Large language model aided automatic high-throughput drug screening using self-controlled cohort study

Shenbo Xu[1], Stan N. Finkelstein[2], Roy E. Welsch[3], Kenney Ng[4], Ioanna Tzoulaki[5], and Lefkos Middleton[6]

[1]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02142, USA.
[1]Email: xushenbo@mit.edu

August 4, 2024

Abstract:

Background: Developing medicine from scratch to governmental authorization and detecting adverse drug reactions (ADR) have barely been economical, expeditious, and risk-averse investments. The availability of large-scale observational healthcare databases and the popularity of large language models offer an unparalleled opportunity to enable automatic high-throughput drug screening for both repurposing and pharmacovigilance.

Objectives: To demonstrate a general workflow for automatic high-throughput drug screening with the following advantages: (i) the association of various exposure on diseases can be estimated; (ii) both repurposing and pharmacovigilance are integrated; (iii) accurate exposure length for each prescription is parsed from clinical texts; (iv) intrinsic relationship between drugs and diseases are removed jointly by bioinformatic mapping and large language model - ChatGPT; (v) causal-wise interpretations for incidence rate contrasts are provided.

Methods: Using a self-controlled cohort study design where subjects serve as their own control group, we tested the intention-to-treat association between medications on the incidence of diseases. Exposure length for each prescription is determined by parsing common dosages in English free text into a structured format. Exposure period starts from initial prescription to treatment discontinuation. A same exposure length preceding initial treatment is the control period. Clinical outcomes and categories are identified using existing phenotyping algorithms. Incident rate ratios (IRR) are tested using uniformly most powerful (UMP) unbiased tests.

Results: We assessed 3,444 medications on 276 diseases on 6,613,198 patients from the Clinical Practice Research Datalink (CPRD), an UK primary care electronic health records (EHR) spanning from 1987 to 2018. Due to the built-in selection bias of self-controlled cohort studies, ingredients-disease pairs confounded by deterministic medical relationships are removed by existing map from RxNorm and nonexistent maps by calling ChatGPT. A total of 16,901 drug-disease pairs reveals significant risk reduction, which

can be considered as candidates for repurposing, while a total of 11,089 pairs showed significant risk increase, where drug safety might be of a concern instead.

Conclusions: This work developed a data-driven, nonparametric, hypothesis generating, and automatic high-throughput workflow, which reveals the potential of natural language processing in pharmacoepidemiology. We demonstrate the paradigm to a large observational health dataset to help discover potential novel therapies and adverse drug effects. The framework of this study can be extended to other observational medical databases.

Keywords: drug screening, drug repurposing, pharmacovigilance, natural language processing, self-controlled cohort study, incidence rate ratio.

1 Introduction

Currently approved treatment options for many diseases are limited, including cancer, Alzheimer's disease, Parkinson's disease, depressive and bipolar disorders, and HIV etc. As a consequence, there is a considerable unmet demand for disease-modifying medications for many disorders. Evolution of therapeutics for disease modification has not been successful due to restricted drug targets, high-expense, and time-consuming experiments. Meanwhile, there are more than 3,000 medications currently being prescribed in the UK, with many employed for specific indications. Extensive opportunities can be taken to repurpose existing medications for new indications, along with possible sequences of treatments for various groups of disorders.

Aside from discovering new uses, existing drugs must be overseen during their whole circulation for adverse drug reactions (ADR), also known as side-effects, and other unintended consequences. In the UK, ADRs may account for 5-8% of impromptu hospitalizations that result in 4-6% hospital beds filled, with an approximate annual bill of £1-2.5bn for the National Health Service (NHS) (Jordan et al., 2018). Before supervisory authorization, randomized clinical trials (RCTs), often considered as the gold standard for causal effects, serve as fundamental sources for ADRs. Though drugs have rigorous preapproval procedures in the UK, many ADRs with low incidence rates may not be discovered during clinical trials due to limited sample size with several thousand people, insufficient follow-up, highly-stratified population without certain diseases and/or history of drug classes, restricting causal effects to a subgroup with poor coincidence in real-world clinical scenarios. Therefore, only a partial list of undesired effects are known when new drugs are approved.

However, unknown ADRs are likely to appear and evolve into a threat to public health after releasing new drugs to the market. As millions of people with heterogeneous medical history may take the same drug, post-marketing pharmacovigilance systems are considered indispensable and worthwhile to detect and report ADRs in a timely manner. Currently, ADRs are detected by spontaneous reports where patients report individual negative symptoms to health professionals. This system is quite competent for instant recurrable reactions to therapies with low intrinsic risks. Since pre-exposure cases are generally not reported, spontaneous reports are not appropriate for ADRs with high inherent rates or delayed responses (Harpaz et al., 2013). Due to limitations of RCTs and spontaneous reports, evidence from observational studies has turned out be a vital

source for post-approval drug surveillance owing to their population size, long pre- and post-exposure follow-up, and wide coverage of patients from all backgrounds.

The secondary use of longitudinal observational databases, including electronic health records (EHRs) and administrative claims, offer the possibility to characterize relationships between drugs and clinical outcomes with real-world insights (Shin et al., 2021). These types of data capture broad healthcare information including diagnoses from physicians, therapies filled for patients, lab tests, and other information, which have been actively used to conduct hypothesis-testing pharmacoepidemiology studies for causal effects of defined exposure on subsequent clinical outcomes. In recent years, there has been increasing interest in adopting these datasets to inform early drug development (Mittal et al., 2017), to identify novel treatment pathways (Yao et al., 2011), to fathom disease etiology as well as prevention, and to discover unknown benefits (Glicksberg et al., 2019) and side-effects of existing medications (Zhou et al., 2018) in a fast, large-scale data-driven, nonparametric, hypothesis generating, and high-throughput method.

Prior work employing identical study design have focused either on specific clinical outcomes (Kern et al., 2019, 2021; Teneralli et al., 2021; Kern et al., 2022) or particular drug class of interest (Cepeda et al., 2019). All these real-world applications concentrated on unacknowledged benefits of existing medications based on US administrative claims data. The other common type of observational data-electronic health records and the other direction of effects-unknown ADRs haven't been investigated.

Empirical performance has been compared with several study designs (Ryan, Stang, et al., 2013; Norén et al., 2013) and assessed as a tool for risk identification in observational healthcare data (Ryan, Schuemie, et al., 2013; Ryan and Schuemie, 2013). Due to inadequate prescription information in claims data, a fixed 30 days gap between consecutive fills were utilized to calculate length of exposure. Previous applications also require manual removal of drug confounded by indication. Above-mentioned investigations focus merely on associations without relating target quantity to causal interpretation.

The year 2023 has seen an explosive growth of artificial intelligence generated content (AIGC) (Y. Cao et al., 2023; Gozalo-Brizuela and Garrido-Merchan, 2023; Bubeck et al., 2023; OpenAI, 2023b), especially the release a powerful large language model (LLM) ChatGPT-4 created by OpenAI (2023a). Though ChatGPT is a general language model (Wolfram, 2023), its application to healthcare has been demonstrated by chatbox (Lee et al., 2023). As a disruptor to the healthcare industry (Eloundou et al., 2023), its utility on epidemiology hasn't been widely explored. We bring the power of ChatGPT into a pharmacoepidemiologic setting, aiming to remove known intertwined drug-disease pairs.

In this work, we aim to establish an automated framework to screen available drugs on possible diseases for both unknown positive and negative clinical signals with more accurate exposure length, auto-removal of drug-indication pairs, and causal-wise interpretation. Drug-disease pairs with significant risk reduction could inform potential treatment options for new indications while those pairs with significant risk increase may be monitored for drug safety.

2 Methods

2.1 Study design

We discuss the major limitation of existing pharmacoepidemiology study design for drug screening in Table 1. Depending on the target population, we can roughly categorize study designs into cohort study, case-control, spontaneous reporting systems, and other specific arrangements. Due to confounding by indication in observational health databases, the external control group in cohort studies has to have the same background conditions with the exposure group in order to ensure overlap. If this is not satisfied, researchers can barely interpret what particular contrasts are made, leading to spurious signals for drug screening. Real-world performance of cohort studies with external control group have been investigated by Ryan, Schuemie, et al. (2013); Norén et al. (2013). Since a self-controlled cohort study does not rely on any form of external control group, it becomes a natural choice for drug screening.

The second category case-based design focuses on cases and can be further divided into case-control and self-controlled case-only depending on the existence of an external control group. Case-based methods are not suitable for drug screening since case-control requires an external control group while a self-controlled case-only study incorporates solely individuals who experience the outcome of interest. Detailed comparison of self-controlled case-only designs are available in Hallas and Pottegård (2014); Takeuchi et al. (2018) and empirical behavior of case-based methods has been studied by Madigan et al. (2013); Suchard et al. (2013). The third type is a disproportionality analysis of drug-disease combinations of spontaneous reports as explained in Huang et al. (2014). As disproportionality analysis depends entirely on case counts without bringing in person time as exposure length, this method can be appropriate for spontaneous reports but not healthcare data. DuMouchel et al. (2013) demonstrated this limitation using real-world data.

Method comparison and detailed demonstration of common pharmacoepidemiology studies are available in Murphy et al. (2011); OHDSI (2020). Schuemie et al. (2012); Ryan and Schuemie (2013); Ryan, Stang, et al. (2013); Ryan et al. (2012); Reps et al. (2013) assessed various study designs for risk identification by large simulation and empirical performance. Some other methods, such as tree-based scan statistic (Kulldorff et al., 2003, 2013) and supervised learning (Reps et al., 2014, 2015), have been devised but they failed to provide risk estimates. Remaining study designs, including case-specular, case-distribution, case-control-specular, case series, case reports, ecological study, and proportional mortality study, are not suitable for pharmacoepidemiology and are not explained here.

Owing to limitations of existing pharmacoepidemiology study design, we focus on the self-controlled cohort for high-throughput drug screening (Kern et al., 2019; Cepeda et al., 2019; Teneralli et al., 2021; Kern et al., 2021, 2022). As illustrated in Figure 1, a self-controlled cohort only uses new users of the drug of interest where individuals serve as their own controls to handle confounding, by contrasting incidence rate after exposure versus before exposure. Based on simulation studies, a self-controlled cohort revealed less biased estimates with better predictive performance than other study designs (Ryan, Schuemie, et al., 2013; Ryan and Schuemie, 2013; Schuemie et al., 2013, 2020).

Table 1: Limitations of existing pharmacoepidemiology study design for drug screening

Study design	Major limitation	Reference
Cohort study		
External-controlled cohort		
New user cohort	External control	Laifenfeld et al. (2021)
Sequential statistical testing		Cook et al. (2012)
MaxSPRT	External control	Brown et al. (2007)
CSSP	External control	Li (2009)
Self-controlled cohort		
Plain		
LGPS-LEOPARD	Shrinks estimates	Norén et al. (2013)
Calibrated self-controlled cohort		
Temporal pattern discovery	External control	Norén et al. (2010)
MUTARA/HUNT	External control	Jin et al. (2006)
Fuzzy logic	External control	Ji et al. (2011)
Prior event rate ratio	External control	Yu et al. (2012)
Case-based		
Case-control		
Nested/matched case-control	External control	Ernster (1994)
Case-cohort	Random control	Breslow et al. (1982)
Case-time-control	External control	Suissa (1995)
Self-controlled case-only		
Case-crossover	Included cases only	Maclure (1991)
Self-controlled case series	Included cases only	Petersen et al. (2016)
Hawkes process modeling	Included cases only	Bao et al. (2017)
Case-case-time-control	Included cases only	Wang et al. (2011)
Sequence symmetry analysis	Included cases only	Hallas (1996)
Disproportionality analysis		
Frequentist methods		
ROR/PRR/IC/χ^2 test	Person time ignored	DuMouchel et al. (2013)
Fixed-margin volume χ^2 test	Person time ignored	H. Cao et al. (2007)
Regression	Person time ignored	Domínguez-Almendros et al. (2011)
LRT	Person time ignored	Huang et al. (2011)
Multiple drug LRT	Person time ignored	Huang et al. (2013)
Stratified LRT	Person time ignored	Nam et al. (2017)
Empirical Bayes methods		
BCPNN	Person time ignored	Bate et al. (1998)
MGPS	Person time ignored	DuMouchel (1999)
Bayesian averaging	Person time ignored	Gibbons et al. (2008)
Bayesian multinomial	Person time ignored	Madigan et al. (2005)
Fully Bayesian methods		
Bayes IC	Person time ignored	Norén et al. (2006)
Simplified Bayes	Person time ignored	Huang et al. (2011)
Bayesian Lasso	Person time ignored	Caster et al. (2010)

Abbreviations in Table 1. MaxSPRT: maximized sequential probability ratio test; CSSP: conditional sequential sampling procedure; LGPS: Longitudinal Gamma Poisson Shrinker; LEOPARD: Longitudinal Evaluation of Observational Profiles of Adverse Events Related to Drugs; MUTARA/HUNT: mining the unexpected temporal association rules (TAR) given the antecedent; HUNT: highlighting TARs negating TARs; ROR: reporting odds ratios; PRR: proportional reporting ratios; IC: information component; LRT: likelihood ratio test; BCPNN: Bayesian confidence propagation neural network; MGPS: multi-item gamma Poisson shrinker.

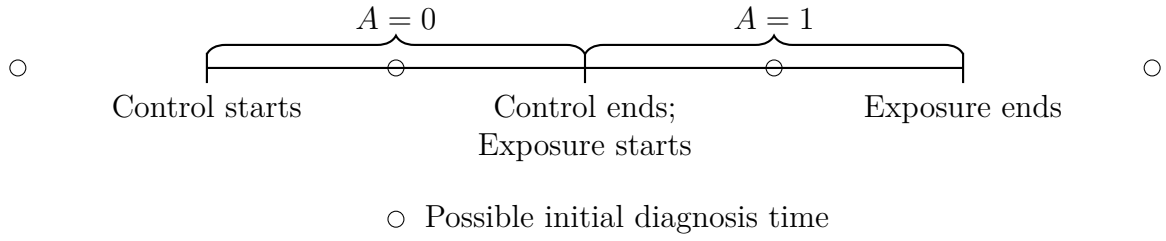

Figure 1: Illustration of self-controlled cohort study design. An example studying the relationship of a drug-disease pair by incorporating all new drug-users into the cohort. Equal person-time are allocated to exposed period after initial prescription and to unexposed period before first treatment for each specific patient. Disease incidence can take place before unexposure starts, during unexposure, during exposure, after exposure ends, or never happens. This arrangement is replicated for available medications on possible diseases in the database.

2.2 Data sources

The screening is conducted on Clinical Practice Research Datalink (CPRD), an ongoing primary care database consisting of more than 60 million participants with 16 million currently registered patients among 674 general practices in the UK (`https://www.cprd.com/`). The follow-up started in 1987 and ended until mortality, transfer-out, or last collection of practice, whichever comes the earliest. The mean and standard deviation of follow-up are 16.77 years and 15.75 years. CPRD include diagnosis (coded in medcode), therapy (coded in prodcode and common dosages), lab tests, consultation, and referral information. The use of CPRD database is approved by Independent Scientific Advisory Committee (ISAC) with protocol 20_000207.

2.3 Exposure lengths

Raw prescription information is available in one of the CPRD tables "therapy". For each prescription, the table contains patient id, "prodcode" for medicinal product, "eventdate" for prescription date, "qty" for total quantity prescribed, and "numdays" for duration entered by prescriber ("CPRD GOLD Data Specification", 2021). By linking to the 'common dosages' table, "dose_duration", estimated duration available for 1% of all data, and raw clinical text can be obtained for every prescription. The "eventdate" in the table is frequently considered as the start date of exposure. Although the stop date is not recorded, it can still be approximated by calculating exposure duration from several sources within existing CPRD data. "numdays" and "dose_duration" are two off-the-shelf variables but failed to work due to large percentage of missing (>95%). Even when present, these two variables are not flexible or variable in determining exposure lengths

when dose frequency (DF), number of doses per day, and dose number (DN), number of tablets to take each time, belong to ranges rather than fixed values. As the percentage of missing is less than 5%, our last resort to calculate exposure period is to divide "qty" by the number of doses to be taken per day, also referred to as numeric daily dose (ndd), which can be computed by

$$\text{ndd} = \frac{\text{DF} \times \text{DN}}{\text{DI}}$$

where DI represents dose interval (number of days between doses). DF, DN, and DI can all be parsed from unstructured free text written by general practitioners following Karystianis et al. (2015); Alfattni et al. (2022) using `R` package `doseminer` (Selby, 2021a). To extend exposure period by reducing "ndd" when clinical texts inform a range of plausible values, we set DF to max, DN and DI to min.

We describe the algorithm along with 10 decision nodes to process drug exposure from therapy data in CPRD in Table 2. There are other plausible options for each decision illustrated in Pye et al. (2018); Yimer et al. (2021b) with detailed description of each decision in Pye et al. (2018, Data S3). Note that these decisions for data preparation can influence risk attribution of clinical outcomes more or less.

The conversion from raw data into a table with exposure length can be roughly realized in 3 broad steps. The initial cleaning step aims to correct missing and implausible values for "qty" and "ndd". Though thresholds for medications can be obtained by medical knowledge and by scraping the British National Formulary (BNF) website `https://bnf.nice.org.uk` following Selby (2021b), the mapping between BNF products and drug substance in CPRD is modest. Thus for simplicity and completeness, we set the maximum of "qty" as 5,000, minimum of "qty" as 1, maximum of "ndd" as 50, and minimum of "ndd" as 1. The second step generates stop dates at the prescription level by "qty" and "ndd". The last step starts by summing durations for the same medication with the same start dates. Then we overlook overlapping prescriptions due to enormous time-complexity when adding overlap to the end of subsequent prescriptions recursively for all drug users. To compensate for possible shorter exposure time, we allow for a maximum of 90-day gap between consecutive refills when constructing the exposure period. The first and second steps are implemented using `R` package `drugprepr` (Yimer et al., 2021a) while the last step leverages `data.table` to boost speed.

Table 2: 10 decision nodes of the exposure preparation algorithm.

Decision nodes	Decisions
Initial cleaning	
Handle implausible qty	Set to population mean
Handle missing qty	Set to population mean
Handle implausible ndd	Set to population mean
Handle missing ndd	Set to population mean
Clean duration variables	Set to 12 months if $>$ 12 months
Prescription length	
Define stop dates	qty/ndd
Missing stop dates	Set to individual mean; if not available use population mean
Continuing prescriptions	
Handle multiple prescriptions	Sum durations
Handle overlapping prescriptions	Ignore overlap
Handle gaps between prescriptions	Assume continuous if gap <= 90 days

The first prescription date is considered as exposure start and the time from treatment initiation until discontinuation is considered as exposure end. Exposure time is then calculated by

$$\begin{aligned}\text{exposure time} = \min\{&30 \text{ days}, \text{exposure start} - \text{frd},\\ &\text{exposure end} - \text{exposure start},\\ &\min(\text{tod}, \text{lcd}) - \text{exposure start}\}\end{aligned}$$

where "frd" stands for first registration date; "tod" represents transfer out date; "lcd" is last collection date ("CPRD GOLD Data Specification", 2021). Then control start = exposure start−exposure time and update exposure end = exposure start+exposure time.

A minimum of 30 days exposure increases the chance to capture clinical outcomes. Once exposure period for each drug user is defined, longitudinal diagnosis history can be combined and assessed.

2.4 Outcome definitions

The first incidence of every disease and category is identified by using code lists phenotyped by validated bioinformatic algorithms from `https://github.com/spiros/chronological-map-phenotypes/tree/master/primary_care` (Kuan et al., 2019). Other code lists, such as `https://www.phpc.cam.ac.uk/pcu/research/research-groups/crmh/cprd_cam/codelists/v11/` (Payne et al., 2020) and `https://clinicalcodes.rss.mhs.man.ac.uk/` (Springate et al., 2014), are not adopted since they are less comprehensive, unified and rigorous. A total of 276 distinct diseases and 16 broad condition categories are tested.

2.5 Removing confounding pairs

A self-controlled cohort study requires that initial exposure is not caused by indication. If prior diagnosis lead to subsequent treatment, bogus protective effect will appear because

the first diagnosis often occurs before initial prescription. Previous studies can remove drug-indication combinations manually based on subject matter knowledge as they focus either on particular diseases (Kern et al., 2019; Cepeda et al., 2019; Teneralli et al., 2021; Kern et al., 2022) or on a specific class of drugs (Kern et al., 2021) with clear primary indication relationship. However, since we aim to screen available drugs on possible diseases, manual removal is laborious, time-consuming, and prone-to-error. Figure 2 demonstrates the medication-indication open loop starting from prodcode and ending by medcode. To the best of our knowledge, there is no existing drug-indication map available within the UK system, and thus we have to turn to the US system and leverage the may_treat relationship between rxcui and Medical Subject Headings (MeSH) according to *RxClass API* (2022). Therefore, we need to map prodcode towards rxcui and MeSH to medcode, respectively.

The open loop starts from prodcode, the only local therapeutic coding system in CPRD which can be mapped towards British National Formulary (BNF) code and gemscript code. In order to connect the UK system to the US system, Systematized Nomenclature of Medicine (SNOMED), an international organized terminology, is selected as the bridge. As the map between gemscript codes and SNOMED drug codes are not actively managed (*Gemscript drug code to SNOMED/DM+D code lookup*, 2020), the UK national BNF code, currently administered by National Institute for Health and Care Excellence (NICE), is adopted instead. Prodcodes are then mapped to the first six digits of BNF codes at the ingredient level (*Prescribing Data: BNF Codes*, 2017). Though BNF codes can only be mapped to UK SNOMED drug codes, the "Has specific active ingredient" attributes further convert UK-only SNOMED drug codes to universal SNOMED ingredient codes, which can be used to match rxcui and rxcui ingredients.

Then we need to map MeSH code towards medcode. As MeSH is US-based while medcode is UK-based, SNOMED is again chosen as the international link. Since SNOMED clinical codes can't be mapped with CPRD-local medcode directly, Readcode, a clinical terminology system that was widely used in UK general practice until 2018, comes into play. SNOMED clinical codes are mapped to Readcode v3 then to Readcode v2. Although Readcode v2 stopped updating in 2016, it is the only version that can be converted to CPRD-local medcode directly. As a result, the drug side, the clinical aspect, along with the rxcui-MeSH drug-indication map can be joined into a comprehensive medcode-prodcode drug-indication table.

After removing drug-disease pairs following the mapped deterministic rules, the remaining drug-disease pairs are still subject to unmappable confounding by indication. To automate the high-throughput screening procedure, we start by calling the ChatGPT API sequentially with the question "is [drug] used to treat [disease]? Just answer yes or no" for all the remaining pairs. This prompt limits the answer from ChatGPT to yes or no without explaining the reasoning of the association. If we modify the last sentence in the prompt to "Just answer yes or no or unknown", then ChatGPT become quite conservative and tend to answer "unknown" but rarely answers "yes". The art of prompt engineering has been explored by John (2023).

After pulling out confonding by indication pairs, the remaining duplets, however, are subject to confonding by risk factors of all indications of the drug of interest. Motivated

by two-stage least squares, we adopt a two step procedure by taking the output from the first stage as part of input in the second stage. For candicate pairs with the potential for drug repurposing, we start by calling the ChatGPT API with the question: "which diseases are [drug] used to treat? Limit answer within eight words" and record the response as [indication.of.drug] besides the drug-disease pairs. We limit the length of the answers since ChatGPT tends to provide explanations which is irrelevant in the next stage. In the second stage, we identify confounding by risk factors of all indications of the drug of interest with the response from the first stage by asking the question: "is any disease in [indication.of.drug] a risk factor of [disease]? Just answer yes or no". Eventually, we can discard all pairs subject to confonding by risk factors of all indications of the drug of interest and the remaining pairs are of our interest.

For pharmacovigilance purposes, the drug-disease pairs still suffer from natural confounding issues. The diseases can be a direct consequence of an indication of the drug, and we remove such pairs by asking ChatGPT "is [disease] caused by any indication of [drug] Just answer yes or no". Though aging does not exacerbate time-varying confounding for drug repurposing in self-controlled cohort studies, it is an major source of bias for drug safety especially for those medications with long exposure. As people getting older after prescribing the drug, the probability of developing aging-related diseases increases regardless of the effect of the medication. Hence, for prescriptions that last longer than a year, we remove pairs with a yes to the question "is [disease] more common as people age? Just answer yes or no".

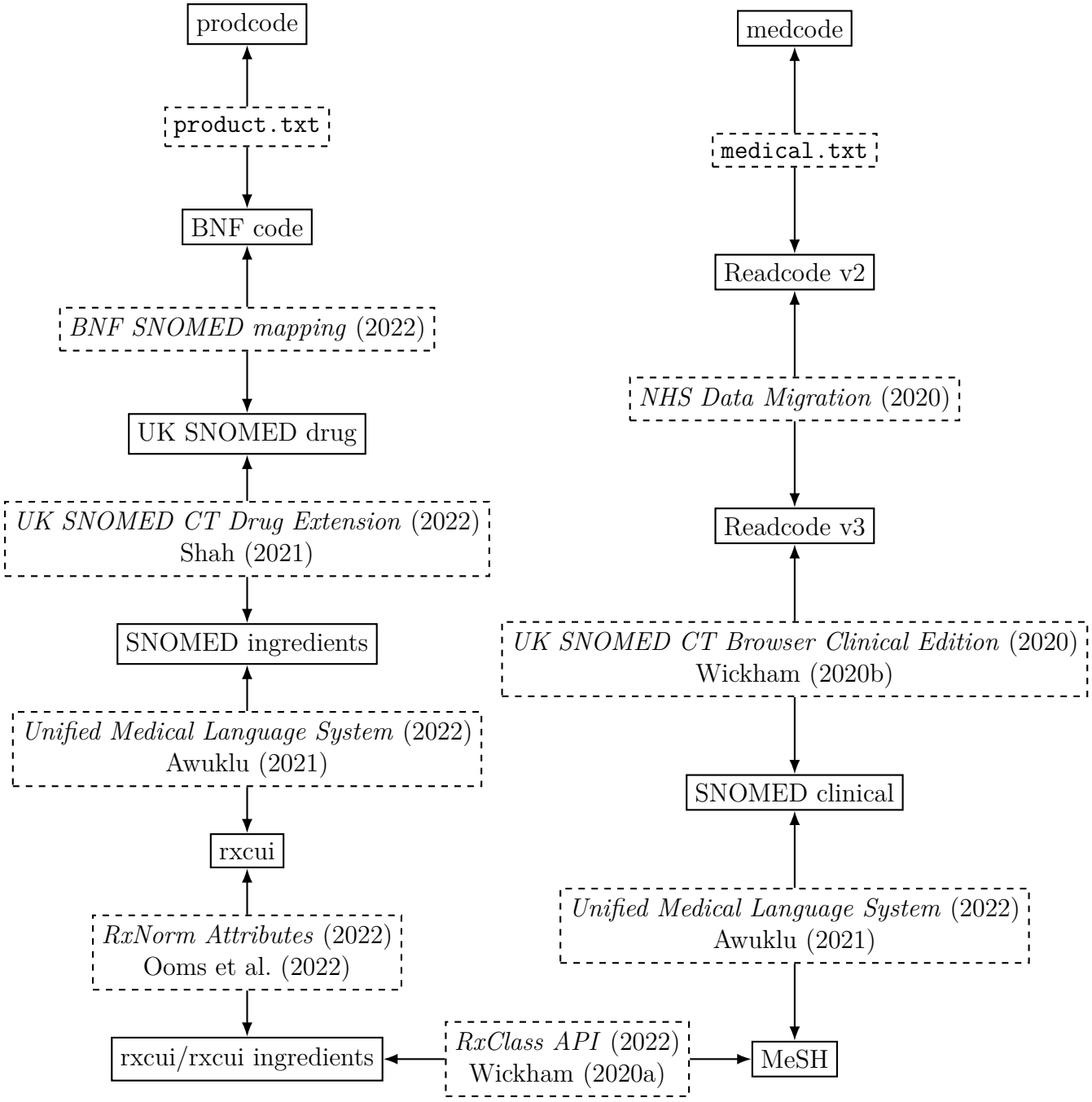

Figure 2: Drug indication map from prodcode to medcode. Solid boxes reveal specific coding system while dashed boxes contain sources of maps between adjacent coding systems along with `R` packages for extraction. If `R` package in a dashed box is missing, then the source of map are in machine-readable format.

2.6 Causal interpretation

To our knowledge, IRR in a self-controlled cohort study has not been couched in explicit counterfactual language, and we discuss causal interpretation of IRR and its additive equivalent, the incident rate difference (IRD) in this section. It can be shown that the interpretability of these quantities relies on untestable common trend assumption between factual rate before exposure and counterfactual rate after treatment initiation had the exposure been removed. This assumption becomes less likely to hold as exposure length

increases, so we conduct sensitivity analysis to inspect how estimates are affected by possible violations of assumption in various extent.

Suppose there are $a*$ ($a = 0, 1, \ldots, a*$) exposures of interest, $j*$ ($j = 1, 2, \ldots, j*$) outcomes of interest, and n_a units who have ever been exposed to treatment a ($i = 1, 2, \ldots, n_a$). Let A_i be the exposure and the time of the first exposure be time zero. Assume $T_{ia,\text{pre}}$ and $T_{ia,\text{post}}$ are control period before time 0 and exposed period after time 0 for treatment a, respectively. Let $Y_{ija,\text{pre}} \in \{0, 1\}$ and $Y_{ija,\text{post}} \in \{0, 1\}$ denote whether unit i experiences non-terminal event j within $[-T_{ia,\text{pre}}, 0]$ and $[0, T_{ia,\text{post}}]$. Note that $Y_{ija,\text{pre}} + Y_{ija,\text{post}} \in \{0, 1\}$ for all i, j, a since a patient can only encounter the event no more than once for each treatment. Define $Y^a_{ij,\text{post}}$ as the counterfactual posttreatment event indicator for outcome j had subject i received treatment a. Note that the potential outcomes for the pre-exposure indicator are not defined since it will never be exposed.

We define the potential posttreatment incidence rate (IR) as

$$\text{IR}^a_{j,\text{post}} = \frac{E(Y^a_{ij,\text{post}})}{E(T_{ia,\text{post}})}$$

Then, the causal incidence rate ratio (IRR) can be defined as

$$\text{IRR}^a_j = \frac{\text{IR}^a_{j,\text{post}}}{\text{IR}^{a=0}_{j,\text{post}}}$$

and the causal incidence rate difference (IRD) as

$$\text{IRD}^a_j = \text{IR}^a_{j,\text{post}} - \text{IR}^{a=0}_{j,\text{post}}$$

The following conditions are required to identify IRR or IRD.

Assumption 1. *Stable unit treatment value assumption (SUTVA): including no interference between subjects after or before exposure* $Y^{(A_1,A_2,\ldots,A_n)}_{ij,post} = Y^{(A'_1,A'_2,\ldots,A'_n)}_{ij,post}$, *if* $A_i = A'_i$, $\forall i$; *and consistency* $Y^a_{ij,post} = Y_{ija,post}$;

Assumption 2. *Common intensity assumption:* $E(Y_{ija,pre})/E(T_{ia,pre}) = E(Y^{a=0}_{ij,post})/E(T_{ia,post})$. *Had the exposure been removed, the population pretreatment intensity equals to the potential population post-exposure intensity;*

Assumption 3. *Positivity of population pre-exposed period* $E(T_{ia,pre}) > 0$, *positivity of population post-exposed period* $E(T_{ia,post}) > 0$, *and positivity of population pretreatment observed outcomes* $E(Y_{ija,pre}) > 0$. *Causal IRD does not require* $E(Y_{ija,pre}) > 0$.

Assumptions 1 and 2 are crucial to identify IRR/IRD but they are both empirically unverifiable. Assumption 2 is similar to parallel trends assumption in difference-in-differences (Abadie, 2005) and rate-change assumptions in calibrated self-controlled cohort study (van Aalst et al., 2021). Note that this assumption is required for self-controlled cohort studies but exchangeability is not needed since its external control group is absent.

Assumption 3 is ensured automatically since the study is designed to be self-controlled. In addition to these requirements, all subjects are assumed to be observable from unexposure starts until exposure ends. Identification issues pertaining to administrative censoring, terminal events such as death, recurrent event, intermittent exposure, and lagtime are beyond the scope of this work (In'T Veld et al., 2001; Power et al., 2015).

Under Assumptions 1, 2, and 3, the causal IRR can be identified and estimated as

$$\mathrm{IRR}_{ja} = \frac{E(Y_{ja,\mathrm{post}})/E(T_{a,\mathrm{post}})}{E(Y_{ja,\mathrm{pre}})/E(T_{a,\mathrm{pre}})}, \quad \widehat{\mathrm{IRR}}_{ja} = \frac{\sum_{i=1}^{n_a} Y_{ija,\mathrm{post}} / \sum_{i=1}^{n_a} T_{ia,\mathrm{post}}}{\sum_{i=1}^{n_a} Y_{ija,\mathrm{pre}} / \sum_{i=1}^{n_a} T_{ia,\mathrm{pre}}}$$

and causal IRD can be identified and estimated as

$$\mathrm{IRR}_{ja} = \frac{E(Y_{ja,\mathrm{post}})}{E(T_{a,\mathrm{post}})} - \frac{E(Y_{ja,\mathrm{pre}})}{E(T_{a,\mathrm{pre}})}, \quad \widehat{\mathrm{IRD}}_{ja} = \frac{\sum_{i=1}^{n_a} Y_{ija,\mathrm{post}}}{\sum_{i=1}^{n_a} T_{ia,\mathrm{post}}} - \frac{\sum_{i=1}^{n_a} Y_{ija,\mathrm{pre}}}{\sum_{i=1}^{n_a} T_{ia,\mathrm{pre}}}$$

Suppose the IRR is a ratio between two rates with Poisson distribution, then the closed-form confidence interval can be computed following Graham et al. (2003) as

$$\mathrm{CI}(\widehat{\mathrm{IRR}}_{ja}) = \frac{\sum_{i=1}^{n_a} T_{ia,\mathrm{pre}} / \sum_{i=1}^{n_a} T_{ia,\mathrm{post}}}{2\left(\sum_{i=1}^{n_a} Y_{ija,\mathrm{pre}}\right)^2} \left[2\sum_{i=1}^{n_a} Y_{ija,\mathrm{pre}} \sum_{i=1}^{n_a} Y_{ija,\mathrm{post}} + \left(z_{\alpha/2}\right)^2 \sum_{i=1}^{n_a} (Y_{ija,\mathrm{pre}} + Y_{ija,\mathrm{post}}) \right.$$
$$\left. \pm \sqrt{\left(z_{\alpha/2}\right)^2 \sum_{i=1}^{n_a} (Y_{ija,\mathrm{pre}} + Y_{ija,\mathrm{post}}) \times \left\{ 4\sum_{i=1}^{n_a} Y_{ija,\mathrm{pre}} \sum_{i=1}^{n_a} Y_{ija,\mathrm{post}} + \left(z_{\alpha/2}\right)^2 \sum_{i=1}^{n_a} (Y_{ija,\mathrm{pre}} + Y_{ija,\mathrm{post}}) \right\}} \right]$$

where $z_{\alpha/2}$ is the z-statistic with type I error rate $\alpha/2$. The closed-form large sample z-test based confidence intervals for IRD between two Poisson rates can be found in Krishnamoorthy and Thomson (2004).

The selection between IRR and IRD depends mainly on research tasks. IRR has the advantage of cancelling background scale such that comparison across treatment a and outcome j can be made directly. IRD focuses on the absolute scale of contrast whose intrinsic incidence rates may differ substantially across a and j such that broader comparisons become less meaningful.

The study results can be controversial in situations especially when $T_{ia,\mathrm{post}}$ or $T_{ia,\mathrm{pre}}$ is large since time-varying factors may affect the validity of IRR/IRD analyses with critical reliance on the untestable Assumption 2 common intensity. Here, we provide sensitivity analysis to examine how violations of various scale would affect estimates. For IRR, suppose that $E(Y^{a=0}_{ij,\mathrm{post}})/E(T_{ia,\mathrm{post}}) \neq E(Y_{ija,\mathrm{pre}})/E(T_{ia,\mathrm{pre}}) = E(Y^{a=0}_{ij,\mathrm{post}})/E(T_{ia,\mathrm{post}}) \times \mathrm{bias}_{\mathrm{IRR}}$, where $\mathrm{bias}_{\mathrm{IRR}} > 0$ is the bias for IRR. Under this sensitivity model, the IRR can be expressed as

$$\mathrm{IRR}_{ja} = \frac{E(Y^{a}_{ij,\mathrm{post}})/E(T_{ia,\mathrm{post}})}{E(Y^{a=0}_{ij,\mathrm{post}})/E(T_{ia,\mathrm{post}})} \times \frac{1}{\mathrm{bias}_{\mathrm{IRR}}} = \mathrm{IRR}^a_j \times \frac{1}{\mathrm{bias}_{\mathrm{IRR}}}$$

Such that when $\mathrm{bias}_{\mathrm{IRR}} = 1$, $\widehat{\mathrm{IRR}}_{ja}$ becomes an unbiased estimator for IRR^a_j; when $0 < \mathrm{bias}_{\mathrm{IRR}} < 1$, $\widehat{\mathrm{IRR}}_{ja}$ serves as an upper bound for IRR^a_j; whereas when $\mathrm{bias}_{\mathrm{IRR}} > 1$, $\widehat{\mathrm{IRR}}_{ja}$ acts as an lower bound for IRR^a_j.

For IRD, we can parameterize the violation as $E(Y_{ija,\text{pre}})/E(T_{ia,\text{pre}}) \neq E(Y^{a=0}_{ij,\text{post}})/E(T_{ia,\text{post}}) = E(Y^{a=0}_{ij,\text{post}})/E(T_{ia,\text{post}}) - \text{bias}_{\text{IRD}}$, where bias_{IRD} is the bias for IRD. Under this sensitivity model, the IRD can be expressed as

$$\text{IRD}_{ja} = \frac{E(Y^{a}_{ij,\text{post}})}{E(T_{ia,\text{post}})} - \frac{E(Y^{a=0}_{ij,\text{post}})}{E(T_{ia,\text{post}})} + \text{bias}_{\text{IRD}} = \text{IRD}^{a}_{j} + \text{bias}_{\text{IRD}}$$

When $\text{bias}_{\text{IRD}} = 0$, $\widehat{\text{IRD}}_{ja}$ becomes an unbiased estimator for IRD^{a}_{j}; when $\text{bias}_{\text{IRD}} > 0$, $\widehat{\text{IRD}}_{ja}$ serves as an upper bound for IRD^{a}_{j}; whereas when $\text{bias}_{\text{IRD}} < 0$, $\widehat{\text{IRD}}_{ja}$ acts as an lower bound for IRD^{a}_{j}.

As neither bias_{IRR} nor bias_{IRD} can be estimated from data, our sensitivity analysis can be conducted by testing a set of values. Note that the conditional counterfactual incidence rate can be defined as $\text{IR}^{a}_{j,\text{post}}(x) = E(Y^{a}_{ij,\text{post}} \mid X_i = x)/E(T_{ia,\text{post}} \mid X_i = x)$, where X must be baseline time-invariant covariates, such that conditional counterfactual IRR/IRD, identification conditions, estimators, along with sensitivity analysis can be adapted and derived accordingly.

3 Application

A total of 6,613,198 patients, 3,444 medications, and 276 diseases were analyzed in this study. We also investigate various exposure lengths, age groups at initial prescription, drug classes, along with more general disease categories. The exposed period is designed to be the same as unexposed period at the patient level for symmetry and simplicity. Only drug-disease pairs satisfying the following conditions are included: (1) drug does not confound with disease through known pathways; (2) after pairing with a specific drug, the total number of outcomes should be more than 100; (3) the number of outcomes during both control and exposure period is larger than 30. The full lists of drug-disease pairs are available upon request.

If there is no association between the exposure and the outcome, the pretreatment incidence rate should be approximately identical to the posttreatment incidence rate such that the estimated IRR should not be significantly away from 1. An upper 95% confidence interval of IRR < 1 reveals potential protective effect while an lower 95% confidence interval of IRR > 1 indicates possible adverse reactions. A total of 16,901 drug-disease pairs are found with significant risk reduction and a total of 11,089 pairs revealed significant risk augmentation.

For repurposing candidates, we focus on dementia and present upper 95% confidence interval of IRR, the number of participants exposed to each drug, exposure period mean, and exposure period standard deviation by increasing upper 95% confidence interval of IRR in Table 3. Compared with results using IBM MarketScan data (Kern et al., 2019), we discover no overlap between two databases. Though results on other diseases are not tabulated explicitly, our results revealed no overlap with IBM MarketScan on bipolar affective disorder, mania, or depression (Teneralli et al., 2021), and post-traumatic stress disorder (PTSD) (Kern et al., 2022). The only common drug-disease pair among all previous finding is that propranolol hydrochloride might postpone the diagnosis but of Parkinson's disease (Cepeda et al., 2019).

Table 3: Candidates for repurposing. upper: upper 95% confidence interval of IRR; N exposed: number of participants; exposure mean: exposure period mean; exposure sd: exposure period standard deviation.

drug	disease	upper	N exposed	exposure mean	exposure sd
chloroform/magnesium oxide light/ magnesium sulfate dried/sodium hydroxide	dementia	0.51	26189	28.31	5.76
folic acid	dementia	0.53	280096	28.43	5.53
omeprazole	dementia	0.53	1408560	28.86	4.72
dipyridamole	dementia	0.54	81372	28.21	5.89
paracetamol	dementia	0.56	1455955	28.53	5.40
promethazine hydrochloride	dementia	0.56	77760	29.31	3.71
quinine bisulfate	dementia	0.58	308277	29.03	4.36
latanoprost	dementia	0.61	101823	28.36	5.64
permethrin	dementia	0.62	118084	321.29	96.81
benzyl alcohol/benzyl benzoate/ benzyl cinnamate/wool fat/zinc oxide	dementia	0.64	146876	196.44	292.15
tamsulosin hydrochloride	dementia	0.64	207294	28.92	4.61
ketoconazole	dementia	0.64	188727	29.66	2.59
benzalkonium chloride/dimeticone	dementia	0.65	72381	242.90	142.84
brinzolamide	dementia	0.66	27258	28.77	4.92
benzyl alcohol/benzyl benzoate/ benzyl cinnamate/wool fat/zinc oxide	dementia	0.66	146583	257.96	137.83
benzalkonium chloride/dimeticone	dementia	0.67	72596	233.27	339.55
cloral betaine	dementia	0.67	15643	28.94	4.57
malathion	dementia	0.68	82949	325.99	92.16
ibuprofen	dementia	0.69	1862997	29.56	2.97
benzyl alcohol/benzyl benzoate/ benzyl cinnamate/wool fat/zinc oxide	dementia	0.69	146583	28.36	5.57

Table 4: Candidates for pharmacovigilance. lower: lower 95% confidence interval of IRR; N exposed: number of participants; exposure mean: average exposure length; exposure sd: standard deviation of exposure length.

drug	disease	lower	N exposed	exposure mean	exposure sd
atenolol	primary pulmonary hypertension	3.90	641137	867.89	1345.64
fusidic acid	anterior and intermediate uveitis	3.77	729364	29.78	2.09
fusidic acid	anterior and intermediate uveitis	3.69	729477	33.11	31.22
naproxen	multiple myeloma and malignant plasma cell neoplasms	3.54	994272	327.66	91.45
nicorandil	anorectal fistula	3.37	82631	753.49	1100.67
simvastatin	aspiration pneumonitis	3.21	1057217	956.02	1199.80
citric acid anhydrous/magnesium oxide/ sodium picosulfate	diverticular disease of intestine	3.16	7285	29.63	2.68
disodium hydrogen phosphate dodecahydrate/ sodium dihydrogen phosphate anhydrous	diverticular disease of intestine	3.16	31785	28.09	5.99
salbutamol	dilated cardiomyopathy	2.95	1160084	290.81	127.41
cyclopenthiazide/potassium chloride	dermatitis	2.91	17013	376.98	372.25
citric acid anhydrous/magnesium oxide/ sodium picosulfate	diverticular disease of intestine	2.88	7291	37.37	71.56
chamomile extract	menorrhagia and polymenorrhoea	2.88	8918	327.82	86.59
malathion	lichen planus	2.86	82966	132.14	106.52
disodium hydrogen phosphate dodecahydrate/ sodium dihydrogen phosphate anhydrous	diverticular disease of intestine	2.78	31861	59.74	140.61
phenobarbital	dermatitis	2.77	8444	843.32	1622.39
disodium hydrogen phosphate dodecahydrate/ sodium dihydrogen phosphate anhydrous	benign neoplasm and cin	2.77	31785	28.09	5.99
malathion	lichen planus	2.75	82949	325.99	92.16
metformin	primary pulmonary hypertension	2.69	358596	984.75	1251.41
aciclovir	trigeminal neuralgia	2.68	382413	332.59	84.53

To the best of our knowledge, no self-controlled cohort study has been conducted to explore unknown adverse effects on various diseases. After removing malignancy outcomes, the lower 95% confidence interval of IRR, the number of participants exposed to each drug, exposure period mean, and exposure period standard deviation are presented partially by decreasing lower 95% confidence interval of IRR in table 4.

4 Discussion

This study established a general workflow to screen medications on clinical outcomes. We aim to provide accessible and economical hypothesis-generating tests with limited validity and rigor for the relationship between exposure and clinical outcomes Rothman et al. (2008, Chapter 6). As identifying candidates is the very first step to discover both new uses and unknown ADRs of existing medications, more stringent and expensive hypothesis-testing confirmatory studies such as observational studies or even RCTs with external control group are required to validate signals found in this investigation, which can inform current medical guidelines.

4.1 Strengths

There are several strengths of this work. The self-controlled cohort study is applicable to both drug repurposing and pharmacovigilance by allowing subjects to act as their own control such that all time-fixed covariates, whether observed or not, such as genetics, are automatically controlled for. Compared with other cohort studies, narrow confidence intervals induced by underestimated variability and erroneous findings provoked by multiple comparisons can be avoided by overshooting risk reduction and risk augmentation. Even though potential effects are likely to be missed, the lack of significant discoveries indicates estimated association is not substantial enough to be incorporated rather than absence of such relationship. Afterall, potential false positive (type 1 error) is not a big concern for hypothesis-screening studies since we are targeting candidates for further research instead of confirming absolute causal effects.

We defined causal IRR/IRD and outlined conditions for identification. The major provision that differentiates from exchangeability-based external control group is the common intensity assumption, which involves the relationship between counterfactual posttreatment occurrence had the exposure been removed and observed factual pre-exposure occurrence. Though this can barely hold in practice, the estimated IRR/IRD can still serve as upper bounds for causal IRR/IRD if the control/exposed period is long enough such that aging becomes a dominant factor which boosts post-treatment incidence. As drug screening requires population-level estimates, we confine our focus to Imbens approach (Imbens, 2003) which targets average treatment effects for sensitivity analysis rather than Rosenbaums approach (Rosenbaum, 2002, Chapter 3) which considers sharp null tests and p-values from randomization inference. A potential benefit of Imbens approach is that detailed subject matter knowledge for unmeasured time-varying confounders or their relationship with observed data is not required (Robins et al., 2000). Moreover, IRR/IRD conditioned on time-fixed covariates can be defined and identified readily by some modifications on the assumptions for unconditional IRR/IRD.

Exchangeability-based methods relying on external control group require positivity, SUTVA

and conditional ignorability such that treatment and control groups become comparable (Hernán and Robins, 2020). Notably, self-controlled cohort studies are inspired by the argument that observational healthcare data sources seldomly meet these strong and unverifiable assumptions. Putting strong and unverifiable assumptions to handle confounding by conditioning on measured covariates aside, the control group must share the same indication with the treatment group to render sufficient common support due to intrinsic confounding by indication in observational healthcare data. Consistency in SUTVA worsens the selection of external control by demanding only one version of treatment/control, which narrows the comparator group to a set of very limited treatment regimes. The choice between external control and internal control depends heavily on the scientific question and when it comes to high-throughput drug screening, self-controlled studies revealed dramatic advantage over external controlled studies.

Exposure periods serve as a critical component to capture clinical events and to compute incidence rates in self-controlled cohort studies. Owing to inherent discrepancies, prescription lengths are not uniform across medication, patient, practice, and region and setting a fixed value for all prescriptions will lead to inaccurate exposure length, inducing biased IRR/IRD estimates. Therefore, more accurate exposure lengths are computed for each prescription by parsing clinical texts using common dosage information in CPRD. Time-varying confounding for exposure continuity, cessation, and switching can't be adjusted for in self-controlled cohort studies which may cause additional bias in either direction. Moreover, as some prescriptions have some variability and resilience, sensitivity analysis may be conducted to demonstrate how analytical decisions could affect IRR/IRD in future studies.

Though the major drawback of self-controlled cohort study is its intrinsic confounding related to indications, contraindications, comorbidities, complications, and off-label uses, where temporal sequences are predetermined by existing clinical guidelines or natural connection between diseases. In such cases, medication-disease pairs should be disposed in that these directional effects are expected to show up. Due to medical ontology incompatibility between the UK and the US, only mappable drug-indication pairs can be removed and without the power of modern LLMs, additional manual removal is required based on medical domain knowledge from physicians. The application of ChatGPT not only handles unmappable confounding by indication and other types of drug-disease relationships, but it also opens a door to generate nonexistent high-quality ontologies and extract clinical information from clinical texts, which will add value to the field of bioinformatics and pharmacoepidemiology (Shue et al., 2023).

The last major benefit of the study design would be computational efficiency, allowing researchers to perform large-scale screening on the entire database at relatively fast rate, which can be extended to other similar observational databases.

4.2 Limitations

Chronic diseases might not be well suited for self-controlled cohort analysis when the amount of time on medication after initial prescription is abridged. If this is true, the IR of diseases before and after treatment are expected to be similar as aging does not play an essential role when the exposure length is short. However, this is not the case

for many drug-disease pairs found in the study. The main reason is that patients with shorter exposure time add less information to capture outcomes compared with those with relatively longer exposure. Medications with shorter exposure are therefore not a big issue to our study though those with longer exposure are likely to be more meaningful for incidence of chronic conditions. Aside from exposure period, the exact onset time of chronic diseases is hard to determine as symptoms start gradually and formal diagnosis may take place years after actual onset. Slow progression and delayed realization can result in temporal misclassification of chronic diseases and eventually, diagnoses are more likely to be captured after exposure even if the actual condition showed up before drug exposure. This bias leads to erroneously increased estimated risks which may reduce true negatives for repurposing intensions and enlarge false positives for drug safety concerns.

Diagnosis codes to phenotype clinical outcomes are based on established studies rather than subjective definition of conditions. To minimize the impact of under-recording in CPRD data, we do not require multiple diagnoses for the same condition to identify clinical outcomes. Therefore, false positive outcomes may appear due to single exclusion, mis-recording, mis-diagnosis, and misclassification. If false positive cases are non-differential with respect to the treatment, then the results should lie around the null which can't be explained by the directional effects found in the analysis. Moreover, only patients in the UK primary care database are evaluated. Medications not marketed in the UK can be assessed but our approach may require modification when applied to other countries, as treatment guidelines may differ outside the UK.

Another potential limitation of this study is the accuracy of the ChatGPT responses to the queries used to identify known drug-disease associations. Large Language Models (LLMs) like ChatGPT are known to occasionally produce confident but inaccurate responses, a phenomenon often referred to as "hallucination." This inherent limitation of LLMs could potentially introduce false positives or negatives in the identification of drug-disease associations, thereby affecting the overall reliability of the pharmacovigilance analysis. While LLMs have demonstrated impressive capabilities in processing and synthesizing vast amounts of biomedical literature, their outputs should be treated with caution and verified against established databases and expert knowledge. Future work in this area should consider implementing additional validation steps, such as cross-referencing ChatGPT's outputs with curated biomedical databases, or developing ensemble approaches that combine LLM-generated insights with traditional bioinformatics methods to mitigate the risk of hallucinated associations and enhance the robustness of the pharmacovigilance findings.

5 Conclusion

There are large unmet medical needs to discover effective disease-modifying therapies and unknown side-effects of existing drugs. The increasing volume and availability of observational healthcare databases provides the basis for detecting unknown benefits and detriments in-silico via large-scale drug screening. We therefore compared current pharmacoepidemiology study design and choose self-controlled cohort study to assess the association between initiation of marketed drugs and the onset of possible diseases in millions of patients using real-world UK primary care EHR data CPRD. Accurate exposure period is calculated by construing unstructured texts. Due to built-in selection bias

of self-controlled cohort study, intended drug-disease pairs are discarded based on cross-ontology maps and answers from the recent LLM - ChatGPT. We also offer causal-wise interpretation of incidence rate contrasts along with Imbens-type sensitivity analysis on the critical common intensity assumption. By screening signals in both directions, this approach identifies 16,901 drug-disease pairs with reduced risk as potential candidates for repurposing and 11,089 pairs with excess risk as possible unknown ADRs. The results of this large-scale analytics can be followed up with more rigorous attention and help generate hypotheses for subsequent observational, preclinical, and clinical research, which examines the validity and efficacy of our paradigm. The general workflow of this work unlocks the potential of AIGC on bioinformatics and pharmacoepidemiology, and can be generalized easily to other observational healthcare databases.

Acknowledgement

This work was supported by IBM Research with award number W1771646 and National Institutes of Health with award number R01AG058063-04. The authors thank Prof. Zach Shahn at the City University of New York, Dr. Bang Zheng and Dr. Bowen Su at Imperial College London for their discussion, explanations, and guidance. The authors are also grateful to the editors and the reviewers for their insightful suggestions.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies*, *72*(1), 1–19.

Alfattni, G., Peek, N., Nenadic, G., and Caskey, F. (2022). Integrating text analytics and statistical modelling to analyse kidney transplant immune suppression medication in registry data. *International Journal of Population Data Science*, *1*(1).

Awuklu, Y. (2021). *getUMLS: Query the UMLS metathesaurus* [Manual]. Retrieved from `https://github.com/yvoawk/getUMLS/releases/tag/v0.1.0` (R package version 0.1.0)

Bao, Y., Kuang, Z., Peissig, P., Page, D., and Willett, R. (2017). Hawkes process modeling of adverse drug reactions with longitudinal observational data. In *Machine learning for healthcare conference* (Vol. 68, pp. 177–190).

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., and De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, *54*, 315–321.

Breslow, N., Day, N., and Schlesselman, J. J. (1982). Statistical methods in cancer research. volume 1the analysis of case-control studies. *Journal of Occupational and Environmental Medicine*, *24*(4), 255–257.

Brown, J. S., Kulldorff, M., Chan, K. A., Davis, R. L., Graham, D., Pettus, P. T., ... others (2007). Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and drug safety*, *16*(12), 1275–1284.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... others (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cao, H., Hripcsak, G., and Markatou, M. (2007). A statistical methodology for analyzing co-occurrence data from a large sample. *Journal of biomedical informatics*, *40*(3), 343–352.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., and Sun, L. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.

Caster, O., Norén, G. N., Madigan, D., and Bate, A. (2010). Large-scale regression-based pattern discovery: the example of screening the who global drug safety database. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *3*(4), 197–208.

Cepeda, M. S., Kern, D. M., Seabrook, G. R., and Lovestone, S. (2019). Comprehensive real-world assessment of marketed medications to guide parkinson's drug discovery. *Clinical Drug Investigation*, *39*, 1067–1075.

Cook, A. J., Tiwari, R. C., Wellman, R. D., Heckbert, S. R., Li, L., Heagerty, P., ... Nelson, J. C. (2012). Statistical approaches to group sequential monitoring of post-market safety surveillance data: current state of the art for use in the mini-sentinel pilot. *pharmacoepidemiology and drug safety*, *21*, 72–81.

CPRD GOLD Data Specification [Manual]. (2021). Retrieved from `https://cprd.com/sites/default/files/CPRD%20GOLD%20Full%20Data%20Specification%20v2.4.pdf` (version 2.4)

Domínguez-Almendros, S., Benítez-Parejo, N., and Gonzalez-Ramirez, A. R. (2011). Logistic regression models. *Allergologia et immunopathologia*, *39*(5), 295–305.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician*, *53*(3), 177–190.

DuMouchel, W., Ryan, P. B., Schuemie, M. J., and Madigan, D. (2013). Evaluation of disproportionality safety signaling applied to healthcare databases. *Drug safety*, *36*, 123–132.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Ernster, V. L. (1994). Nested case-control studies. *Preventive medicine*, *23*(5), 587–590.

Gemscript drug code to SNOMED/DM+D code lookup. (2020). `https://www.whatdotheyknow.com/request/gemscript_drug_code_to_snomed_dm`. (Accessed: 2022-01-18)

Gibbons, J., Cox, G., Wood, A., Craigon, J., Ramsden, S., Tarsitano, D., and Crout, N. (2008). Applying bayesian model averaging to mechanistic models: An example and comparison of methods. *Environmental Modelling & Software*, *23*(8), 973–985.

Glicksberg, B. S., Li, L., Chen, R., Dudley, J., and Chen, B. (2019). Leveraging big data to transform drug discovery. *Bioinformatics and Drug Discovery*, 91–118.

Gozalo-Brizuela, R., and Garrido-Merchan, E. C. (2023). Chatgpt is not all you need. a state of the art review of large generative ai models. *arXiv preprint arXiv:2301.04655*.

Graham, P., Mengersen, K., and Morton, A. (2003). Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. *Statistics in medicine*, *22*(12), 2071–2083.

Hallas, J. (1996). Evidence of depression provoked by cardiovascular medication: a prescription sequence symmetry analysis. *Epidemiology*, *7*(5), 478–484.

Hallas, J., and Pottegård, A. (2014). Use of self-controlled designs in pharmacoepidemiology. *Journal of internal medicine*, *275*(6), 581–589.

Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P., and Shah, N. H. (2013). Performance of pharmacovigilance signal-detection algorithms for the fda adverse event reporting system. *Clinical Pharmacology & Therapeutics*, *93*(6), 539–546.

Hernán, M., and Robins, J. (2020). *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC.

Huang, L., Guo, T., Zalkikar, J. N., and Tiwari, R. C. (2014). A review of statistical methods for safety surveillance. *Therapeutic Innovation & Regulatory Science*, *48*(1), 98–108.

Huang, L., Zalkikar, J., and Tiwari, R. C. (2011). A likelihood ratio test based method for signal detection with application to fda's drug safety data. *Journal of the American Statistical Association*, *106*(496), 1230–1241.

Huang, L., Zalkikar, J., and Tiwari, R. C. (2013). Likelihood ratio test-based method for signal detection in drug classes using FDA's AERS database. *Journal of biopharmaceutical statistics*, *23*(1), 178–200.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, *93*(2), 126–132.

In'T Veld, B. A., Ruitenberg, A., Hofman, A., Launer, L. J., van Duijn, C. M., Stijnen, T., ... Stricker, B. H. (2001). Nonsteroidal antiinflammatory drugs and the risk of alzheimer's disease. *New England Journal of Medicine*, *345*(21), 1515–1521.

Ji, Y., Ying, H., Dews, P., Mansour, A., Tran, J., Miller, R. E., and Massanari, R. M. (2011). A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine*, *15*(3), 428–437.

Jin, H., Chen, J., Kelman, C., He, H., McAullay, D., and OKeefe, C. M. (2006, 4). Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases. In *Advances in knowledge discovery and data mining: 10th pacific-asia conference, pakdd 2006, singapore, april 9-12, 2006. proceedings 10* (Vol. 3918, pp. 867–876).

John, I. (2023). *The art of asking chatgpt for high-quality answers*.

Jordan, S., Logan, P. A., Panes, G., Vaismoradi, M., and Hughes, D. (2018). Adverse drug reactions, power, harm reduction, regulation and the adre profiles. *Pharmacy*, *6*(3), 102.

Karystianis, G., Sheppard, T., Dixon, W. G., and Nenadic, G. (2015). Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. *BMC medical informatics and decision making*, *16*(1), 1–10.

Kern, D. M., Cepeda, M. S., Flores, C. M., and Wittenberg, G. M. (2021). Application of real-world data and the REWARD framework to detect unknown benefits of memantine and identify potential disease targets for new NMDA receptor antagonists. *CNS drugs*, *35*, 243–251.

Kern, D. M., Cepeda, M. S., Lovestone, S., and Seabrook, G. R. (2019). Aiding the discovery of new treatments for dementia by uncovering unknown benefits of existing medications. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *5*, 862–870.

Kern, D. M., Teneralli, R. E., Flores, C. M., Wittenberg, G. M., Gilbert, J. P., and Cepeda, M. S. (2022). Revealing unknown benefits of existing medications to aid the discovery of new treatments for post-traumatic stress disorder. *Psychiatric Research and Clinical Practice*, *4*(1), 12–20.

Krishnamoorthy, K., and Thomson, J. (2004). A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*, *119*(1), 23–35.

Kuan, V., Denaxas, S., Gonzalez-Izquierdo, A., Direk, K., Bhatti, O., Husain, S., ... others (2019). A chronological map of 308 physical and mental health conditions from 4 million individuals in the English national health service. *The Lancet Digital Health*, *1*(2), e63–e77.

Kulldorff, M., Dashevsky, I., Avery, T. R., Chan, A. K., Davis, R. L., Graham, D., ... others (2013). Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiology and drug safety*, *22*(5), 517–523.

Kulldorff, M., Fang, Z., and Walsh, S. J. (2003). A tree-based scan statistic for database disease surveillance. *Biometrics*, *59*(2), 323–331.

Laifenfeld, D., Yanover, C., Ozery-Flato, M., Shaham, O., Rosen-Zvi, M., Lev, N., ... Grossman, I. (2021). Emulated clinical trials from longitudinal real-world data efficiently identify candidates for neurological disease modification: examples from parkinsons disease. *Frontiers in pharmacology*, *12*, 631584.

Lee, P., Goldberg, C., and Kohane, I. (2023). *The ai revolution in medicine: Gpt-4 and beyond*. Pearson.

Li, L. (2009). A conditional sequential sampling procedure for drug safety surveillance. *Statistics in medicine*, *28*(25), 3124–3138.

Maclure, M. (1991). The case-crossover design: a method for studying transient effects on the risk of acute events. *American journal of epidemiology*, *133*(2), 144–153.

Madigan, D., Genkin, A., Lewis, D. D., and Fradkin, D. (2005). Bayesian multinomial logistic regression for author identification. In *Aip conference proceedings* (Vol. 803, pp. 509–516).

Madigan, D., Schuemie, M. J., and Ryan, P. B. (2013). Empirical performance of the case–control method: lessons for developing a risk identification and analysis system. *Drug Safety*, *36*, 73–82.

Mittal, S., Bjørnevik, K., Im, D. S., Flierl, A., Dong, X., Locascio, J. J., ... others (2017). β2-adrenoreceptor is a regulator of the α-synuclein gene driving risk of Parkinson's disease. *Science*, *357*(6354), 891–898.

Murphy, S. N., Castro, V., Colecchi, J., Dubey, A., Gainer, V., Herrick, C., and Sordo, M. (2011). *Partners healthcare OMOP study report.*

Nam, K., Henderson, N. C., Rohan, P., Woo, E. J., and Russek-Cohen, E. (2017). Logistic regression likelihood ratio test analysis for detecting signals of adverse events in post-market safety surveillance. *Journal of biopharmaceutical statistics*, *27*(6), 990–1008.

Norén, G. N., Bate, A., Orre, R., and Edwards, I. R. (2006). Extending the methods used to screen the who drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in medicine*, *25*(21), 3740–3757.

Norén, G. N., Bergvall, T., Ryan, P. B., Juhlin, K., Schuemie, M. J., and Madigan, D. (2013). Empirical performance of the calibrated self-controlled cohort analysis within temporal pattern discovery: lessons for developing a risk identification and analysis system. *Drug safety*, *36*, 107–121.

Norén, G. N., Hopstadius, J., Bate, A., Star, K., and Edwards, I. R. (2010). Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, *20*, 361–387.

OHDSI. (2020). *The book of OHDSI.*

Ooms, J., Lang, D. T., and Hilaiel, L. (2022). *jsonlite: A simple and robust JSON parser and generator for R* [Manual]. Retrieved from `https://cran.r-project.org/web/packages/jsonlite/index.html` (R package version 1.7.3)

OpenAI. (2023a). *Chatgpt-4.* `https://chat.openai.com/chat`. (Accessed: 2023-03-20)

OpenAI. (2023b). *Gpt-4 technical report.*

Payne, R. A., Mendonca, S. C., Elliott, M. N., Saunders, C. L., Edwards, D. A., Marshall, M., and Roland, M. (2020). Development and validation of the cambridge multimorbidity score. *Cmaj*, *192*(5), E107–E114.

Petersen, I., Douglas, I., and Whitaker, H. (2016). Self controlled case series methods: an alternative to standard epidemiological study designs. *bmj*, *354*.

Power, M. C., Weuve, J., Sharrett, A. R., Blacker, D., and Gottesman, R. F. (2015). Statins, cognition, and dementiasystematic review and methodological commentary. *Nature Reviews Neurology*, *11*(4), 220–229.

Prescribing data: Bnf codes. (2017). `https://www.thedatalab.org/blog/2017/04/prescribing-data-bnf-codes/`. (Accessed: 2022-01-18)

Pye, S. R., Sheppard, T., Joseph, R. M., Lunt, M., Girard, N., Haas, J. S., ... others (2018). Assumptions made when preparing drug exposure data for analysis have an impact on results: A n unreported step in pharmacoepidemiology studies. *Pharmacoepidemiology and drug safety*, *27*(7), 781–788.

Reps, J. M., Garibaldi, J. M., Aickelin, U., Gibson, J. E., and Hubbard, R. B. (2015). A supervised adverse drug reaction signalling framework imitating bradford hill's causality considerations. *Journal of Biomedical Informatics*, *56*, 356–368.

Reps, J. M., Garibaldi, J. M., Aickelin, U., Soria, D., Gibson, J., and Hubbard, R. (2013). Comparison of algorithms that detect drug side effects using electronic healthcare databases. *Soft Computing*, *17*, 2381–2397.

Reps, J. M., Garibaldi, J. M., Aickelin, U., Soria, D., Gibson, J. E., and Hubbard, R. B. (2014). Signalling paediatric side effects using an ensemble of simple study designs. *Drug Safety*, *37*, 163–170.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. *IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS*, *116*, 1–94.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer-Verlag.

Rothman, K., Greenland, S., and Lash, T. (2008). *Modern Epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins.

RxClass API. (2022). `https://lhncbc.nlm.nih.gov/RxNav/APIs/api-RxClass.getClassByRxNormDrugName.html`. (Accessed: 2022-01-18)

RxNorm Attributes. (2022). `https://www.nlm.nih.gov/research/umls/rxnorm/docs/appendix4.html`. (Accessed: 2022-01-18)

Ryan, P. B., Madigan, D., Stang, P. E., Marc Overhage, J., Racoosin, J. A., and Hartzema, A. G. (2012). Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership. *Statistics in medicine*, *31*(30), 4401–4415.

Ryan, P. B., and Schuemie, M. J. (2013). Evaluating performance of risk identification methods through a large-scale simulation of observational data. *Drug safety*, *36*, 171–180.

Ryan, P. B., Schuemie, M. J., Gruber, S., Zorych, I., and Madigan, D. (2013). Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug safety*, *36*, 59–72.

Ryan, P. B., Stang, P. E., Overhage, J. M., Suchard, M. A., Hartzema, A. G., DuMouchel, W., ... Madigan, D. (2013). A comparison of the empirical performance of methods for a risk identification system. *Drug safety*, *36*, 143–158.

Schuemie, M. J., Cepeda, M. S., Suchard, M. A., Yang, J., Tian, Y., Schuler, A., ... Hripcsak, G. (2020). How confident are we about observational findings in health care: a benchmark study. *Harv Data Sci Rev*, *2*(1), 10.

Schuemie, M. J., Coloma, P. M., Straatman, H., Herings, R. M., Trifirò, G., Matthews, J. N., ... others (2012). Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Medical care*, *50*, 890–897.

Schuemie, M. J., Madigan, D., and Ryan, P. B. (2013). Empirical performance of LGPS and LEOPARD: lessons for developing a risk identification and analysis system. *Drug safety*, *36*, 133–142.

Selby, D. (2021a). *doseminer: Extract drug dosages from free-text prescriptions* [Manual]. Retrieved from `https://cran.r-project.org/web/packages/doseminer/index.html` (R package version 0.1.2)

Selby, D. (2021b). *Web scraping for drug safety* [Manual]. Retrieved from `https://personalpages.manchester.ac.uk/staff/david.selby/rthritis/2021-11-05-web-scraping/`

Shah, A. (2021). *Rdiagnosislist: Manipulate SNOMED CT diagnosis lists* [Manual]. Retrieved from `https://cran.r-project.org/web/packages/Rdiagnosislist/index.html` (R package version 1.0)

Shin, H., Cha, J., Lee, C., Song, H., Jeong, H., Kim, J.-Y., and Lee, S. (2021). The 2011–2020 trends of data-driven approaches in medical informatics for active pharmacovigilance. *Applied Sciences*, *11*(5), 2249.

Shue, E., Liu, L., Li, B., Feng, Z., Li, X., and Hu, G. (2023). Empowering beginners in bioinformatics with chatgpt. *bioRxiv*, 2023–03.

Springate, D. A., Kontopantelis, E., Ashcroft, D. M., Olier, I., Parisi, R., Chamapiwa, E., and Reeves, D. (2014). ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PloS one*, *9*(6), e99825.

Suchard, M. A., Zorych, I., Simpson, S. E., Schuemie, M. J., Ryan, P. B., and Madigan, D. (2013). Empirical performance of the self-controlled case series design: lessons for developing a risk identification and analysis system. *Drug safety*, *36*, 83–93.

Suissa, S. (1995). The case-time-control design. *Epidemiology*, *6*(3), 248–253.

Takeuchi, Y., Shinozaki, T., and Matsuyama, Y. (2018). A comparison of estimators from self-controlled case series, case-crossover design, and sequence symmetry analysis for pharmacoepidemiological studies. *BMC medical research methodology*, *18*(1), 1–15.

Teneralli, R. E., Kern, D. M., Cepeda, M. S., Gilbert, J. P., and Drevets, W. C. (2021). Exploring real-world evidence to uncover unknown drug benefits and support the discovery of new treatment targets for depressive and bipolar disorders. *Journal of Affective Disorders*, *290*, 324–333.

Unified medical language system. (2022). `https://www.nlm.nih.gov/research/umls/index.html`. (Accessed: 2022-01-18)

BNF SNOMED mapping. (2022). `https://www.nhsbsa.nhs.uk/prescription-data/understanding-our-data/bnf-snomed-mapping`. (Accessed: 2022-01-18)

NHS data migration. (2020). `https://isd.digital.nhs.uk/trud/users/authenticated/group/0/pack/1/subpack/9/releases`. (Accessed: 2022-01-18)

UK SNOMED CT browser clinical edition. (2020). `https://snomedbrowser.com/`. (Accessed: 2022-01-18)

UK SNOMED CT drug extension. (2022). `https://isd.digital.nhs.uk/trud/users/authenticated/filters/0/categories/26/items/105/releases`. (Accessed: 2022-01-18)

van Aalst, R., Thommes, E., Postma, M., Chit, A., and Dahabreh, I. J. (2021). On the causal interpretation of rate-change methods: the prior event rate ratio and rate difference. *American Journal of Epidemiology*, *190*(1), 142–149.

Wang, S., Linkletter, C., Maclure, M., Dore, D., Mor, V., Buka, S., and Wellenius, G. A. (2011). Future-cases as present controls to adjust for exposure-trend bias in case-only studies. *Epidemiology*, *22*(4), 568-574.

Wickham, H. (2020a). *httr: Tools for working with URLs and HTTP* [Manual]. Retrieved from `https://cran.r-project.org/web/packages/httr/index.html` (R package version 1.4.2)

Wickham, H. (2020b). *rvest: Easily harvest (scrape) web pages* [Manual]. Retrieved from `https://cran.r-project.org/web/packages/rvest/index.html` (R package version 1.0.2)

Wolfram, S. (2023). *What is chatgpt doing... and why does it work?*

Yao, L., Zhang, Y., Li, Y., Sanseau, P., and Agarwal, P. (2011). Electronic health records: Implications for drug discovery. *Drug discovery today*, *16*(13-14), 594–599.

Yimer, B. B., Selby, D., Jani, M., Nenadic, G., Lunt, M., and Dixon, W. G. (2021a). *drugprepr: Prepare electronic prescription record data to estimate drug exposure* [Manual]. Retrieved from `https://cran.r-project.org/web/packages/drugprepr/index.html` (R package version 0.0.4)

Yimer, B. B., Selby, D. A., Jani, M., Nenadic, G., Lunt, M., and Dixon, W. G. (2021b). *Introduction to drugprepr* [Manual]. Retrieved from `https://cran.r-project.org/web/packages/drugprepr/vignettes/introduction.pdf`

Yu, M., Xie, D., Wang, X., Weiner, M. G., and Tannen, R. L. (2012). Prior event rate ratio adjustment: numerical studies of a statistical method to address unrecognized confounding in observational studies. *pharmacoepidemiology and drug safety*, *21*, 60–68.

Zhou, X., Bao, W., Gaffney, M., Shen, R., Young, S., and Bate, A. (2018). Assessing performance of sequential analysis methods for active drug safety surveillance using observational data. *Journal of Biopharmaceutical Statistics*, *28*(4), 668–681.