly more frequently observed in cohort X as compared to background. To assess the prevalence of a term in cohort X, we first construct 2×2 contingency tables $N(\mathbf{t}, \mathbf{X})$. Then, for each cohort X (IPV, Preg or control cohorts), and term t, we compute a prevalence score $\mathrm{P}_X(t)$ which is equal to the log-odds ratio $\mathrm{LOR}(t, X|BG)$:

$$\begin{aligned}\mathrm{LOR}(t, X|\mathrm{BG}) &= \log_2 \left(\frac{N(t,X)N(\neg t,\neg X)}{N(\neg t,X)N(t,\neg X)} \right) \\ &= \log_2 \left((N - N(t) - N(X) + N(t,X) \right) \\ &+ \log_2 N(t,X) - \log_2 \left(N(t) - N(t,X) \right) \\ &- \log_2 \left(N(X) - N(t,X) \right)\end{aligned} \tag{1}$$

As illustrated in Fig. 1d, $\mathrm{LOR}(t, X|\mathrm{BG})$ increases as the frequency of term t in cohort X goes up in relation to the frequency of term t in the background cohort and the size of cohort X.

2.3.2 Accounting for variance

In order to assess the variability in the estimation of $\mathrm{LOR}(t, X|BG)$, we compute the standard error as follows:

$$\begin{aligned}se(t,X) = \frac{1}{\ln(2)} \Bigg\{ &\frac{1}{N(t,X)} + \frac{1}{N(t,\neg X)} \\ &+ \frac{1}{N(\neg t,X)} + \frac{1}{N(\neg t,\neg X)} \Bigg\}^{1/2}\end{aligned} \tag{2}$$

Then, we compute the 95% confidence interval:

$$\begin{aligned}\mathrm{LOR}_{\min}(t,X|\mathrm{BG}) &= \mathrm{LOR}(t,X|\mathrm{BG}) - 1.96se(t,X) \\ \mathrm{LOR}_{\max}(t,X|\mathrm{BG}) &= \mathrm{LOR}(t,X|\mathrm{BG}) + 1.96se(t,X)\end{aligned} \tag{3}$$

2.3.3 Accounting for measurement error

The confidence interval shown in Equation 3 accounts for variance but does not consider the measurement error due to the rounding of frequencies. For example, if $N(t, X) = 10$, this indicates that the frequency of term t in cohort X is between 5 and 15, which is a sufficiently large range that could potentially affect the log-odds ratio a considerable amount for relatively rare terms. To

Table 1 The cohort sizes and number of diagnostic terms that appear in each cohort. Cohort size refers to the number of records (patients) with the corresponding diagnosis or finding (e.g., pregnant). "Number of terms" refers to the number of observed terms with non-zero frequency in the corresponding cohort. IPV: Intimate Partner Violence, Preg: Pregnancy, CKD: Chronic Kidney Disease, Endo: Endometriosis, IUD: Intra-Uterine Device, MC: Miscarriage. The background cohort corresponds to the population of females 18-65 years of age that has a diagnosis record in Explorys database. Note that, Explorys provides all numbers (including cohort sizes) rounded to the nearest ten.

	Cohorts of Interest		
Query/Cohort	**Preg**	**IPV**	**Preg&IPV**
Cohort Size	891780	6880	1170
Number of Terms	15488	5465	2960
	Control Cohorts		
Query/Cohort	**CKD**	**IPV&CKD**	**Preg&CKD**
Cohort Size	150240	160	8010
Number of Terms	13954	1589	7017
Query/Cohort	**Endo**	**IPV&Endo**	**Preg&Endo**
Cohort Size	251700	340	24070
Number of Terms	13713	2037	8486
Query/Cohort	**IUD**	**IPV&IUD**	**Preg&IUD**
Cohort Size	210460	210	73190
Number of Terms	12432	1250	10102
Query/Cohort	**MC**	**IPV&MC**	**Preg&MC**
Cohort Size	284640	460	130960
Number of Terms	12785	1864	11489
Query/Cohort	**Background Cohort**		
Cohort Size	13164960		
Number of Terms	18863		

account for the additional uncertainty due to rounding, we construct an *augmented confidence interval*: We consider all possible values of frequency (number of records) to construct the corresponding LOR confidence intervals and report the minimum and maximum bounds among all computed intervals. The log-odds ratio in Equation 1 is monotonic with respect to the counts i.e.,

- $\mathrm{LOR}(t, X|BG)$ increases as N or $N(t, X)$ increases.
- $\mathrm{LOR}(t, X|BG)$ decreases as $N(t)$ or $N(X)$ increases.

Thus, it is sufficient to consider the corner cases of frequency to compute the augmented confidence intervals:

- *Min* bound: Compute interval for $\bar{N} = N - \Delta$, $\bar{N}(t) = N(t) + \Delta$, $\bar{N}(X) = N(X) + \Delta$ and $\bar{N}(t, X) = N(t, X) - \Delta$
- *Max* bound: Compute interval for $\bar{N} = N + \Delta$, $\bar{N}(t) = N(t) - \Delta$, $\bar{N}(X) = N(X) - \Delta$ and $\bar{N}(t, X) = N(t, X) + \Delta$

where $\Delta = 5$ since the rounding is to the tenth digit.

2.3.4 Identification of prevalent terms

In this study, we are interested in identifying diagnostic terms that are over-represented in both IPV and Preg cohorts. In order to identify terms that are prevalent in *both* cohorts of interest X, Y (e.g., IPV, Preg), we first compute $\mathrm{P}_X(t) = \mathrm{LOR}(t, X|\mathrm{BG})$ and $\mathrm{P}_Y(t) = \mathrm{LOR}(t, Y|\mathrm{BG})$ as well as the corresponding augmented confidence intervals. Then, we utilize the following combined prevalence scoring function $\mathrm{P}(t)$:

$$\begin{aligned}\mathrm{P}(t) &= \min\left(\mathrm{P}_X(t), \mathrm{P}_Y(t)\right) \\ &= \min\left(\mathrm{LOR}(t, X|\mathrm{BG}),\ \mathrm{LOR}(t, Y|\mathrm{BG})\right)\end{aligned} \tag{4}$$

We compute an augmented confidence interval for $\mathrm{P}(t)$ as follows:

$$\begin{aligned}\mathrm{P}_{\min}(t) &= \min\left(\mathrm{LOR}_{\min}(t, X|\mathrm{BG}),\ \mathrm{LOR}_{\min}(t, Y|\mathrm{BG})\right) \\ \mathrm{P}_{\max}(t) &= \min\left(\mathrm{LOR}_{\max}(t, X|\mathrm{BG}),\ \mathrm{LOR}_{\max}(t, Y|\mathrm{BG})\right)\end{aligned} \tag{5}$$

Note that, a high $\mathrm{P}(t)$ value will necessarily imply a high value for $\mathrm{P}_X(t)$ and $\mathrm{P}_Y(t)$ as well since $\mathrm{P}(t) \geq \mathrm{P}_X(t)$ and $\mathrm{P}(t) \geq \mathrm{P}_Y(t)$. Thus, a high $\mathrm{P}(t)$ value indicates high prevalence in both X and Y cohorts.

2.3.5 Computing co-occurrence score

We define the *co-occurrence* between two conditions X and Y (e.g., IPV and pregnancy) as a tendency of both conditions to occur together, which can be measured using the observed frequencies. To assess the overall co-occurrence of two conditions X and Y in the background population, we first construct a 2×2 contingency table $N(\mathbf{X}, \mathbf{Y})$. Then, we compute an the overall co-occurrence score $\mathrm{C}_{\mathrm{BG}}(X, Y)$ which is equal to the log-odds ratio $\mathrm{LOR}(X, Y|\mathrm{BG})$:

$$\begin{aligned}\mathrm{C}_{\mathrm{BG}}(X, Y) &= \mathrm{LOR}(X, Y|\mathrm{BG}) \\ &= \log_2\left(\frac{N(X, Y)N(\neg X, \neg Y)}{N(\neg X, Y)N(X, \neg Y)}\right)\end{aligned} \tag{6}$$

Similarly, we assess the *conditional co-occurrence* $\mathrm{C}_t(X, Y)$ of two conditions X and Y in a sub-population where term t is present. For this purpose, we construct 2×2 contingency tables $N(t, \mathbf{X}, \mathbf{Y})$ for each term t and measure the log-odds ratio $\mathrm{LOR}(X, Y|t)$:

$$\begin{aligned}\mathrm{C}_t(X, Y) &= \mathrm{LOR}(X, Y|t) \\ &= \log_2\left(\frac{N(t, X, Y)N(t, \neg X, \neg Y)}{N(t, \neg X, Y)N(t, X, \neg Y)}\right)\end{aligned} \tag{7}$$

Both co-occurrence scores $\mathrm{LOR}(X, Y|\mathrm{BG})$ and $\mathrm{LOR}(X, Y|t)$ compares the risk (measured by odds) of X when Y is present to the risk of X when Y is not present. Thus, as illustrated in Fig. 1e, $\mathrm{C}_t(X, Y)$ goes up as the frequency of term t in the $X\&Y$ cohort (e.g., IPV&Preg) increases in relation to the frequencies

of t in the X (e.g., IPV) and Y (e.g., Preg) cohorts. Similarly, $\mathrm{C_{BG}}(X,Y)$ goes up as the size of $X\&Y$ cohort increases in relation to the sizes of the X and Y cohorts.

We compute 95% augmented confidence intervals for co-occurrence scores $\mathrm{LOR}(X,Y|\mathrm{BG})$ and $\mathrm{LOR}(X,Y|t)$ in a similar manner as we do for prevalence score $\mathrm{LOR}(t,X|\mathrm{BG})$. We replace the contingency table items $N(\mathbf{t},\mathbf{X})$ in Equation 2 with appropriate contingency table items: $N(\mathbf{X},\mathbf{Y})$ for $\mathrm{LOR}(X,Y|\mathrm{BG})$ and with $N(t,\mathbf{X},\mathbf{Y})$ for $\mathrm{LOR}(X,Y|t)$. Then, we repeat the remaining steps described in *Accounting for variance* and *Accounting for measurement error* sections.

2.3.6 Identification of synergistic terms

Here, our aim is to identify *synergistic* terms whose presence increase the co-occurrence of X and Y (specifically IPV and pregnancy), i.e., terms that make X and Y more likely to occur together. To this end, we first compute the overall co-occurrence score $\mathrm{C_{BG}}(X,Y)$ and a conditional co-occurrence score $\mathrm{C}_t(X,Y)$ for each term t. We are interested in terms that have co-occurrence scores higher than the background co-occurrence $\mathrm{C_{BG}}(X,Y)$. Thus, we utilize the following synergy score $\mathrm{S}(t)$ for a term t, which is adjusted for overall co-occurrence of X and Y conditions:

$$\begin{aligned}\mathrm{S}(t) &= \mathrm{C}_t(X,Y) - \mathrm{C_{BG}}(X,Y)\\ &= \mathrm{LOR}(X,Y|t) - \mathrm{LOR}(X,Y|\mathrm{BG})\end{aligned} \tag{8}$$

Note that, $\mathrm{C_{BG}}(X,Y)$ does not depend on a term. Thus, rankings of the terms are the same for $\mathrm{C}_t(X,Y)$ and $\mathrm{S}(t)$.

We compute the 95% augmented confidence interval for $\mathrm{S}(t)$ using the confidence intervals of co-occurrences scores as follows:

$$\begin{aligned}\mathrm{S_{min}}(t) &= \mathrm{LOR_{min}}(X,Y|t) - \mathrm{LOR_{max}}(X,Y|\mathrm{BG})\\ \mathrm{S_{max}}(t) &= \mathrm{LOR_{max}}(X,Y|t) - \mathrm{LOR_{min}}(X,Y|\mathrm{BG})\end{aligned} \tag{9}$$

RESULTS

2.4 Assessment of Prevalence

We analyze a total of 18 863 diagnostic terms. For each term t among all 18 863 terms, we compute a prevalence score $\mathrm{P}_X(t)$ (Equation 1) for each cohort X which is IPV, Preg or one of the control groups (MC, Endo, IUD, CKD). To assess the statistical significance of these scores, we compute 95% augmented confidence intervals (Equation 3). We consider a prevalence score for a term t *valid* if the corresponding augmented confidence interval has a finite range i.e., when the term frequency is non-zero. There are a total of 5 464 and 15 445 valid terms for IPV and Preg respectively.

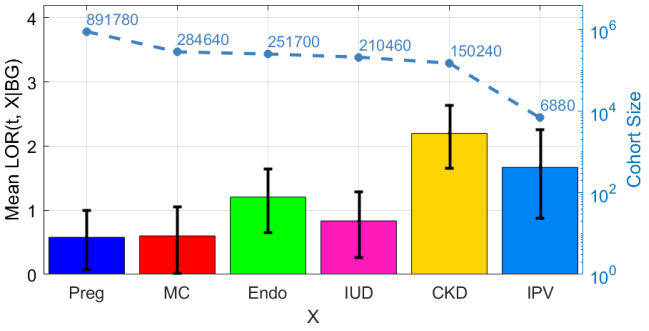

Fig. 2 Cohort sizes and mean prevalence score (log-odds ratio) for IPV, Preg and all 4 control groups. The black lines indicate the 95% augmented confidence interval for the mean prevalence of all terms in the corresponding cohort. The blue line and numbers of top indicate the cohort size. The cohorts are sorted in descending order according to the cohort sizes.

2.4.1 Estimation of the bias in reporting frequencies

We hypothesize that there may be a bias in reporting frequencies in cohorts describing serious conditions like IPV due to more meticulous record keeping. To estimate the effect of such a bias in reporting frequencies, we measure the mean prevalence score of all valid terms for each cohort (including the control cohorts). As seen in Fig. 2, the mean prevalence score for each cohort is positive and the two cohorts with smallest cohort size (CKD and IPV) have the highest mean prevalence scores. We argue that the dependence between the cohort size and this over-representation in mean prevalence scores may be because of the invalid (censored) terms with zero frequencies. That is, some rare terms that are negatively associated with IPV are simply not observed (zero number of records) in our sample population, thus, are removed from the analysis. As it can be seen from Fig. 3, it is not possible to observe rare terms (with less than 10 000 records in the background cohort) that are negatively associated with IPV. Whereas, this effect is much weaker in the Preg cohort (only very rare terms with less than 100 records in background are not observed as negatively associated with Preg) because of its higher cohort size.

Overall, from our analysis on the mean prevalence scores, we estimate that the reporting frequency of a term is roughly 3.2 times as high in the IPV cohort compared to background (95% confidence interval for odds ratio: [1.84, 4.79]) and roughly 1.5 times as high in the Preg cohort compared to background (95% confidence interval for odds ratio: [1.05, 2.00]).

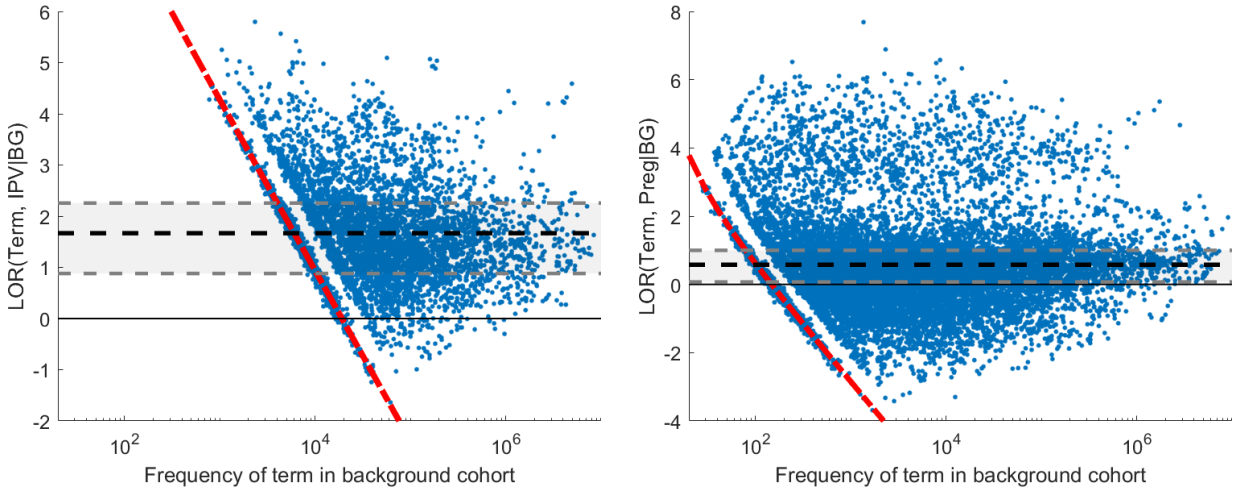

Fig. 3 Distribution of prevalence scores across all terms. The distribution of the terms' (Left) IPV prevalence score: LOR(t, IPV—BG), and (Right) Preg prevalence score: LOR(t, Preg—BG) are shown with respect to the frequency of the term in the background cohort (i.e., rarity of the term) in log-scale. The black dashed line is the mean log-odds ratio (LOR) and grayed area is the corresponding augmented confidence interval. The red line indicates the minimum observeable log-odds ratio (for a frequency of 10) given the corresponding rarity level of a term. For each term (point), a small jitter is added to help see the overlapping points.

2.4.2 Adjusting for the bias in reporting frequencies

To avoid over-estimating the prevalence scores for a cohort of interest due to a bias on reporting frequencies, we focus on terms that are more prevalent than the average prevalence of that cohort. Thus, we assign a confidence level for each computed prevalence score $\mathrm{P}_X(t)$ for cohort X and term t by applying a cohort-specific threshold:

- $\mathrm{P}_X(t)$ has "high" confidence level if the minimum bound of its confidence interval is higher than the maximum bound of mean prevalence score for cohort X.
- $\mathrm{P}_X(t)$ has "medium" confidence level if the minimum bound of its confidence interval is higher than mean prevalence score for cohort X.
- $\mathrm{P}_X(t)$ has "low" confidence level if it does not have a high or medium confidence level.

2.4.3 Prevalent terms in IPV and Preg cohorts

For the IPV cohort, we identify 668 and 611 terms with respectively high and medium confidence among 5 464 valid terms. For the Preg cohort, there are 2750 and 2024 terms with respectively high and medium confidence among 15 445 valid terms. In Supplementary Data 2, we provide terms identified as prevalent for IPV, Preg, and control cohorts.

Next, we analyze terms that are prevalent in both IPV and Preg cohorts. For this purpose, we compute a 95% augmented confidence interval for the combined prevalence score $\mathrm{P}(t)$ for each term t (Equation 4, Equation 5). For each term t, we assign the following confidence levels to their combined prevalence score:

- $\mathrm{P}(t)$ has "high" confidence level if both $\mathrm{P}_{\mathrm{IPV}}(t)$ and $\mathrm{P}_{\mathrm{Preg}}(t)$ have high confidence levels.
- $\mathrm{P}(t)$ has "medium" confidence level if both $\mathrm{P}_{\mathrm{IPV}}(t)$ and $\mathrm{P}_{\mathrm{Preg}}(t)$ have at least medium confidence levels.
- $\mathrm{P}(t)$ has "low" confidence level if $\mathrm{P}_{\mathrm{IPV}}(t)$ or $\mathrm{P}_{\mathrm{Preg}}(t)$ has a low confidence level.

All valid terms in the IPV cohort are also valid for the Preg cohort, thus we compute confidence intervals for combined prevalence scores of 5 464 terms.

Among these terms, 279 are prevalent in *both* IPV and Preg cohorts with high confidence (see Fig. 4 for a distribution of high and medium confidence terms). We sort the terms with respect to the lower bound of the combined prevalence score $\mathrm{P}(t)$ (Equation 5) and report the top 20 terms in Table 2. We provide the combined prevalence scores for all identified terms in Supplementary Data 3.

We repeat the prevalence analysis for control cohorts. For comparison, the number of terms that exhibit significant (with high and medium confidence) combined prevalence in two cohorts for all pairs of cohorts is shown in Fig. 5. As expected, for control cohorts that are directly related to pregnancy, namely intra-uterine device (IUD) and miscarriage (MC), the number of prevalent terms identified with high confidence is much higher than it is for other cohorts. Furthermore, among all analyzed cohorts, pregnancy shares the largest number of high-confidence prevalent terms with IPV. We report

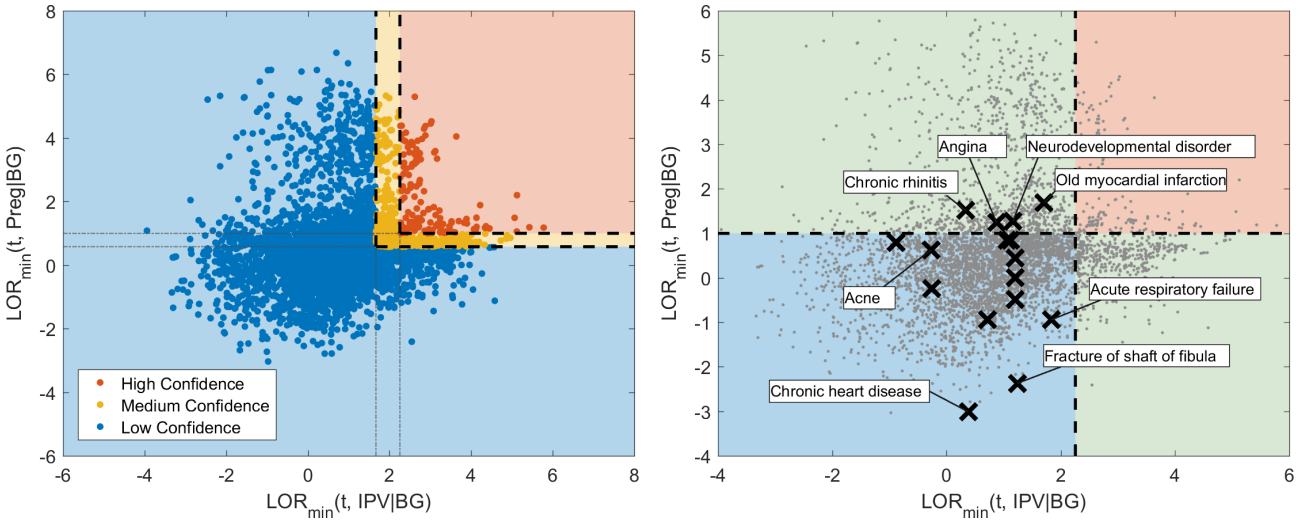

Fig. 4 Distribution of prevalence scores for high confidence prevalent and synergistic terms. X-axis and Y-axis indicate the minimum bounds of 95% augmented confidence intervals for IPV prevalence score: LOR(t, IPV—BG), and pregnancy prevalence score: LOR(t, Preg—BG) respectively. (Left) High and medium confidence prevalent terms are shown in red and yellow regions respectively. (Right) The IPV and pregnancy prevalence score distributions of high confidence synergistic terms are shown. Red region indicates high confidence prevalent terms, and green region indicates terms that are prevalent with high confidence in IPV or Preg cohorts (but not both).

Table 2 Top 20 terms identified with high confidence as prevalent in both IPV and Preg cohorts. Within square brackets, the 95% augmented confidence intervals of the corresponding log-odds ratios (LOR) and prevalence scores P(t) are provided. The terms are sorted according the minimum bound of prevalence score P(t). All 20 reported terms are identified with high confidence (higher than average prevalence in both IPV and Preg cohorts in a statistically significant manner). See Supplementary Data 3 for a complete list of terms identified.

	Term Description	LOR(Term, X|BG)		Number of records			Prevalence
		Preg	IPV	BG	Preg	IPV	P(t)
1	TID11770: Mental disorder during pregnancy - baby not yet delivered	4.07 [4.05, 4.10]	3.82 [3.64, 4.00]	50170	27230	350	3.82 [3.64, 4.00]
2	TID11772: Mental disorder in mother complicating pregnancy	4.07 [4.05, 4.10]	3.82 [3.64, 4.00]	50170	27230	350	3.82 [3.64, 4.00]
3	TID11773: Mental disorder in the puerperium - baby delivered	3.37 [3.35, 3.40]	3.36 [3.17, 3.55]	58440	24720	300	3.36 [3.17, 3.40]
4	TID7481: Fetal or neonatal effect of maternal use of tobacco	3.60 [3.58, 3.62]	3.27 [3.12, 3.42]	108830	49610	510	3.27 [3.12, 3.42]
5	TID7479: Fetal or neonatal effect of maternal transmission of substance	3.56 [3.54, 3.58]	3.24 [3.10, 3.39]	113140	50730	520	3.24 [3.10, 3.39]
6	TID11599: Maternal drug exposure	4.56 [4.52, 4.60]	3.36 [3.03, 3.68]	22860	14360	120	3.36 [3.03, 3.68]
7	TID6700: Drug dependence during pregnancy - baby not yet delivered	4.51 [4.45, 4.58]	3.52 [3.01, 4.00]	10120	6300	60	3.52 [3.01, 4.00]
8	TID6701: Drug dependence during pregnancy, childbirth and the puerperium	4.42 [4.37, 4.46]	3.28 [2.88, 3.65]	18080	10950	90	3.28 [2.88, 3.65]
9	TID11777: Mental disorders during pregnancy, childbirth and the puerperium	4.09 [4.08, 4.11]	3.01 [2.88, 3.14]	168840	89340	650	3.01 [2.88, 3.14]
10	TID6702: Drug dependence in mother complicating pregnancy, childbirth AND/OR puerperium	4.34 [4.29, 4.40]	3.31 [2.85, 3.74]	13690	8130	70	3.31 [2.85, 3.74]
11	TID11769: Mental disorder during pregnancy - baby delivered	3.99 [3.97, 4.00]	2.96 [2.80, 3.13]	106540	55440	410	2.96 [2.80, 3.13]
12	TID719: Acute gonococcal cervicitis	3.45 [3.33, 3.58]	3.92 [2.77, 4.82]	2550	1130	20	3.45 [2.77, 3.58]
13	TID11884: Mild hyperemesis-not delivered	3.54 [3.52, 3.56]	2.94 [2.74, 3.13]	78120	35040	300	2.94 [2.74, 3.13]
14	TID6699: Drug dependence during pregnancy - baby delivered	4.00 [3.91, 4.09]	3.53 [2.69, 4.23]	5030	2700	30	3.53 [2.69, 4.09]
15	TID18493: Urogenital infection caused by Trichomonas vaginalis	2.70 [2.68, 2.72]	2.95 [2.74, 3.15]	69440	21950	270	2.70 [2.68, 2.72]
16	TID605: Acute cervicitis	3.37 [3.24, 3.49]	3.82 [2.67, 4.72]	2730	1170	20	3.37 [2.67, 3.49]
17	TID724: Acute gonorrhea of lower genitourinary tract	3.60 [3.57, 3.64]	3.01 [2.66, 3.35]	26550	12370	110	3.01 [2.66, 3.35]
18	TID8252: Gonorrhea	3.51 [3.47, 3.54]	2.98 [2.66, 3.28]	32260	14480	130	2.98 [2.66, 3.28]
19	TID8156: Genitourinary tract infection in pregnancy - not delivered	3.65 [3.64, 3.67]	2.81 [2.66, 2.95]	149320	68580	510	2.81 [2.66, 2.95]
20	TID10721: Late pregnancy vomiting - not delivered	3.35 [3.31, 3.39]	3.01 [2.66, 3.34]	26640	11270	110	3.01 [2.66, 3.34]

the prevalence scores and the confidence levels for the combinations of IPV and Pregnancy with each of the control cohorts (MC, Endo, CKD, IUD) in Supplementary Data 4.

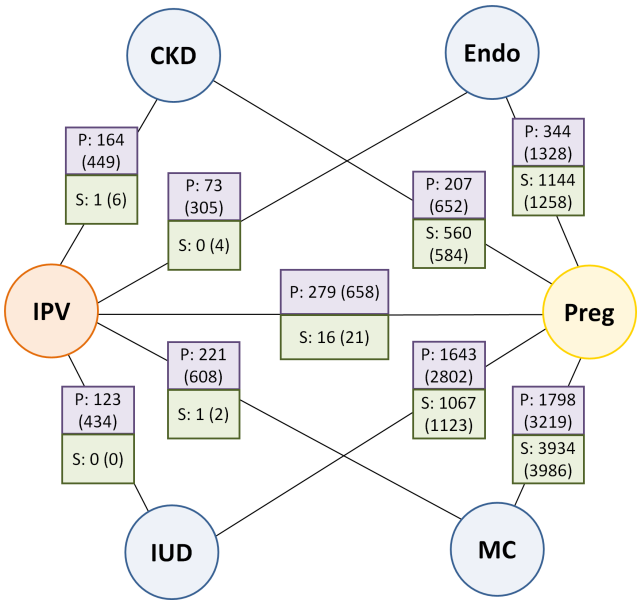

Fig. 5 Number of prevalent (P) and synergistic (S) terms identified with high (at least medium) confidence between IPV, Preg and each of the 4 control cohorts. The number of terms identified with high or medium confidence level are given in parentheses.

2.5 Assessment of Co-occurrence

To study our second research question, we assess the co-occurrence between IPV and pregnancy. Namely, we investigate if the frequency of IPV among pregnant women is higher than the frequency of IPV among women who are not pregnant. For this purpose, we compute an overall co-occurrence score $\mathrm{C}_{\mathrm{BG}}(\mathrm{IPV}, \mathrm{Preg})$ (Equation 6) as well as a conditional co-occurrence score $\mathrm{C}_t(\mathrm{IPV}, \mathrm{Preg})$ (Equation 7) for each term t. Similar to the prevalence analysis, we assess the statistical significance of these scores by computing 95% augmented confidence intervals. There are 2818 terms that have a valid co-occurrence score (which requires a non-zero frequency in the IPV&Preg cohort).

The overall co-occurrence score C_{BG} between IPV and pregnancy is 1.50 [1.40, 1.60]. Similarly, the mean co-occurrence score $\mathrm{C}_t(\mathrm{IPV}, \mathrm{Preg})$ of IPV and pregnancy is 1.26 [0.16, 2.22]. Thus, both numbers suggest that there is overall a significant positive co-occurrence between IPV and pregnancy: With 95% confidence, we estimate that the odds (risk) of IPV among pregnant women is between 2.4 ($2^{1.26}$) to 2.8 ($2^{1.5}$) times more than the risk of IPV among women who are not pregnant. However, this observation may be affected by selection bias: During pregnancy, women interact with the healthcare system more frequently. Therefore, IPV may be documented more frequently among pregnant women.

2.6 Assessment of Synergy

To further elucidate the association between pregnancy and IPV in terms of their health correlates, we identify *synergistic* terms that indicate higher conditional co-occurrence between IPV and pregnancy than the overall co-occurrence in the general population (measured 1.50 [1.40, 1.60]). Thus, for each term t, we compute a synergy score $\mathrm{S}(t) = \mathrm{C}_t(\mathrm{IPV}, \mathrm{Preg}) - \mathrm{C}_{\mathrm{BG}}(\mathrm{IPV}, \mathrm{Preg})$ and assign a confidence level for the synergy of each term as follows:

- Synergy $\mathrm{S}(t)$ has "high" confidence level if the minimum bound of the confidence interval for conditional co-occurrence score $\mathrm{C}_t(\mathrm{IPV}, \mathrm{Preg})$ is higher than the maximum bound of the overall co-occurrence score $C_{\mathrm{BG}}(\mathrm{IPV}, \mathrm{Preg})$ (i.e., 1.60).
- Synergy $\mathrm{S}(t)$ has "medium" confidence level if the minimum bound of the confidence interval for conditional co-occurrence score $\mathrm{C}_t(\mathrm{IPV}, \mathrm{Preg})$ is higher than the point estimate of the overall co-occurrence score $\mathrm{C}_{\mathrm{BG}}(\mathrm{IPV}, \mathrm{Preg})$ (i.e., 1.50).
- Synergy $\mathrm{S}(t)$ has "low" confidence level if it does not have high or medium confidence level.

We identify 16 and 5 terms that are synergistically associated with IPV and pregnancy with respectively high and medium confidence . These terms are shown on Table 3. We provide the synergy and co-occurrence analysis results of all valid terms in Supplementary Data 5. We also visualize the distribution of IPV and pregnancy prevalence scores for the 16 high-confidence synergistic terms in Fig. 4. As seen in the figure, most of the terms that are identified as synergistic with the co-occurrence of IPV and pregnancy are not significantly prevalent in the IPV or pregnancy cohorts.

Next, we repeat the synergy analysis for all pairs of control cohorts with IPV and pregnancy (e.g., IPV-CKD). The number of high-confidence and medium-confidence synergistic terms for all pairs of cohorts involving IPV or pregnancy are shown in Fig. 5. We provide the results of synergy and co-occurrence analyses for control cohorts in Supplementary Data 6.

3 Discussion

3.1 Commonly Prevalent Terms

Three key themes emerge from the collection of commonly prevalent terms: (i) poor mental health during pregnancy, (ii) maternal substance use, and (iii) genitourinary infections and complications.

Poor Mental Health During Pregnancy. Our first finding that poor mental health is significant in both

Table 3 Top 21 terms identified as synergistic for IPV and Preg cohorts with high or medium confidence. Within square brackets, the 95% augmented confidence intervals of the corresponding log-odds ratio (LOR) and synergy scores S(t) are provided. The terms are sorted according to the minimum bound of the synergy score S(t). See Supplementary Data 5 for the synergy analysis results of all terms. *Both refers to the number of records diagnosed with the term in IPV&Preg cohort.

	Term Description	LOR(Preg, IPV|Term)	Number of Records			Synergy	Conf.
			Preg	IPV	Both*	S(t)	Level
1	TID398: Acne	2.36 [1.83, 2.90]	56950	250	100	0.87 [0.23, 1.50]	High
2	TID3725: Chronic heart disease	3.21 [1.80, 4.44]	3040	90	20	1.72 [0.21, 3.04]	High
3	TID5759: Developmental disorder of motor function	2.19 [1.79, 2.58]	32300	420	140	0.69 [0.19, 1.18]	High
4	TID2005: Attention deficit hyperactivity disorder	2.19 [1.79, 2.58]	32260	420	140	0.69 [0.19, 1.18]	High
5	TID6562: Disorders of attention and motor control	2.19 [1.79, 2.58]	32260	420	140	0.69 [0.19, 1.18]	High
6	TID7875: Fracture of shaft of fibula	3.69 [1.78, 6.08]	1490	40	20	2.19 [0.18, 4.68]	High
7	TID13221: Old myocardial infarction	3.08 [1.76, 4.17]	1330	140	20	1.58 [0.16, 2.77]	High
8	TID5758: Developmental disorder	2.08 [1.73, 2.43]	45010	510	170	0.59 [0.13, 1.04]	High
9	TID12790: Neurodevelopmental disorder	2.10 [1.73, 2.48]	37240	470	150	0.61 [0.13, 1.08]	High
10	TID6425: Disorder of sebaceous gland	2.16 [1.69, 2.63]	67240	320	110	0.66 [0.09, 1.23]	High
11	TID5427: Current knee cartilage tear	2.74 [1.69, 3.71]	5350	120	30	1.24 [0.09, 2.31]	High
12	TID12808: Neurologic disorder associated with type II diabetes mellitus	3.07 [1.68, 4.25]	2100	100	20	1.57 [0.09, 2.85]	High
13	TID5985: Disease of possible viral origin	2.72 [1.67, 3.83]	26110	90	40	1.23 [0.07, 2.43]	High
14	TID3882: Chronic rhinitis	2.48 [1.64, 3.32]	16080	130	50	0.99 [0.05, 1.93]	High
15	TID930: Acute respiratory failure	2.42 [1.63, 3.15]	4150	220	40	0.92 [0.03, 1.76]	High
16	TID1525: Angina	2.98 [1.62, 4.13]	2510	110	20	1.48 [0.03, 2.73]	High
17	TID7148: Essential hypertension	1.80 [1.57, 2.02]	99210	1850	260	0.30 [-0.03, 0.62]	Medium
18	TID5765: Developmental mental disorder	2.82 [1.54, 4.16]	6240	70	30	1.32 [-0.06, 2.77]	Medium
19	TID1967: Atrial fibrillation	2.90 [1.52, 4.09]	3080	100	20	1.41 [-0.07, 2.69]	Medium
20	TID3497: Child attention deficit disorder	2.21 [1.51, 2.89]	15900	180	60	0.71 [-0.09, 1.50]	Medium
21	TID7514: Fibrillation	2.88 [1.50, 4.06]	3240	100	20	1.38 [-0.10, 2.66]	Medium

Pregnancy and IPV populations is suggestive of the increased incidence of mental health issues in each of these cohorts independently. Depression and post-traumatic stress disorder (PTSD) are the most common mental disorders that affect women who experience IPV during pregnancy [11, 19]. Compared to their non-abused counterparts, pregnant mothers who experience IPV have about four times higher odds of experiencing antenatal and seven times higher odds of experiencing postnatal depression [10]. A similar pattern is reported in the prevalence of PTSD, which sees an increase of up to four-fold in IPV and pregnancy, compared to pregnancy alone [12, 19].

Maternal Substance Use. Our second finding is that maternal substance use is significantly prevalent in both Pregnancy and IPV populations. Substance use during pregnancy spells harmful consequences for both the mother and child. In the US, less than 5% of pregnant women have been reported using drugs during pregnancy [26]. Due to the stigma and risks associated with self -reporting substance use and IPV, an estimated prevalence of substance use in the pregnant and IPV population is difficult to ascertain; however, studies confirm that IPV significantly increases drinking, smoking and substance abuse behaviors in pregnant women [16]. These studies also suggest that chronic stress and substance abuse as a stress-coping mechanism that predisposes vulnerable women to drug dependence [16].

Genitourinary Infections and Complications: Our third finding is that genitourinary complication is significantly prevalent in both Pregnancy and IPV populations. These complications include the presence of unspecified cervical and uterine inflammatory disease, as well as genitourinary inflammation due to sexually transmitted infections (STIs), notably gonorrhea and trichomoniasis. This association becomes especially dangerous in the context of pregnancy with IPV because of the increased risk of pregnancy complication that brings harm to the mother and child. Maternal infection with gonorrhea and/or trichomoniasis specifically predisposes the child to low birth weight and preterm birth complications, both of which spell adverse long-term health complications for the child [15].

3.2 Synergistic Health Correlates

We define the synergistic health correlates as the terms whose appearance in a record significantly increases the likelihood that pregnancy and IPV will occur together in that record. By conservatively assessing significance based on the lower end of the confidence interval for the odds ratio, we identify 21 (16 high confidence, 5 medium confidence) synergistic terms. The observed synergy for these terms can be due to various reasons, including the following:

1. The term is a direct consequence of trauma caused by IPV in a pregnant woman.
2. The term is a risk factor that increases the vulnerability of a woman to IPV during pregnancy.
3. The term is more prevalent in the demographics (particularly young women) on which IPV during pregnancy is observed more frequently as compared to other demographics.
4. The term is associated with the occurrence of IPV during pregnancy potentially as a long-term consequence.
5. The term is synergistically related to IPV and pregnancy (i.e., either falls into one of the above categories or is otherwise related), but this link is not previously reported.
6. The term is a false positive i.e., the observed synergy between the term and IPV/pregnancy is a statistical artefact.

Note that, with the observational data that is utilized in this study, it is not possible to distinguish the terms that are in categories 5 and 6. To understand whether a term is a previously undiscovered synergistic factor or a false positive, additional analyses using more detailed (patient-specific and/or time-course) data are required. Thus the results we provide here can be used to seed detailed analyses utilizing additional data.

While it is also not possible to conclusively categorize the identified synergistic terms into one of the first four categories, many of the terms we identify can be putatively interpreted as belonging to one of these categories. Interestingly, the 21 terms we identify include representatives that potentially belong to each of these categories:

Category 1: Among the conditions identified by the aggregation of synergistically prevalent terms, acute musculoskeletal injuries (i.e. tibial fracture and meniscal tear) can be explained by patterns of mechanical trauma found in IPV [13, 20]. The remaining conditions—integumentary, neurological, cardiovascular, pulmonary, and immunological in origin—have a more multifactorial etiology.

Category 2: The synergistic terms that can be considered in this category include: developmental disorder of motor function, attention deficit hyperactivity disorder (ADHD), disorders of attention and motor control, developmental disorder, and neuro-developmental disorder (observe that the terms are also quite variable in their specificity). Our findings are corroborated by the literature, which identify childhood ADHD as a significant predictor of IPV, and find that persistence of ADHD into adulthood compounds the risk of IPV victimization [4, 9]. Neuro-developmental and attention deficit disorders often co-occur with motor disorders; motor dysfunction results from impaired neuro-muscular maturation and motor planning mechanisms [7, 24]. The strong association between neuro-developmental disorders and IPV might be attributed to the increased vulnerability experienced due to the presence of these unique motor and neurological deficits.

Category 3: Some terms may be identified as synergistic due to demography-associated factors that are related to both IPV and pregnancy. For example, acne and problems with sebaceous glands can be considered in this category. Although our population does not include teenagers, many young adults might still have lingering problems with their skins. Skin problems can also be due to stress, bacteria, hormones, medication and genetics [6, 27]. It is also important to note that acne and its etiopathogenesis can uniquely be linked to anxiety and depression, common comorbidities of IPV and pregnancy [2, 22].

Category 4: Other synergistic terms can be attributed to increased risk of certain medical consequences due to exposure to IPV during pregnancy. Among these, chronic rhinitis can be associated with repeated assault to neck, face, and head (potentially due to deviation). In addition, many cardiopulmonary pathologies are identified as synergistic terms. These terms include conditions such as myocardial infarction, angina, fibrillation, acute respiratory failure, and essential hypertension. Cardiovascular disease (CVD) can be congenital or occur secondary to smoking, alcohol abuse, stress, and a number of comorbidities, such as hyperlipidemia and diabetes. Additional research also reports strong association of IPV with CVD risk factors, smoking, abdominal obesity, hyperlipidemia, and hypertension [21]. Unfortunately, there are not many studies investigating the direct link between IPV and cardiopulmonary pathologies. These conditions require more investigation to explore this link and its mechanisms to facilitate comprehensive understanding of the impact of IPV on the victims' health.

3.3 Generalizability of findings

An important step in data analysis is the validation of findings in terms of their reliability and generalizability. These are typically assessed through established strategies like cross-validation, which measure the reproducibility of findings using hold-out samples. However, since these strategies are based on separating a subset of observations for validation, they require knowledge of individual observations. Such strategies cannot be employed in this study, since we analyze data that is presented in summaries, i.e., our data consists of frequencies (number of records) of terms in a specified

population. As a result, our validation strategy is based on (i) application of conservative statistical significance levels, in combination with confidence intervals that take into account the inherent noise and systematic bias that can be detected from data, (ii) utilization of control cohorts as a comparison for women with different health concerns, including chronic and acute conditions, and (iii) incorporation of expert opinion to assess the clinical relevance of the key identified terms. From a methodological perspective, we acknowledge the limitations on assessing the generalizability of the findings based on such summary-level data. We recognize the development of alternative strategies that can function on summary-level data as an important open problem.

3.4 Limitations

Electronic health records (EHR) data utilized in this study are observational. Answering causal questions with this type of data is tricky (if not impossible) without a comprehensive understanding of possible confounding variables. Any apparent associations in the data may stem from an unobserved confounder, or simply be due to selection bias in the collection of the records. Also, due to the employed privacy protocols, we do not have access to individual record information. There could be multiple records of a single patient, which can bias our observations. Thus, with the available data, it is a challenging task to draw conclusions on the mechanisms of the observed associations.

4 Conclusions

The main research question we raised in this study was whether utilization of large-scale EHR data can provide new insights into the health correlates of IPV during pregnancy. As expected, our results demonstrated that IPV significantly affects the well-being of the mother and can intensify complications during pregnancy. Beyond these expected results, our proposed notion of synergy led to the identification of many significant terms that can be potentially related to the occurrence of IPV during pregnancy. By interpreting these terms in the light of the literature, we categorized these terms in terms of their potential relationship with the interplay between IPV and pregnancy. These categories included:

- direct consequences of trauma
- risk factors for exposure to IPV during pregnancy (in particular neuro-developmental disorders)
- markers for the demographics that are at risk (e.g., acne, indicating that young women can be particularly at risk of IPV during pregnancy)
- long-term consequences of IPV during pregnancy.

It was particularly striking that representatives from each of these categories were present among the terms that were found to be synergistically associated with IPV and pregnancy. Further investigation of these findings on more detailed EHR data can further elucidate the interplay between these conditions and IPV during pregnancy. After further validation, the terms that we identify can serve as potential markers of the presence of IPV among pregnant women.

From a methodological perspective, our results clearly demonstrated that EHR data contains information that can provide new insights into the relationships between different health conditions. The sets of prevalent and synergistic terms we identified were highly parsimonious and most of these terms were directly related to the relationship between IPV and pregnancy. The terms that were not known to be associated with the relationship between IPV and pregnancy revealed indirect relationships that can also provide insights into markers, risk factors, and potential consequences of IPV during pregnancy. These results showed that carefully designed data analysis techniques can reliably extract "hidden" patterns despite the limitations of EHR data. The methodology presented here will also be useful in the investigation of the synergy between other medical conditions using EHR databases with privacy constraints.

5 Supplementary Materials

The supplementary data are available at:
`github.com/serhan-yilmaz/IPV_pregnant_women`

Acknowledgements This work is supported by the National Institutes of Health (NIH) grant R01-LM012518 from the National Library of Medicine.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Announcement, I.U.S.S.: Ibm explorys cohort discovery (2016). 216-191
2. Bagatin, E., Freitas, T.H.P.d., Machado, M.C.R., Ribeiro, B.M., Nunes, S., Rocha, M.A.D.d.: Adult female acne: a guide to clinical practice. Anais brasileiros de dermatologia **94**(1), 62–75 (2019)

3. Breiding, M., Basile, K.C., Smith, S.G., Black, M.C., Mahendra, R.R.: Intimate partner violence surveillance: uniform definitions and recommended data elements. version 2.0 (2015)
4. Buitelaar, N.J., Posthumus, J.A., Buitelaar, J.K.: Adhd in childhood and/or adulthood as a risk factor for domestic violence or intimate partner violence: a systematic review. Journal of attention disorders p. 1087054715587099 (2015)
5. Chisholm, C.A., Bullock, L., Ferguson II, J.E.J.: Intimate partner violence and pregnancy: epidemiology and impact. American journal of obstetrics and gynecology **217**(2), 141–144 (2017)
6. Clayton, R., Göbel, K., Niessen, C., Paus, R., van Steensel, M., Lim, X.: Homeostasis of the sebaceous gland and mechanisms of acne pathogenesis. British Journal of Dermatology (2019)
7. Dahan, A., Reiner, M.: Evidence for deficient motor planning in adhd. Scientific reports **7**(1), 9631 (2017)
8. García-Moreno, C., Pallitto, C., Devries, K., Stöckl, H., Watts, C., Abrahams, N.: Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence. World Health Organization (2013)
9. Guendelman, M.D., Ahmad, S., Meza, J.I., Owens, E.B., Hinshaw, S.P.: Childhood attention-deficit/hyperactivity disorder predicts intimate partner victimization in young women. Journal of abnormal child psychology **44**(1), 155–166 (2016)
10. Halim, N., Beard, J., Mesic, A., Patel, A., Henderson, D., Hibberd, P.: Intimate partner violence during pregnancy and perinatal mental disorders in low and lower middle income countries: a systematic review of literature, 1990–2017. Clinical psychology review **66**, 117–135 (2018)
11. Kessler, R.C., Angermeyer, M., Anthony, J.C., De Graaf, R., Demyttenaere, K., Gasquet, I., De Girolamo, G., Gluzman, S., Gureje, O., Haro, J.M., et al.: Lifetime prevalence and age-of-onset distributions of mental disorders in the world health organization's world mental health survey initiative. World psychiatry **6**(3), 168 (2007)
12. Khoramroudi, R.: The prevalence of posttraumatic stress disorder during pregnancy and postpartum period. Journal of family medicine and primary care **7**(1), 220 (2018)
13. Kivelä, S., Leppäkoski, T., Ruohoniemi, J., Puolijoki, H., Paavilainen, E.: The documentation and characteristics of hospitalized ipv patients using electronic medical records data: a follow-up descriptive study. Journal of Family Violence pp. 1–9 (2019)
14. Martin, S.L., Mackie, L., Kupper, L.L., Buescher, P.A., Moracco, K.E.: Physical abuse of women before, during, and after pregnancy. Jama **285**(12), 1581–1584 (2001)
15. Mullick, S., Watson-Jones, D., Beksinska, M., Mabey, D.: Sexually transmitted infections in pregnancy: prevalence, impact on pregnancy outcomes, and approach to treatment in developing countries. Sexually transmitted infections **81**(4), 294–302 (2005)
16. Organization, W.H., et al.: Intimate partner violence during pregnancy: Information sheet. Tech. rep., Geneva: World Health Organization (2011)
17. Organization, W.H., et al.: Understanding and addressing violence against women: Intimate partner violence. Tech. rep., World Health Organization (2012)
18. Platform, T.I.E.: Solution brief (2016)
19. Rose, L., Alhusen, J., Bhandari, S., Soeken, K., Marcantonio, K., Bullock, L., Sharps, P.: Impact of intimate partner violence on pregnant women's mental health: Mental distress and mental strength. Issues in Mental Health Nursing **31**(2), 103–111 (2010)
20. Sheridan, D.J., Nash, K.R.: Acute injury patterns of intimate partner violence victims. Trauma, Violence, & Abuse **8**(3), 281–289 (2007)
21. Stene, L.E., Jacobsen, G.W., Dyb, G., Tverdal, A., Schei, B.: Intimate partner violence and cardiovascular risk in women: a population-based cohort study. Journal of women's health **22**(3), 250–258 (2013)
22. Stewart, D.E., Vigod, S.N.: Mental health aspects of intimate partner violence. Psychiatric Clinics **40**(2), 321–334 (2017)
23. Truman, J.L., Morgan, R.E.: Nonfatal domestic violence, 2003-2012. Journal of Current Issues in Crime, Law & Law Enforcement **8**(4) (2015)
24. Udo, I.E., Sharps, P., Bronner, Y., Hossain, M.B.: Maternal intimate partner violence: Relationships with language and neurological development of infants and toddlers. Maternal and child health journal **20**(7), 1424–1431 (2016)
25. Van Parys, A.S., Verhamme, A., Temmerman, M., Verstraelen, H.: Intimate partner violence and pregnancy: A systematic review of interventions. PloS one **9**(1), e85084 (2014)
26. Wendell, A.D.: Overview and epidemiology of substance abuse in pregnancy. Clinical obstetrics and gynecology **56**(1), 91–96 (2013)
27. Yim, I.S., Kofman, Y.B.: The psychobiology of stress and intimate partner violence. Psychoneuroendocrinology **105**, 9–24 (2019)