Foundational Model-aided Automatic High-throughput Drug Screening Using Self-controlled Cohort Study

Shenbo Xu
Institute for Data, Systens, and Society
Massachusetts Institute of Technology
Cambridge, MA 02142
xushenbo@mit.edu

Raluca Cobzaru
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02142
rcobzaru@mit.edu

Stan Finkelstein
Institute for Data, Systens, and Society
Massachusetts Institute of Technology
Cambridge, MA 02142
snfinkel@mit.edu

Roy Welsch
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02142
rwelsch@mit.edu

Kenney Ng
Center for Computational Health
IBM Research
Cambridge, MA 02142
Kenney.Ng@us.ibm.com

Abstract

The process of developing new drugs, from initial discovery to obtaining regulatory approval, has historically been neither cost-efficient, expeditious, nor free from risk. The growing availability of large-scale observational healthcare databases, combined with the rise of foundational models, offer an unparalleled opportunity to enable automatic high-throughput drug screening for both repurposing and pharmacovigilance. In this work, we present a general workflow for automatic high-throughput drug screening which estimates the association between various drug exposures and disease outcomes. We provide frameworks for parsing the accurate exposure length for each prescription from clinical texts and removing confounding relationships between drugs and diseases using bioinformatic mapping and foundational models. Using a self-controlled cohort study design, we tested the intention-to-treat association between 3,444 medications and 276 diseases across 6.6 million UK patients from the Clinical Practice Research Datalink (CPRD). Our analysis revealed 16,901 drug-disease pairs with significant risk reduction, indicating candidates for repurposing, as well as 11,089 pairs with significant risk increase which raise drug safety concerns. Our data-driven, nonparametric, hypothesis-generating, and automatic approach demonstrates the potential of foundational models in drug discovery and provides a scalable framework for drug repurposing that can be extended to other observational medical databases.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Approved treatment options for many diseases, such as cancer, Alzheimer's, or HIV, remain limited, with restricted drug targets, high costs, and long development times hindering the development of new therapies. As a result, there is considerable unmet demand for disease-modifying medications for various groups of disorders. Meanwhile, there are 3,000 medications currently prescribed in the UK, offering valuable opportunities to repurpose existing treatments for new indications. Aside from the potential for discovering new uses, existing drugs must be monitored for adverse drug reactions (ADR), also known as side effects, and other unintended consequences. In 2018, ADRs accounted for 5-8% of impromptu UK hospitalizations that resulted in 4-6% hospital beds filled, with an approximate annual bill of £1-2.5bn for the National Health Service (NHS) [1]. Longitudinal observational databases, including electronic health records (EHRs) and administrative claims, offer real-world insights into the relationships between drugs and clinical outcomes [2]. These types of data capture broad healthcare information, including physician diagnoses, therapies filled for patients, and lab tests, which have been actively used to conduct hypothesis-testing pharmacoepidemiology studies for causal effect estimation in clinical settings. In recent years, there has been increasing interest in adopting such databases to inform early drug development [3], identify novel treatment pathways [4], and discover unknown benefits [5] and side-effects of existing medications [6] in a fast, large-scale, data-driven, nonparametric, and high-throughput method.

Prior work employing identical study design has focused on specific clinical outcomes [7–10] or a particular drug class of interest [11] using US administrative claims data. However, other contexts of observational data, such as EHRs, remain underexplored. Moreover, the empirical performance of several study designs has been assessed as a tool for risk identification and analysis in healthcare data [12, 13]. Due to inadequate prescription information in claims data, a fixed 30-days gap between consecutive fills was utilized to calculate length of exposure. Previous applications also relied on manual removal of drugs confounded by indication and focused on drug-disease associations without relating target quantity to causal interpretation.

Additionally, the past two years have seen an explosive growth of artificial intelligence-generated content (AIGC) [14], especially through the release of the powerful large language model (LLM) ChatGPT-4 developed by OpenAI [15]. While ChatGPT is considered a disruptor to the healthcare industry [16], having demonstrated applicability to healthcare settings (e.g., [17]), its utility in drug discovery has not been widely explored [18, 19].

In this paper, we establish an automated framework for high-throughput drug screening on potential disease groups to detect beneficial clinical signals, leveraging foundational models to improve exposure length estimation, remove drug-indication confounding pairs, and provide causal-wise interpretations. We then apply this approach to identify drug-disease pairs with potential therapeutic benefits, offering a novel and scalable approach to drug repurposing. We also include results on pharmacovigilance in the Appendix.

2 Methods

2.1 Study Design

Owing to limitations of existing pharmacoepidemiology study designs, we focus on the self-controlled cohort for high-throughput drug screening [7–11]. As illustrated in Figure 1, a self-controlled cohort only utilizes new users of the drug of interest where individuals serve as their own controls. This setup handles potential confounding issues in the treatment allocation and ensures exposures are randomized automatically. This approach can be exemplified by studying the relationship of a drug-disease pair when all new drug-users are incorporated into the cohort. For each specific patient, equal person-time is allocated to the exposed period after initial prescription and to the unexposed period before first treatment. This arrangement is then replicated for available medications on possible diseases in the database, enabling a comprehensive analysis. Notably, simulation studies have demonstrated that the self-controlled cohort method yields less biased estimates and better predictive performance compared to other study designs [12, 13, 20, 21].

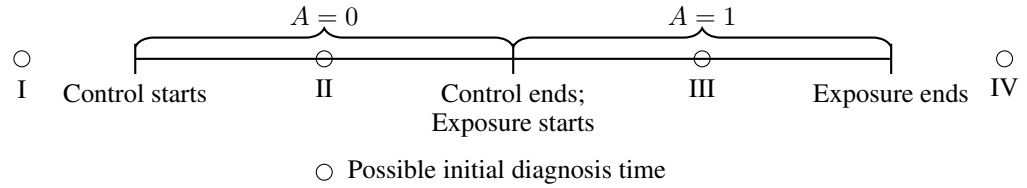

Figure 1: Illustration of self-controlled cohort study design. Disease incidence can take place before unexposure starts (I), during unexposure (II), during exposure (III), after exposure ends (IV), or never happens.

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 [22]. Patient follow-up starts in 1987 and ends at the earliest between mortality, transfer-out, or last collection of practice. The mean and standard deviation of follow-up are 16.77 years and 15.75 years. CPRD includes 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 the "therapy" table of CPRD. For each prescription, the table contains the patient id, "prodcode" for medicinal product, "eventdate" for prescription date, "qty" for total quantity prescribed, and "numdays" for duration entered by prescriber [23]. 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, we approximate it by dividing "qty" by the number of doses to be taken per day, also referred to as numeric daily dose (ndd). The "ndd" can be computed as $\mathrm{ndd} = \frac{\mathrm{DF} \times \mathrm{DN}}{\mathrm{DI}}$, where where DF represents the dose frequency (number of doses per day), DN the dose number (number of tablets to take each time), and DI the dose interval (number of days between doses). DF, DN, and DI can all be parsed from unstructured free text written by general practitioners following [24, 25] using `R` package `doseminer` [26]. To extend exposure period by reducing "ndd" when clinical texts inform a range of plausible values, we set DF to max, and DN and DI to min by drug.

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". 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. We overlook overlapping prescriptions due to enormous time-complexity when recursively adding overlap to the end of subsequent prescriptions 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` [27] while the last step leverages `data.table` to boost speed.

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, and "lcd" is last collection date GOL [23]. It follows that control start $=$ exposure start $-$ exposure time and 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 Definition

The first incidence of every disease and category is identified by using code lists phenotyped by validated bioinformatic algorithms from [28]. We test a total of 276 distinct diseases and 16 broad condition categories.

2.5 Removing Confounding Pairs

A self-controlled cohort study requires that initial exposure is not caused by indication. For example, if previous hypertension diagnosis (which happen to be both the indication and the clinical outcome at the same time) led to subsequent anti-hypertensive treatment, pre-exposure incidence rate will always be higher than post-exposure incidence rate. Then the spurious protective effect will appear because the first hypertension diagnosis often occurs before (and thus impacts) initial anti-hypertensive exposure. We can manually remove drug-indication combinations using subject-matter knowledge from previous studies focused on particular diseases [7, 9–11] or on specific drug classes [8] with clear relationship to the primary indication. However, since we aim to screen available drugs on possible diseases, manual removal is laborious, time-consuming, and prone to error. To address this issue, we propose a systematic framework for automatically identifying potential confounding pairs by leveraging prodcode-medcode associations with established relationships. This approach allows for efficient screening of drug-indication combinations across a wide range of diseases and medications.

Figure 2 demonstrates the medication-indication open loop starting from potential therapies (coded by prodcode) and ending at potential targets, or diagnoses (coded by medcode). 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. 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 [29]. In order to map prodcode to rxcui and MeSH to medcode, respectively, the Systematized Nomenclature of Medicine (SNOMED), an international organized terminology, is selected as the bridge. As the map between gemscript codes and SNOMED drug codes is not actively managed [30], 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 [31]. 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.

On the drug branch of Figure 2, we need to map MeSH code to medcode. As MeSH is US-based while medcode is UK-based, SNOMED is again chosen as the international link. Since SNOMED clinical codes cannot 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 deterministic mapping rules above, 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. This approach to prompting ChatGPT demonstrates the nuanced impact of prompt wording on AI responses. By allowing for an "unknown" option, we inadvertently encourage a more conservative response pattern, where the AI tends to default to "unknown" rather than committing to a "yes" or "no" answer. This behavior likely stems from the AI’s training to avoid making definitive statements when uncertainty exists. The observation highlights the importance of carefully crafting prompts to elicit the desired type of response, balancing between encouraging definitive answers and allowing for

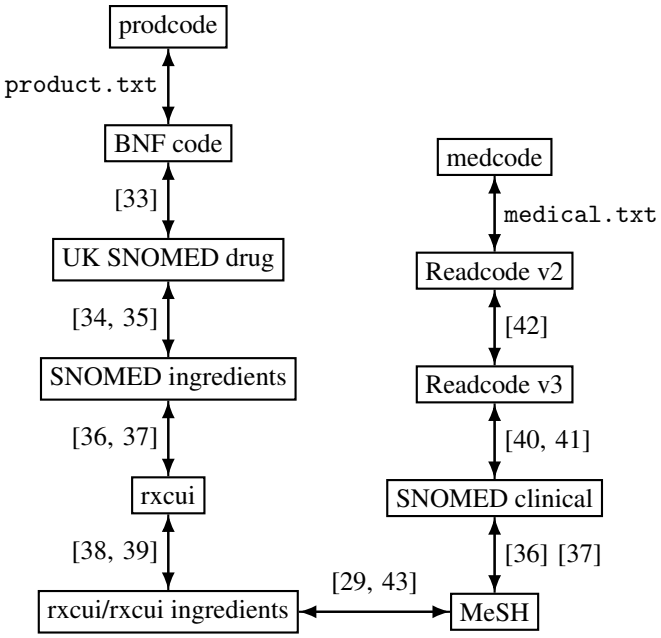

Figure 2: Drug indication map from prodcode to medcode. Solid boxes reveal specific coding system while dashed boxes contain sources of maps between adjacent coding systems along with `R` packages for extraction. If `R` package in a dashed box is missing, then the source of map are in machine-readable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

appropriate uncertainty. As noted by John [32], the art of prompt engineering involves understanding these subtle interactions between prompt structure and AI behavior, and tailoring the prompts to achieve the most useful and accurate responses for the task at hand.

After pulling out confounding by indication pairs, the remaining duplets are still subject to confounding 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 candidate 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 are 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 confounding by risk factors of all indications of the drug of interest.

Finally, for pharmacovigilance purposes, the drug-disease pairs may 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."

2.6 Causal Interpretation

To our knowledge, IRR in a self-controlled cohort study has not been clearly expressed in counterfactual language. In this section, we discuss a causal interpretation of IRR and its additive equivalent, the incident rate difference (IRD). It can be shown that the interpretability of these quantities relies on the 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 the assumption to various extent.

Suppose there are $a*$ exposures of interest, $j*$ outcomes of interest, and n_a units who have ever been exposed to treatment $a = 0, 1, \ldots, a*$. Let A_i be the exposure ($i = 1, 2, \ldots, n_a$) and the time of the first exposure be time 0. For treatment a, assume $T_{ia,\text{pre}}$ is the control period before time 0 and $T_{ia,\text{post}}$ is the exposed period after time 0. 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) and the causal incidence rate difference (IRD) can be defined as

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

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 assumptions)**.** *Positivity holds for the population in the following periods: pre-exposed period* $E(T_{ia,pre}) > 0$*, post-exposed period* $E(T_{ia,post}) > 0$*, and pretreatment observed outcomes for the population* $E(Y_{ija,pre}) > 0$*. Note that causal IRD does not require* $E(Y_{ija,pre}) > 0$*.*

Assumptions 1 and 2 are crucial to identify IRR/IRD but are both empirically unverifiable. Assumption 2 is similar to the parallel trends assumption in difference-in-differences [44] and rate-change assumptions in calibrated self-controlled cohort studies [45]. 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 lag-time are beyond the scope of this work [46, 47].

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

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

and causal IRD can be identified and estimated as

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

Suppose the IRR is a ratio between two rates with Poisson distribution [48]. Then, the closed-form confidence interval can be computed as

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

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 [49].

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

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

When $\text{bias}_{\text{IRR}} = 1$, $\widehat{\text{IRR}}_{ja}$ becomes an unbiased estimator for IRR^a_j; when $0 < \text{bias}_{\text{IRR}} < 1$, $\widehat{\text{IRR}}_{ja}$ serves as an upper bound for IRR^a_j; whereas when $\text{bias}_{\text{IRR}} > 1$, $\widehat{\text{IRR}}_{ja}$ 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, and 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 occurrences in the data should be more than 100; (3) the number of outcomes during both control and exposure period is larger than 30. Depending on the specification, the analyses require 208-256 CPU cores and 2-3TB memory, with execution times ranging from several to less than 10 hours.

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 1. The results presented in Table 1 reveal several promising drug candidates for potential repurposing in dementia prevention or treatment. Notably, all listed medications show significant protective effects, with

upper bounds of the 95% confidence intervals for IRR well below 1. These drugs, ranging from common over-the-counter medications like paracetamol and folic acid to prescription drugs such as omeprazole and latanoprost, are typically used for diverse conditions including pain relief, acid reflux, and glaucoma. The consistent protective signals across this varied group of medications underscore the potential for repurposing existing drugs in novel approaches to dementia management, opening up exciting avenues for further research and clinical investigation.

Table 1: Summary of self-controlled cohort study to repurpose multiple drugs candidates for dementia. For each drug, "Upper" denotes the upper bound in the 95% confidence interval of IRR, "N exposed" the number of participants exposed to the drug, "Exposure mean" the length of the exposure period (in days), and "Exposure SD" the standard deviation of the exposure period.

Drug	Upper	N exposed	Exposure mean	Exposure SD
chloroform / magnesium oxide light / magnesium sulfate dried / sodium hydroxide	0.51	26,189	28.31	5.76
folic acid	0.53	280,096	28.43	5.53
omeprazole	0.53	1,408,560	28.86	4.72
dipyridamole	0.54	81,372	28.21	5.89
paracetamol	0.56	1,455,955	28.53	5.40
promethazine hydrochloride	0.56	77,760	29.31	3.71
quinine bisulfate	0.58	308,277	29.03	4.36
latanoprost	0.61	101,823	28.36	5.64
permethrin	0.62	118,084	321.29	96.81

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 2.

4 Discussion

4.1 Strengths

There are several strengths of this work. The self-controlled cohort study allows subjects to act as their own control, automatically accounting for all time-fixed covariates (whether observed or not) sycg as genetics. Moreover, by overshooting risk reduction and risk augmentation, it avoids the pitfalls of narrow confidence intervals induced by underestimated variability and erroneous findings resulting from multiple comparisons often seen in other cohort studies. While some potential effects may be missed, the lack of significant discoveries indicates that the estimated associations are not substantial, rather than entirely absent. Potential false positives (type 1 errors) are less concerning for hypothesis-screening studies, we we are targeting candidates for futher research instead of confirming causal effects.

Additionally, we defined causal IRR/IRD and outlined conditions for identification, with the key distinction from exchangeability-based external control group methods being the common intensity Assumption 2. Although this assumption is challenging in practice, the estimated IRR/IRD can still serve as upper bounds for causal IRR/IRD if the control/exposed periods are long enough for aging to become a dominant factor that boosts posttreatment incidence. We focused on Imbens' approach [50] which targets average treatment effects for sensitivity analysis and does not require detailed subject-matter knowledge of unmeasured time-varying confounders. Moreover, IRR/IRD conditioned on time-fixed covariates can be readily defined and identified by some modifications to the assumptions for unconditional IRR/IRD.

Accurate exposure lengths are critical for capturing clinical events and computing incidence rates in self-controlled cohort studies. Since prescription lengths are not uniform across medication, patient, practice, and region, setting a fixed value for all prescriptions would result in biased IRR/IRD

estimates. We computed more accurate exposure lengths for each prescription by parsing clinical texts, using common dosage information in CPRD.

The major drawback in self-controlled cohort studies is intrinsic confounding due to indications, contraindications, comorbidities, complications, and off-label uses, where temporal sequences are predetermined by existing clinical guidelines or natural connections between diseases. Due to medical ontology incompatibilities between the UK and the US, only mappable drug-indication pairs can be removed and often requiring manual input from physicians. Our application of foundational models not only addresses unmappable confounding by indication and other types of drug-disease relationships, but also enables the creation of new high-quality ontologies and extraction of clinical information from text data, adding value to the fields of bioinformatics and pharmacoepidemiology [51]. Finally, our framework demonstrates computational efficiency, allowing large-scale screening of rich databases such as CPRD at a 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 short. In such cases, the incidence rates before and after treatment are expected to be similar, as aging is not an essential factor with short exposure times. However, this was not a major issue for many drug-disease pairs found in the study, as patients with shorter exposure times have a lesser contribution to capturing outcomes. Chronic diseases also pose challenges due to their gradual onset and delayed formal diagnosis, which can result in temporal misclassifications and erroneously increased estimated risks, potentially reducing true negatives for repurposing intentions.

Clinical outcomes are phenotyped using diagnosis codes based on established studies, rather than subjective definition of conditions. We avoid under-recording in CPRD data by not requiring multiple diagnoses for the same condition to identify clinical outcomes, which may result in false positives due to single exclusion, misdiagnosis, or misclassification. If these false positive cases are non-differential with respect to the treatment, then the results should lie around the null which cannot be explained by the directional effects found in the analysis. Additionally, this study is based on UK primary care data, which may impact its generalizability to other countries with different treatment guidelines.

Another limitation is the potential for inaccuracies in ChatGPT responses used to identify known drug-disease associations, as foundational models are known to occasionally produce "hallucinations"– confident but incorrect answers. This could introduce false positives or negatives in the analysis. Future work in this area should implement validation steps such as cross-referencing ChatGPT outputs with curated databases or developing ensemble approaches that combine foundational model-generated insights with traditional bioinformatics methods to mitigate the risks of hallucination errors and enhance the reliability of pharmacovigilance findings.

5 Conclusion

The growing availability of observational databases enables the detection of unknown benefits via large-scale in-silico drug screening, which helps to address the large unmet medical needs for effectove disease-modifying therapies. We utilized a self-controlled cohort study to assess the association between marketed drug initiation and disease onset in millions of patients using UK primary care data from CPRD. We determined accurate exposure periods using unstructured text analysis. To remedy built-in selection bias issues of the self-controlled cohort study, we discard drug-disease pairs based on cross-ontology maps and insights from ChatGPT. We also offered causal-wide interpretation of incidence rate contrasts along with an Imbens-type sensitivity analysis on the critical common intensity assumption. After screening for positive signals, our approach identified 16,901 drug-disease pairs with reduced risk as potential candidates for repurposing. The results of this large-scale analysis can help generate hypotheses for subsequent observational, preclinical, and clinical research, which would further the validity and efficacy of our findings. The general workflow of this work demonstrates the potential of AIGC in bioinformatics and pharmacoepidemiology, and can be easily applied to other observational healthcare databases.

References

[1] Sue Jordan, Patricia A Logan, Gerwyn Panes, Mojtaba Vaismoradi, and David Hughes. Adverse drug reactions, power, harm reduction, regulation and the adre profiles. *Pharmacy*, 6(3):102, 2018.

[2] Hyunah Shin, Jaehun Cha, Chungchun Lee, Hyejin Song, Hyuntae Jeong, Jong-Yeup Kim, and Suehyun Lee. The 2011–2020 trends of data-driven approaches in medical informatics for active pharmacovigilance. *Applied Sciences*, 11(5):2249, 2021.

[3] Shuchi Mittal, Kjetil Bjørnevik, Doo Soon Im, Adrian Flierl, Xianjun Dong, Joseph J Locascio, Kristine M Abo, Elizabeth Long, Ming Jin, Bing Xu, et al. β2-adrenoreceptor is a regulator of the α-synuclein gene driving risk of Parkinson's disease. *Science*, 357(6354):891–898, 2017.

[4] Lixia Yao, Yiye Zhang, Yong Li, Philippe Sanseau, and Pankaj Agarwal. Electronic health records: Implications for drug discovery. *Drug discovery today*, 16(13-14):594–599, 2011.

[5] Benjamin S Glicksberg, Li Li, Rong Chen, Joel Dudley, and Bin Chen. Leveraging big data to transform drug discovery. *Bioinformatics and Drug Discovery*, pages 91–118, 2019.

[6] Xiaofeng Zhou, Warren Bao, Mike Gaffney, Rongjun Shen, Sarah Young, and Andrew Bate. Assessing performance of sequential analysis methods for active drug safety surveillance using observational data. *Journal of Biopharmaceutical Statistics*, 28(4):668–681, 2018.

[7] David M Kern, M Soledad Cepeda, Simon Lovestone, and Guy R Seabrook. 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, 2019.

[8] David M Kern, M Soledad Cepeda, Christopher M Flores, and Gayle M Wittenberg. 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, 2021.

[9] Rachel E Teneralli, David M Kern, M Soledad Cepeda, James P Gilbert, and Wayne C Drevets. 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, 2021.

[10] David M Kern, Rachel E Teneralli, Christopher M Flores, Gayle M Wittenberg, James P Gilbert, and M Soledad Cepeda. 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, 2022.

[11] M Soledad Cepeda, David M Kern, Guy R Seabrook, and Simon Lovestone. Comprehensive real-world assessment of marketed medications to guide parkinson's drug discovery. *Clinical Drug Investigation*, 39:1067–1075, 2019.

[12] Patrick B Ryan, Martijn J Schuemie, Susan Gruber, Ivan Zorych, and David Madigan. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug safety*, 36:59–72, 2013.

[13] Patrick B Ryan and Martijn J Schuemie. Evaluating performance of risk identification methods through a large-scale simulation of observational data. *Drug safety*, 36:171–180, 2013.

[14] OpenAI. Gpt-4 technical report, 2023.

[15] OpenAI. Chatgpt-4. `https://chat.openai.com/chat`, 2023. Accessed: 2023-03-20.

[16] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

[17] Peter Lee, Carey Goldberg, and Isaac Kohane. *The AI Revolution in Medicine: GPT-4 and Beyond*. Pearson, 2023.

[18] Jesus de la Fuente Cedeño, Guillermo Serrano, Uxía Veleiro, Mikel Casals, Laura Vera, Marija Pizurica, Antonio Pineda-Lucena, Idoia Ochoa, Silve Vicent, Olivier Gevaert, et al. Towards a more inductive world for drug repurposing approaches. In *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*.

[19] Sepehr Asgarian, Sina Akbarian, and Jouhyun Jeon. All you need is love: Large optimized vector embeddings network for drug repurposing. In *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*.

[20] Martijn J Schuemie, David Madigan, and Patrick B Ryan. Empirical performance of LGPS and LEOPARD: lessons for developing a risk identification and analysis system. *Drug safety*, 36: 133–142, 2013.

[21] Martijn J Schuemie, M Soledad Cepeda, Marc A Suchard, Jianxiao Yang, Yuxi Tian, Alejandro Schuler, Patrick B Ryan, David Madigan, and George Hripcsak. How confident are we about observational findings in health care: a benchmark study. *Harv Data Sci Rev*, 2(1):10, 2020.

[22] Achim Wolf, Daniel Dedman, Jennifer Campbell, Helen Booth, Darren Lunn, Jennifer Chapman, and Puja Myles. Data resource profile: clinical practice research datalink (cprd) aurum. *International journal of epidemiology*, 48(6):1740–1740g, 2019.

[23] *CPRD GOLD Data Specification*, 2021. URL `https://cprd.com/sites/default/files/CPRD%20GOLD%20Full%20Data%20Specification%20v2.4.pdf`. version 2.4.

[24] George Karystianis, Therese Sheppard, William G Dixon, and Goran Nenadic. 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, 2015.

[25] Ghada Alfattni, Niels Peek, Goran Nenadic, and Fergus Caskey. Integrating text analytics and statistical modelling to analyse kidney transplant immune suppression medication in registry data. *International Journal of Population Data Science*, 1(1), 2022.

[26] David Selby. *doseminer: Extract Drug Dosages from Free-Text Prescriptions*, 2021. URL `https://cran.r-project.org/web/packages/doseminer/index.html`. R package version 0.1.2.

[27] Belay Birlie Yimer, David Selby, Meghna Jani, Goran Nenadic, Mark Lunt, and William G. Dixon. *drugprepr: Prepare Electronic Prescription Record Data to Estimate Drug Exposure*, 2021. URL `https://cran.r-project.org/web/packages/drugprepr/index.html`. R package version 0.0.4.

[28] Valerie Kuan, Spiros Denaxas, Arturo Gonzalez-Izquierdo, Kenan Direk, Osman Bhatti, Shanaz Husain, Shailen Sutaria, Melanie Hingorani, Dorothea Nitsch, Constantinos A Parisinos, et al. 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, 2019.

[29] RxClass API. `https://lhncbc.nlm.nih.gov/RxNav/APIs/api-RxClass.getClassByRxNormDrugName.html`, 2022. Accessed: 2022-01-18.

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

[31] Prescribing data: Bnf codes. `https://www.thedatalab.org/blog/2017/04/prescribing-data-bnf-codes/`, 2017. Accessed: 2022-01-18.

[32] Ibrahim John. *The Art of Asking ChatGPT for High-Quality Answers*. 2023.

[33] BNF SNOMED mapping. `https://www.nhsbsa.nhs.uk/prescription-data/understanding-our-data/bnf-snomed-mapping`, 2022. Accessed: 2022-01-18.

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

[35] Anoop Shah. *Rdiagnosislist: Manipulate SNOMED CT Diagnosis Lists*, 2021. URL https://cran.r-project.org/web/packages/Rdiagnosislist/index.html. R package version 1.0.

[36] Unified medical language system. https://www.nlm.nih.gov/research/umls/index.html, 2022. Accessed: 2022-01-18.

[37] Yvon Awuklu. *getUMLS: Query the UMLS metathesaurus*, 2021. URL https://github.com/yvoawk/getUMLS/releases/tag/v0.1.0. R package version 0.1.0.

[38] RxNorm Attributes. https://www.nlm.nih.gov/research/umls/rxnorm/docs/appendix4.html, 2022. Accessed: 2022-01-18.

[39] Jeroen Ooms, Duncan Temple Lang, and Lloyd Hilaiel. *jsonlite: A Simple and Robust JSON Parser and Generator for R*, 2022. URL https://cran.r-project.org/web/packages/jsonlite/index.html. R package version 1.7.3.

[40] UK SNOMED CT browser clinical edition. https://snomedbrowser.com/, 2020. Accessed: 2022-01-18.

[41] Hadley Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2020. URL https://cran.r-project.org/web/packages/rvest/index.html. R package version 1.0.2.

[42] NHS data migration. https://isd.digital.nhs.uk/trud/users/authenticated/group/0/pack/1/subpack/9/releases, 2020. Accessed: 2022-01-18.

[43] Hadley Wickham. *httr: Tools for Working with URLs and HTTP*, 2020. URL https://cran.r-project.org/web/packages/httr/index.html. R package version 1.4.2.

[44] Alberto Abadie. Semiparametric difference-in-differences estimators. *The review of economic studies*, 72(1):1–19, 2005.

[45] Robertus van Aalst, Edward Thommes, Maarten Postma, Ayman Chit, and Issa J Dahabreh. On the causal interpretation of rate-change methods: the prior event rate ratio and rate difference. *American Journal of Epidemiology*, 190(1):142–149, 2021.

[46] Bas A In'T Veld, Annemieke Ruitenberg, Albert Hofman, Lenore J Launer, Cornelia M van Duijn, Theo Stijnen, Monique MB Breteler, and Bruno HC Stricker. Nonsteroidal antiinflammatory drugs and the risk of alzheimer's disease. *New England Journal of Medicine*, 345(21):1515–1521, 2001.

[47] Melinda C Power, Jennifer Weuve, A Richey Sharrett, Deborah Blacker, and Rebecca F Gottesman. Statins, cognition, and dementia—systematic review and methodological commentary. *Nature Reviews Neurology*, 11(4):220–229, 2015.

[48] PL Graham, Kerrie Mengersen, and AP Morton. Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. *Statistics in medicine*, 22(12):2071–2083, 2003.

[49] Kalimuthu Krishnamoorthy and Jessica Thomson. A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*, 119(1):23–35, 2004.

[50] Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

[51] Evelyn Shue, Li Liu, Bingxin Li, Zifeng Feng, Xin Li, and Gangqing Hu. Empowering beginners in bioinformatics with chatgpt. *bioRxiv*, pages 2023–03, 2023.

A Appendix

Table 2: Summary of self-controlled cohort study for multiple drug-disease pairs for pharmacovigilance. For each drug-disease pair, "Lower" denotes the lower bound in the 95% confidence interval of IRR, "N exposed" the number of participants exposed to the drug, "Exposure mean" the length of the exposure period (in days), and "Exposure SD" the standard deviation of the exposure period.

Drug	Disease	Lower	N exposed	Exposure mean	Exposure SD
atenolol	primary pulmonary hypertension	3.90	641,137	867.89	1345.64
naproxen	multiple myeloma and malignant plasma cell neoplasms	3.54	994,272	327.66	91.45
nicorandil	anorectal fistula	3.37	82,631	753.49	1100.67
simvastatin	aspiration pneumonitis	3.21	1,057,217	956.02	1199.80
salbutamol	dilated cardiomyopathy	2.95	1,160,084	290.81	127.41
cyclopenthiazide / potassium chloride	dermatitis	2.91	17,013	376.98	372.25
chamomile extract	menorrhagia and polymenorrhoea	2.88	8,918	327.82	86.59
malathion	lichen planus	2.86	82,966	132.14	106.52
metformin	primary pulmonary hypertension	2.69	358,596	984.75	1251.41
aciclovir	trigeminal neuralgia	2.68	382,413	332.59	84.53

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper provides an automated framework for high-throughput drug screening on potential disease groups, leveraging foundational models to improve exposure length estimation (section 2.3), remove drug-indication confounding pairs (section 2.5), and provide causal-wide interpretations (section 2.6). The study design is outlined in section 2.1 and the results of our analysis are included in section 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4.2 is dedicated to discussing the potential limitations of our study.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We state the theoretical assumptions behind IRR/IRD identification in section 2.6, which assists in the causal interpretation of our results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper introduces a framework for screening drug-disease pairs in observational databases for potential repurposing or pharmacovigilance candidates. While some aspects of the study are dataset-specific and are not directly applicable to other clinical databases (e.g., computing the drug dosages from CPRD columns), we provide details on each step of the screening process that can be reproduced in other experimental studies.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The CPRD data used in this study is proprietary and cannot be accessed without ISAC approval.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The study design do not require any model thus no sample splitting. Prompts for foundational models are described and discussed in detail.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide the 95% confidence interval upper bound, number of datapoints, exposure length mean, and exposure length SD, which are important to assess the significance of our IRR/IRD estimations, in Table 1 and Table 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed in section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The use of CPRD data complied with the Institutional Review Board (IRB) process, ensuring the ethical use of de-identified patient health records. This is in alignment with the NeurIPS Code of Ethics.]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The use of foundational models for drug repurposing and pharmacovigilance enables new avenues for automated, fast, and economical signal identification from existing clinical data. However, this work is subject to limitations and should only be used to inform future confirmatory studies and clinical investigations.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: As we do not release the data and models from CPRD, which is subject to strict approval guidelines by ISAC, the risk of misuse by researchers applying our general framework is extremely low.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets are properly credited and the license and terms of use explicitly mentioned are properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not directly crowdsourse or perform research with human subjects beyond the approved use of past observational data as approved by ISAC.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The CPRD data was collected used in accordance with IRB. We did not contribute to the collection or processing of CPRD data.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.