Received: 23 February 2021 | Revised: 7 April 2021 | Accepted: 22 May 2021

DOI: 10.1002/psp4.12673

ARTICLE

Random control selection for conducting high-throughput adverse drug events screening using large-scale longitudinal health data

Chien-Wei Chiang[1] | **Penyue Zhang**[2] | **Macarius Donneyong**[3] | **You Chen**[4] | **Yu Su**[5] | **Lang Li**[1]

[1]Department of Biomedical Informatics, Ohio State University, Columbus, Ohio, USA

[2]Department of Biostatistics and Health Data Science, Indiana University, Bloomington, Indiana, USA

[3]Division of Outcomes and Translational Sciences, College of Pharmacy, Ohio State University, Columbus, Ohio, USA

[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[5]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA

Correspondence
Pengyue Zhang, Department of Biostatistics and Health Data Science, Indiana University, 410 W 10th St., Indianapolis, IN, 46202 USA.
Email zhangpe@iu.edu

Funding information
No funding was received for this work.

Abstract

Case-control design based high-throughput pharmacoinformatics study using large-scale longitudinal health data is able to detect new adverse drug event (ADEs) signals. Existing control selection approaches for case-control design included the dynamic/super control selection approach. The dynamic/super control selection approach requires all individuals to be evaluated at all ADE case index dates, as the individuals' eligibilities as control depend on ADE/enrollment history. Thus, using large-scale longitudinal health data, the dynamic/super control selection approach requires extraordinarily high computational time. We proposed a random control selection approach in which ADE case index dates were matched by randomly generated control index dates. The random control selection approach does not depend on ADE/enrollment history. It is able to significantly reduce computational time to prepare case-control data sets, as it requires all individuals to be evaluated only once. We compared the performance metrics of all control selection approaches using two large-scale longitudinal health data and a drug-ADE gold standard including 399 drug-ADE pairs. The F-scores for the random control selection approach were between 0.586 and 0.600 compared to between 0.545 and 0.562 for dynamic/super control selection approaches. The random control selection approach was ~ 1000 times faster than dynamic/super control selection approach on preparing case-control data sets. With large-scale longitudinal health data, a case-control design-based pharmacoinformatics study using random control selection is able to generate comparable ADE signals than the existing control selection approaches. The random control selection approach also significantly reduces computational time to prepare the case-control data sets.

Chien-Wei Chiang and Pengyue Zhang contributed equally to this work.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Large-scale longitudinal health data and adverse drug event (ADE) phenotyping algorithms have become increasingly available. Traditional methods are highly computationally intensive to conduct high-throughput ADE screening using large-scale longitudinal health data.

WHAT QUESTION DID THIS STUDY ADDRESS?

We propose a computationally efficient control selection approach to conduct case-control design based on high-throughput ADE screening using large-scale longitudinal health data.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

A case-control design-based pharmacoinformatics study using randomly selected index dates as controls (i.e., random control selection) has comparable or higher performance metrics compared with existing control selection approaches, whereas the random control selection approach is able to significantly reduce the time.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

Using the random control selection approach, a case-control design-based pharmacoinformatics study can be upscaled to screen several hypotheses in short period of time (e.g., 15 min), and identify single drugs and drug combinations with increased ADE risks.

INTRODUCTION

Adverse drug events (ADEs), the unintended pharmacological consequences of correctly administered dugs, are a significant challenge for healthcare practice. Currently, in the United States, ADEs cause ~ 125,000 hospital admissions each year, complicate 53% hospital stays, and cause up to 4.6% of deaths.[1–3] Many serious ADEs cannot be detected prior to the drug approval. For instance, in the United States, the times from approval to withdrawal due to safety concerns were 3.4 years for valdecoxib, 4.7 years for tegaserod, and 5.4 years for efalizumab.[4] Traditionally, a pharmacoepidemiological study has been used to investigate prespecified ADE hypothesis from real-world health data. For instance, a pharmacoepidemiological study can be driven by suspicious ADE case reports. Unlike a pharmacoepidemiological study, a pharmacoinformatic study is not driven by any prespecified hypothesis. A pharmacoinformatic study is a discovery-driven approach.[5] It screens signals from a large number of drug-ADE pairs.[6] Thus, a pharmacoinformatic study is able to generate ADE hypotheses for the subsequent pharmacoepidemiological studies, and hence accelerates translational ADE research. Currently, regulatory agencies collect postmarket ADE reports through Spontaneous Reporting Systems (SRSs) for identifying ADE hypothesis. An SRS report usually includes drug usages, ADE outcomes, and other information (i.e., patient demographics). Using SRS, a pharmacoinformatic study is able to screen ADE signals under the case-control design setting, as the reports can be summarized into a two-by-two contingency table by drug status (yes/no) and ADE status (yes/no).

Pharmacoinformatic studies have successfully identified ADE signals from SRS databases.[7] Pharmacoinformatic approaches based on two-by-two contingency tables (i.e., the case-control design setting) are also known as disproportion analysis (DPA), as they measure ADE signal by the outcome (i.e., total count of a drug-ADE pair) to expectation (i.e., expected count of a drug-ADE pair assuming no association) ratio. Frequentist DPA approaches include the proportional reporting ratio (PRR) and the reporting odds ratio (ROR).[8,9] The Empirical Bayesian geometric mean (EBGM) is an empirical Bayesian DPA approach and the information component (IC) is a Bayesian DPA approach.[10,11] Zhang et al. proposed a three-component mixture model (3CMM), which provided false discovery rate estimation for DPA signals.[12] In addition to DPA approaches, multivariable approaches, such as multiple logistic regression or regulated logistic regression, have been used to adjust potential confounding variables (i.e., comedications).[13] All these pharmacoinformatic studies have their own validations, and many promising discoveries have been successfully validated.[7,14]

Pharmacoinformatic studies have also identified ADE signals from longitudinal health data including

electronic health record (EHR) data and administrative claims data.[15–17] Unlike SRS, EHR data and administrative claims data contain individual level and longitudinal information, including clinical outcomes (i.e., diagnoses) and medications (i.e., pharmacy prescriptions/claims). Compared to SRS, EHR data and administrative claims data contain more population groups (i.e., individuals without any ADE), variables (i.e., health conditions other than ADE), and detailed temporal information. Although the utilization of this additional information can improve ADE screening, pharmacoinformatic studies using longitudinal health data require sophisticated epidemiological study designs, such as the case-control design.[18] Under the case control-design, the aforementioned pharmacoinformatic approaches for SRS (e.g., DPAs) can be directly applied to longitudinal health data. For instance, Wang et al. identified drug combinations with higher myopathy (i.e., a common muscular ADE) risks from an EHR database using the case-control design.[16]

Although pharmacoinformatic studies generated valuable ADE signals from longitudinal health data, they shall be expanded to investigate: (i) more large-scale longitudinal health databases, and (ii) much larger numbers of ADEs.[19,20] First, large-scale longitudinal health databases become more common and available. For instance, the MarketScan commercial claims and encounters database contains over 40 million patients' information per year,[21] and it has been cited more than 10,000 times according to the Google Scholar. Second, informatic resources allow a high-throughput pharmacoinformatics study to screen a large number of ADEs. For instance, algorithms to annotate different coding systems allow more than a hundred ADEs to be identified from longitudinal health data.[22–24] Currently, the development of large-scale longitudinal health data and informatics resources facilitate a high-throughput pharmacoinformatic study using large-scale longitudinal health data. However, the control selection process in the case-control design requires a tremendous amount of time for data preparation. For instance, under the existing incidence density sampling approach, at each ADE case index date (i.e., the health encounter date with an ADE diagnosis), other individuals who had not yet developed the ADE were eligible as controls.[25,26] Thus, the incidence density sampling approach requires all individuals to be evaluated at all ADE case index dates. Such a process requires a significant amount of computational time. If the data set contains a few million individuals, our experiences show that the incidence density sampling approach may require approximately 1 week to prepare the case-control data set for a common ADE using a standard computer. Additionally, using the incidence density sampling approach, the selected controls for one ADE cannot be used for another ADE, as the control selection approach depends on the history of the ADE. Thus, the computational time to prepare case-control data sets is further increased for investigating a large number of ADEs. For over a hundred ADEs, the projected time to prepare all case-control data sets is over 1 month on a computer cluster or over 1 year on a standard computer. The extraordinarily high computational time to prepare case-control data sets is a significant challenge, which cannot be addressed by existing approaches.

In this study, we propose a random sampling approach (i.e., random control selection) for case-control design, which is computationally efficient for investigating a large number of ADEs using large-scale longitudinal health data. To reduce the computational time, we propose to select a random control pool by using random generated control index dates. The proposed approach is able to significantly reduce the computational time, as it only requires all individuals to be evaluated once. Additionally, the random control pool can be used to prepare multiple case-control data sets without additional computational time. On the contrary, controls for one ADE cannot be used as controls for other ADEs under existing approaches. Thus, the random control selection approach is more computationally efficient for high throughput ADE screening. We used two large-scale longitudinal health data sets and the Observational Medical Outcomes Partnership (OMOP) gold standard[7] to evaluate the performance metrics of the proposed random control selection approach and existing control selection approaches under the case-control design.

METHODS

Control selection approaches

As individuals were enrolled in different dates in longitudinal health data, the case-control design required a baseline period (i.e., 3- or 6-month). The case index date was defined as the date of health encounter with an ADE diagnosis given no ADE diagnosis in the baseline period prior to the encounter date. Under the case-control design, control index dates were selected for each of the case index dates. For instance, the control index date could be the same date as the corresponding case index date (i.e., matching by ADE case index date), as long as the selected individual was eligible as a control at the case index date. Noting that a control index date should have duration of enrollment longer than the baseline period prior to the control index date. Thus, an individual's eligibility as a control changed over time depending on the individual's enrollment history. We investigated the drug exposures prior to all case index dates and control index dates. The existing control selection approaches and the proposed random control selection approach are illustrated in Figure 1, and are summarized below:

- Dynamic (i.e., the incidence density sampling) control selection (Figure 1a): for an ADE, individuals eligible as control at a case index date should have duration of enrollment longer than the baseline period and had not yet developed the ADE. Control index dates were matched to the corresponding case index dates.[26] Control index dates should be separately generated for different ADEs, as the control selection process for a specific ADE depends on the ADE's history.
- Super control selection (Figure 1a): for an ADE, individuals eligible as controls at a case index date should have duration of enrollment longer than the baseline period and had never developed the ADE during the entire enrollment period. Control index dates were matched to the corresponding case index dates.[25] Control index dates should be separately generated for different ADEs, as the control selection process for a specific ADE depends on the ADE's history.
- Random control selection (Figure 1b): randomly selected index dates with duration of enrollment longer than the baseline period were gathered as a control pool. Control index dates were not matched to the case index dates. Control index dates in the control pool could be used for all ADEs, as the control selection process did not depend on the ADEs' histories.

In a short summary, the proposed random control selection approach relaxed two restrictions: (i) matching by case index date, and (ii) matching by ADE history.

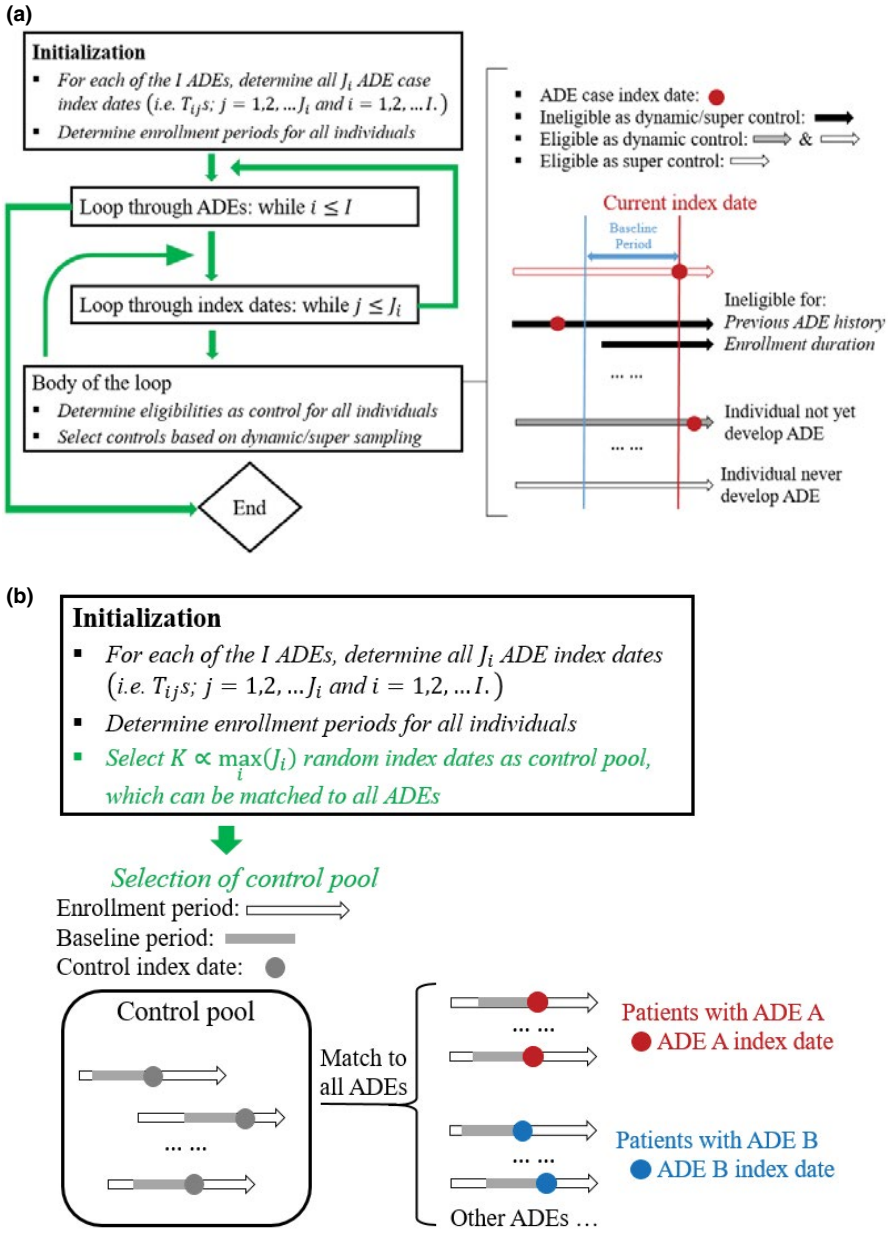

FIGURE 1 (a) Algorithm for dynamic/super control selection approach. (b) Algorithm for random control selection approach. ADE, adverse drug event

Data preparation

We used two large-scale longitudinal health databases in this study. The first database was the MarketScan Commercial Claims and Encounters database from 2012 to 2017. The MarketScan database included ~ 43 million individuals per year. The second database was the Indiana Network for Patient Care Common Data Model (INPC-CDM) from 2004 to 2015. The INPC-CDM database included ~ 5 million individuals per year. Both databases included individual-level demographic information (i.e., age and gender), administrative information (i.e., date of hospital visit), diagnoses, procedures, and pharmacy records. The MarketScan database used International Classification of Disease (ICD)-9/ICD-10 codes to record diagnoses and National Drug Codes (NDCs) to record pharmacy claims. The INPC-CDM database used OMOP concept IDs to record diagnoses and RxNorm to record pharmacy prescriptions. More details of the MarketScan database are presented in the Supplementary Data and Codes.

Our outcomes were acute myocardial infarction, acute renal failure, acute liver injury, and gastrointestinal bleeding. These ADEs were identified by using ICD-9/ICD-10 codes (algorithm given in Table S1). For each ADE, we defined a case as: an ADE diagnosis after a 180-baseline period (Figure S1). For instance, the first ADE diagnosis of an individual after 180-day enrollment was considered as a case; a subsequent ADE diagnosis of an individual was considered as a case if all previous ADEs were diagnosed 180 days prior to the current diagnosis. From the MarketScan database, we identified 203,797 acute myocardial infarction cases, 295,956 acute renal failure cases, 227,755 liver injury cases, and 137,420 gastrointestinal bleeding cases. From INPC data, we identified 137,439 acute myocardial infarction cases, 165,469 acute renal failure cases, 200,956 liver injury cases, and 235,056 gastrointestinal bleeding cases.

We conducted case-control designs for all four ADEs using the two databases. We used the ADE diagnosis dates as the case index dates. We generated the control index dates by using the dynamic control selection approach, super control selection approach, and random control selection approach. Similar as the cases, individuals eligible as controls must be enrolled 180 days prior to the control index dates (Figure S1). In addition, we also applied gender and age matching for all three control selection approaches, and we fixed the case-control ratio as 1:50. Last, we examined the drug exposure statuses within 30-day prior to the index dates. For both databases, the drug names were normalized to generic drug names. As we did not have the authority to share the original data, we created a mock data set that had similar structures as the administrative claim data. Additionally, we provided sample codes to prepare case-control data sets from the mock data set. Please see Supplementary Data and Codes.

Gold standard and DPA analysis

The OMOP drug-ADE gold standard was designed to establish a reference set for a pharmacovigilance study.[7] It included 399 drug-ADE pairs that were based on 181 drugs and the aforementioned four ADEs (acute myocardial infarction, acute renal failure, acute liver injury, and gastrointestinal bleeding). These 399 drug-ADE pairs were classed as 165 true positive test cases and 234 true negative test cases.

For each of the drug-ADE pair in the gold standard, the case-control data set (i.e., all case index dates and control index dates) was summarized into a two-by-two contingency table by statuses of the ADE (Yes/No) and the drug (Yes/No; i.e., a, b, c, and d in the 2-by-2 table). In this study, we selected two frequentist DPA approaches and one Bayesian DPA approach for evaluation. They were PRR, ROR, and IC (Table 1). Additionally, we computed PRR025, ROR025, and IC025, which were the lower bounds of the 95% confidence intervals for the aforementioned quantities.

TABLE 1 Pharmacoinformatics approaches: PRR, ROR and IC (a, b, c, and d are the four counts in a two-by-two contingency table)

DPA	Formula	Quantity of estimation and description
PRR[8]	$\frac{a}{(a+b)} / \frac{c}{(c+d)}$	$\frac{P(\text{ADE} \mid \text{Drug})}{P(\text{ADE} \mid \text{No Drug})}$
ROR[9]	$\frac{a}{b} / \frac{c}{d}$	$\left(\frac{P(\text{ADE}\mid\text{Drug})}{P(\text{No ADE}\mid\text{Drug})}\right) / \left(\frac{P(\text{ADE} \mid \text{No}\mathit{Drug})}{P(\text{No ADE} \mid \text{No Drug})}\right)$
IC[10]	$\log_2\left[\frac{a+1}{\frac{(a+c)\times(a+b)}{a+b+c+d}+1}\right]$	$\log_2\left(\frac{P(\text{ADE \& Drug})}{P(\text{ADE})\times P(\text{Drug})}\right)$ By adding 1 on both the numerator and the denominator, infrequent drug-ADE pairs will have penalized IC values.

Note: a = count (ADE = Yes and drug = Yes), b = count (ADE = No and drug = Yes), c = count (ADE = Yes and drug = No), and d = count (ADE = No and drug = No).

Abbreviations: IC, information component; PRR, proportional reporting ratio; ROR, reporting odds ratio.

RESULTS

There were 107,509,200 unique individuals in the MarketScan database. For the four ADEs, there were 864,928 distinct case index dates. We selected 20,000,000 random index dates as the control pool for random control selection approach. Subsequently, we generated case-control data sets by using the dynamic control selection approach, super control selection approach, and random control selection approach. We computed PRR, PRR025 (i.e., the lower bound of 95% confidence interval of PRR), ROR, ROR025, IC, and IC025 for all gold standard drug-ADE pairs. The rules to determine ADE signals for the aforementioned quantities were: PRR greater than 1, PRR025 greater than 1, ROR greater than 1, ROR025 greater than 1, IC greater than 0, and IC025 greater than 0. We computed precision, recall, and F-score (i.e., performance metrics) using the signal detection rules and the OMOP gold standard. The performance metrics for using all drug-ADE pairs are shown in Figure 2. First, F-scores of random control selection approach were either close to or slightly higher than the F-scores of dynamic/super control selection approaches. Specifically, random control selection approach had F-scores between 0.586 and 0.600; and dynamic/super control selection approaches had F-scores between 0.545 and 0.562. Second, the random control selection approach had higher recall values (0.854–0.961) compared to dynamic/super control selection approaches (0.720–0.804). Last, all three control selection approaches had similar precision values. Specifically, the random control selection approach had precision values between 0.436 and 0.446; and dynamic/super control selection approaches had precision values between 0.430 and 0.449. The ADE-specific performance metrics are shown in Figure S2. In a short summary, all three control selection approaches had similar F-scores for acute renal failure, acute liver injury, and gastrointestinal bleeding. However, the random control selection approach had higher F-scores for acute myocardial infarction. Additionally, in the INPC data analysis, the performance metrics of the random control selection were either close to or higher than dynamic/super control selection approaches (Figures S3 and S4).

Due to the random nature of the control selection process (i.e., the controls were randomly sampled), we replicated the control selection process 50 times to investigate the consistency of control selection. Using acute myocardial infarction as the ADE, we generated 50 case-control data sets for each of the control selection approaches. Subsequently, we computed precision, recall, and F-score; and their 95% empirical confidence intervals (i.e., the 2.5% and the 97.5% quantiles of the performance metrics). The results are shown in Figure 3. First, the random control selection approach had higher performance metrics than the dynamic/super control selection approaches. Dynamic control selection and super control selection had similar performance metrics. Second, all three control selection approaches yield consistent performance metrics with narrow empirical confidence intervals.

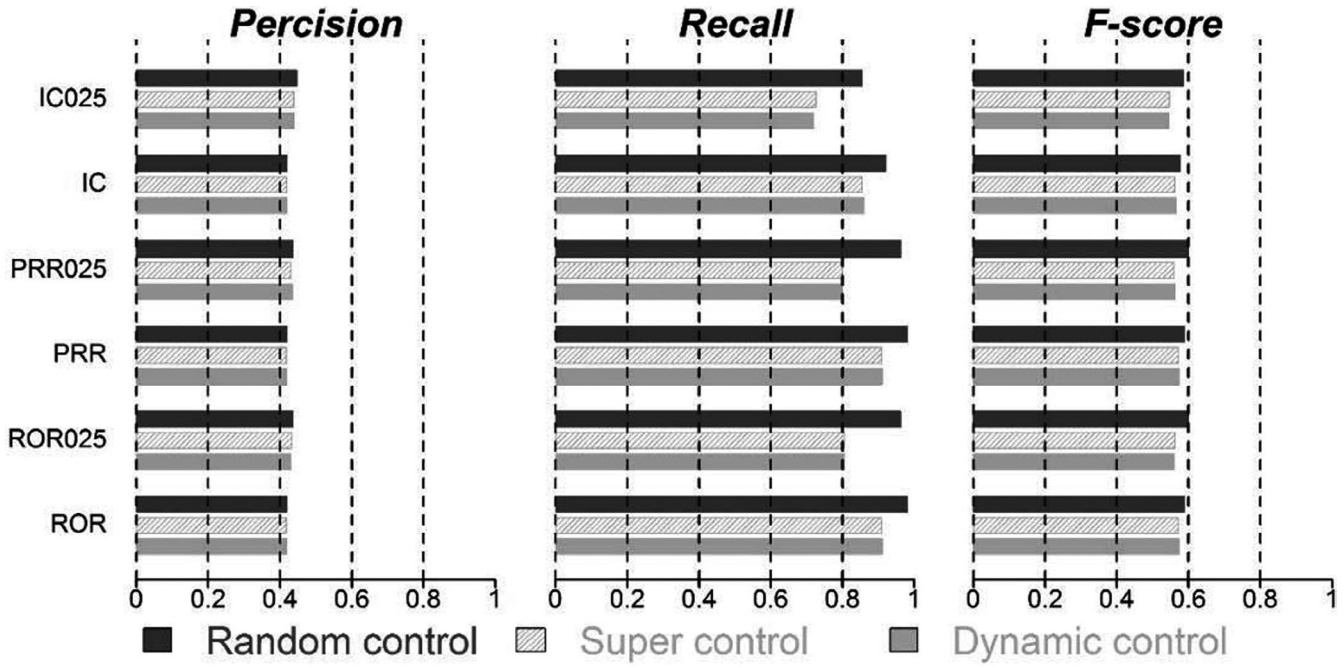

FIGURE 2 Precision, recall, and F-score in MarketScan data analysis. IC, information component; PRR, proportional reporting ratio; ROR, reporting odds ratio

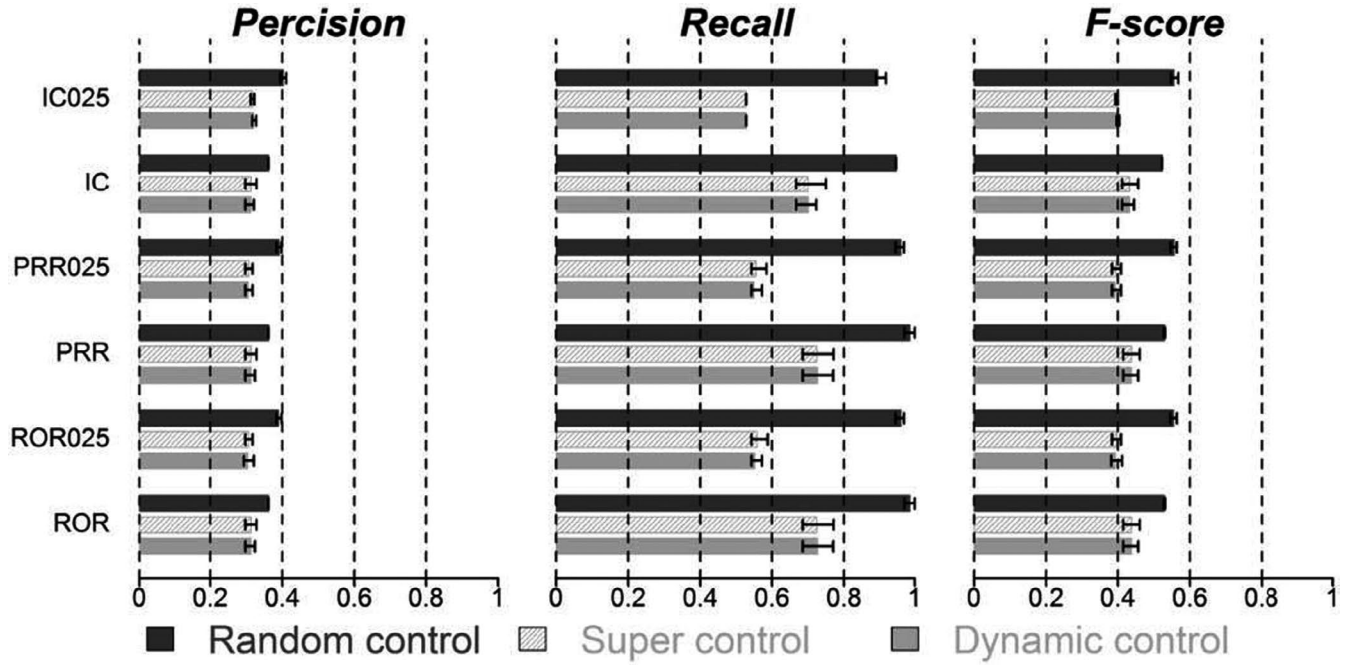

FIGURE 3 Performances for 50 independent replications using MarketScan data and acute myocardial infarction as ADE. ADE, adverse drug event; IC, information component; PRR, proportional reporting ratio; ROR, reporting odds ratio

We also evaluated the actual computation time for all control selection approaches using the MarketScan data on our local server. Because all control selection approaches had the same initialization steps (i.e., determine ADE case index dates and enrollment periods), we only compared the computational time to generate the control index dates. We fixed the case-control ratio as 1:50. With 107,509,200 individuals, we evaluated the computational time to generate: (i) 500 control index dates for 10 cases; (ii) 5000 control index dates for 100 cases; (iii) 50,000 control index dates for 1000 cases; and (iv) 500,000 control index dates for 10,000 cases. The computational times for 10, 100, 1000, and 10,000 cases were: (i) 0.01, 0.14, 1.41, and 13.94 h for the dynamic control selection approach; (ii) 0.01, 0.12, 1.14 and 10.86 h for the super control selection approach; and (iii) only 0.03, 0.35, 3.21, and 33.32 s for the random control selection approach (Figure 4). The random control selection approach was ~ 1000 times faster than the dynamic/super control selection approaches.

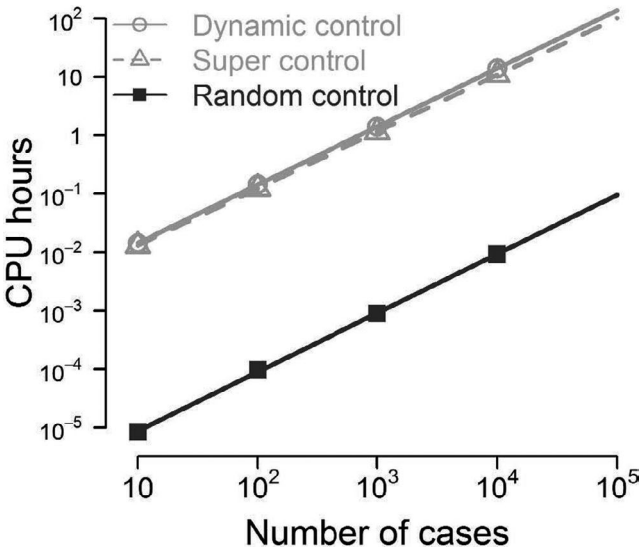

FIGURE 4 Actual and projected computation time for random control selection approach and dynamic/super control selection approach using MarketScan data

DISCUSSION

We propose a random control selection approach to conduct case-control design-based high-throughput ADE screening using large-scale longitudinal health data. Under the random control selection approach, randomly selected index dates are gathered as a random control pool, which can be used to prepare case-control data sets for multiple ADEs. Compared with existing dynamic/super control selection approaches, the random control selection approach relaxes the matching by the case index date restriction and the matching by ADE history restriction. We evaluated the performance metrics of all control selection approaches by using a large-scale administrative claims data set and drug-ADE gold standard. Using precision, recall, and F-score as

metrics, we identified that the proposed random control selection approach had similar performance metrics as the dynamic/super control selection approaches in three ADEs: acute liver injury, acute renal failure, and gastrointestinal bleeding (Figure 2), and better performance metrics than the dynamic/super control selection approaches in acute myocardial infarction (Figure 3). We replicated the control selection process 50 times, and observed the random control selection approach had consistent performance metrics (Figure 3). Thus, the reproducibility of the random control selection approach has been confirmed by 50 replications (Figure 3), and another EHR data analysis (Figures S3 and S4). These results suggested that the case-control design-based pharmacoinformatics study using the random control selection approach was comparable to using the dynamic/super control selection approach for screening ADEs.

Our primary motivation for proposing the random control selection approach is to reduce computational time for high-throughput ADE screening using large-scale longitudinal health data. At each case index date, the dynamic/super control selection approaches require all individuals to be evaluated for eligibility, as the control selection approaches depend on ADE/enrollment history. For instance, at a case index date, eligible dynamic controls are individuals who have not yet developed the ADE and have been enrolled over a period (Figure 1). Thus, the total computational time for dynamic/super control selection approaches is proportional to the total number of case index dates. Given the large amount of distinct case index dates in large-scale longitudinal health data, dynamic/super control selection approaches require a considerable amount of computational time. Moreover, for investigating multiple ADEs, the selected controls using dynamic/super control selection approaches for one ADE cannot be used as controls for another ADE, as the dynamic/super control selection approaches depend on the ADE history (i.e., the controls are ADE specific). Thus, computational time to select controls is further increased for screening multiple ADEs, as multiple case-control data sets are required. In contrast, the random control selection approach does not depend on the health history, nor depend on the ADE history. For random control selection, individuals are randomly selected to from a control pool, and control index dates are randomly selected as well. Thus, once the control pool has been formed, the total computational complexity remains fixed as the number of case index dates increases. Moreover, the random control pool can be used to generate case-control data sets for screening multiple ADEs without additional computational time expense. For large-scale longitudinal health data like MarketScan data (i.e., $N > 40$ million per year), dynamic/super control approaches required 100 h to prepare a case-control data set for an ADE that has a rate of 0.1% (Figure 4). Thus, it required 1000 h to prepare 10 case-control data sets for 10 ADEs with similar rates. Alternatively, the random control selection approach required only 5 min to generate the random control pool. Subsequently, control index dates were random sampled from the control pool. The total time to prepare one case-control data set was ~ 1 min. For 10 ADEs, the random control selection approach only required 15 min (i.e., 5 min for preparing the random control pool and 10 min for selecting control index dates for 10 ADEs).

We would like to point out that the primary aim for ADE screening is to prioritize true ADE signals (i.e., ADE signal ranking). Thus, bias is not a significant concern for screening ADE signals, as long as the true ADEs can be prioritized.[27] The proposed random control selection approach reduces computational time by relaxing the matching by case index date restriction and the matching by ADE history restriction. In traditional pharmacoepidemiological study, these two restrictions are used to reduce potential biases. Matching by case index date is able to reduce potential temporal bias.[28] Additionally, matching by ADE history restriction is able to reduce selection bias.[29] We conducted additional simulation studies to evaluate the impact of relaxation the aforementioned restrictions (details give in Supplementary Simulation Results). Based on 5000 simulations, we observed the estimated PRR, ROR, and IC values under the random control approach were closed to the values under the dynamic/super control selection approach (relative differences less than 1%). Thus, relaxing these restrictions may not induce significant biases. In fact, the biased control selection approaches were widely used for practical or scientific reasons.[25,30] For instance, the super control selection approach is a biased control selection approach.[25] Please note that even carefully conducted case-control design is subject to bias.[31,32] Currently, the active comparator design is considered as the gold standard for accessing drug outcome.[33] In the active comparator design, the ADE rate among patients exposed to the candidate drug is compared to the ADE rate among patients exposed to the comparator drug (i.e., a drug similar to the candidate drug). However, the active comparator design is highly computationally expensive, as it requires all drug exposures to be assessed. Thus, it is natural to first screen ADE signals by using a computationally efficient approach, and subsequently to validate the ADE signals by using a more rigorous approach. In this study, we observed: (i) the random control selection approach had similar or better performance metrics compared with the dynamic/super control selection approaches; and (ii) the random control selection approach required much less computational time. Based on these two reasons, the random control selection approach is able to accelerate ADE screening process and generate comparable ADE signals. We would like to point out that the proposed random control selection approach can reduce

bias by using stratified matching. This can be accomplished by generating separate random control pools for each of the strata. Although the actual computation time for data preparation in case-control design may depend on many factors (i.e., hardware and matching process), the proposed random control selection approach shall significantly reduce the computation time with stratified matching too.

In these studies, we selected four ADEs for performance evaluation. They were acute myocardial infarction, acute renal failure cases, liver injury, and gastrointestinal bleeding. We selected these ADEs as they were in the OMOP drug-ADE gold standard. These ADEs were also highly frequent and had significant clinical consequences (i.e., causing emergency department visit).[34] Thus, these ADEs have been continuously monitored by the US Food and Drug Administration (FDA).[35] Although we are not primarily focusing on the performance metrics of the DPA methods, we would like to discuss ADE screening with respect to different types of databases and ADEs. First, the performance metrics of a pharmacoinformatics study using different types of longitudinal health databases may be different. In administrative claims data analysis, the random control selection approach had F-score ~ 0.6 (precision ~ 0.45 and recall ~ 0.90) for all four ADEs. In other words, the random control selection approach was able to select ~ 90% true ADEs. In EHR data analysis, the random control selection approach had F-score ~ 0.5 (precision ~ 0.45 and recall ~ 0.40) for all four ADEs. Administrative claims data and EHR data had different informatics structure due to their nature structures.[36] Thus, an algorithm to identify an ADE may perform differently in these two types of data. Subsequently, the performance metrics of pharmacoinformatic studies may differ. Second, we identified that the performance metrics among four ADEs were different. In administrative claims data analysis, acute liver injury had F-score ~ 0.8; whereas acute myocardial infarction, acute renal failure, and gastrointestinal bleeding had F-scores between 0.4 and 0.6. The performance metrics of case-control design-based pharmacoinformatics study depend on the length of the window to examine drug exposure (i.e., drug exposure window).[18] In our study, we used 1 month drug exposure window for all four ADEs. However, the 1 month window may not be the best window for all ADEs. For a high-throughput pharmacoinformatics study of multiple ADEs, an ADE-specific drug exposure window can be used. The duration of the drug exposure window can be determined by clinical knowledge or sensitivity analysis. Additionally, the performance metrics of a pharmacoinformatic study also depend on confounding control.[37,38] Longitudinal health data contain a variety of variables, including both categorical and numerical confounders. Both categorical confounders and numerical confounders can be controlled by using multivariable analysis. Additionally, confounders can be controlled in the case-control design by using stratified case-control matching (i.e., matching by gender or dichotomized age). The performance of a pharmacoinformatic study also depends on the quality of phenotyping algorithms. Currently, ADE phenotyping is a fast growing field. We expect to see more accurate ADE phenotyping algorithms in the near future.

The scope of this work is to identify a computational efficient control selection algorithm for screening ADE signals from large-scale longitudinal health database. Currently, large-scale longitudinal health databases and ADE phenotyping algorithms become increasingly available. Using the random control selection approach, a case-control design-based pharmacoinformatic study can be upscaled to screen a tremendous amount of hypotheses without spending a tremendous amount of time, and identify single drugs and drug combinations with increased ADE risks. The random control selection approach is able to accelerate the subsequent validation studies as well. Ultimately, high-throughput adverse drug events screening using large-scale longitudinal health data will find better ways in promoting health. One limitation of this study is that the association between drug dosage and ADE was not investigated. Although dosage information is available in the MarketScan database and INPC database, its utilization requires sophisticated text mining algorithms. Another limitation is that whereas 1:4 to 1:10 case-control ratios were used in the current study, the optimal case-control ratio for ADE screening, remains unclear for large-scale health data mining.

With large-scale longitudinal health data, a case-control design-based pharmacoinformatic study using randomly selected index dates as controls (i.e., random control selection approach) has similar or higher performance metrics compared with existing control selection approaches. Compared with existing control selection approaches, the random control selection approach is able to significantly reduce the time to prepare the case-control data sets.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

AUTHORS CONTRIBUTIONS

All authors wrote the manuscript. L.L., P.Z., and M.D. designed the research. C.C. and P.Z. performed the research. C.C., P.Z., Y.S. and Y.C. analyzed the data.

REFERENCES

1. de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review. *Qual Saf Health Care*. 2008;17(3):216-223.

2. Hall MJ, DeFrances CJ, Williams SN, Golosinskiy A, Schwartzman A. National Hospital Discharge Survey: 2007 summary. *Natl Health Stat Report*. 2010;29:1-20, 4.
3. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*. 1998;279(15):1200-1205.
4. Downing NS, Shah ND, Aminawung JA, et al. Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration between 2001 and 2010. *J Am Med Assoc*. 2017;317(18):1854-1863.
5. Goldmann D, Montanari F, Richter L, Zdrazil B, Ecker GF. Exploiting open data: a new era in pharmacoinformatics. *Future Med Chem*. 2014;6(5):503-514.
6. Hallas J, Wang SV, Gagne JJ, Schneeweiss S, Pratt N, Pottegård A. Hypothesis-free screening of large administrative databases for unsuspected drug-outcome associations. *Eur J Epidemiol*. 2018;33(6):545-555.
7. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clin Pharmacol Ther*. 2013;93(6):539-546.
8. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidem Drug Saf*. 2001;10(6):483-486.
9. van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidem Drug Saf*. 2002;11(1):3-11.
10. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol*. 1998;54(4):315-321.
11. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat*. 1999;53(3):177-190.
12. Zhang P, Li M, Chiang C, et al. Three-component mixture model-based adverse drug event signal detection for the adverse event reporting system. *CPT Pharmacometrics Syst Pharmacol*. 2018;7(8):499-506.
13. Caster O, Norén GN, Madigan D, Bate A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. *Statistical Analy Data Mining*. 2010;3(4):197-208.
14. Patadia VK, Schuemie MJ, Coloma P, et al. Evaluating performance of electronic healthcare records and spontaneous reporting data in drug safety signal detection. *Int J Clin Pharm*. 2015;37(1):94-104.
15. Zhu A, Zeng D, Shen L, Ning X, Li L, Zhang P. A super-combo-drug test to detect adverse drug events and drug interactions from electronic health records in the era of polypharmacy. *Stat Med*. 2020;39(10):1458-1472.
16. Wang X, Zhang P, Chiang CW, et al. Mixture drug-count response model for the high-dimensional drug combinatory effect on myopathy. *Stat Med*. 2018;37(4):673-686.
17. Li Y, Ryan PB, Wei Y, Friedman C. A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. *Drug Saf*. 2015;38(10):895-908.
18. Hennessy S, Leonard CE, Gagne JJ, et al. Pharmacoepidemiologic methods for studying the health effects of drug-drug interactions. *Clin Pharmacol Ther*. 2016;99(1):92-100.
19. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med*. 2016;71:57-61.
20. Zhang X, Pérez-Stable EJ, Bourne PE, et al. Big Data Science: opportunities and challenges to address minority health and health disparities in the 21st century. *Ethn Dis*. 2017;27(2):95-106.
21. Caster O, Sandberg L, Bergvall T, Watson S, Norén GN. vigiRank for statistical signal detection in pharmacovigilance: first results from prospective real-world use. *Pharmacoepidemiol Drug Saf*. 2017;26(8):1006-1010.
22. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data*. 2016;3:160026.
23. Lee S, Han J, Park RW, et al. Development of a controlled vocabulary-based adverse drug reaction signal dictionary for multicenter electronic health record-based pharmacovigilance. *Drug Saf*. 2019;42(5):657-670.
24. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *J Biomed Inform*. 2012;45(4):689-696.
25. Lubin JH, Gail MH. Biased selection of controls for case-control analyses of cohort studies. *Biometrics*. 1984;40(1):63-75.
26. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986;73(1):1-11.
27. Vilar S, Ryan PB, Madigan D, et al. Similarity-based modeling applied to signal detection in pharmacovigilance. *CPT Pharmacometrics Syst Pharmacol*. 2014;3:e137.
28. Yuan W, Beaulieu-Jones BK, Yu KH, et al. Temporal bias in case-control design: preventing reliable predictions of the future. *Nat Commun*. 2021;12(1):1107.
29. Cheung YB, Ma X, Lam KF, Li J, Milligan P. Bias control in the analysis of case-control studies with incidence density sampling. *Int J Epidemiol*. 2019;48(6):1981-1991.
30. Robins JM, Gail MH, Lubin JH. More on "Biased selection of controls for case-control analyses of cohort studies". *Biometrics*. 1986;42(2):293-299.
31. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, editors. *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. AHRQ Methods for Effective Health Care; 2013.
32. Schuemie MJ, Ryan PB, Man KKC, Wong ICK, Suchard MA, Hripcsak G. A plea to stop using the case-control design in retrospective database studies. *Stat Med*. 2019;38(22):4199-4208.
33. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol*. 2015;11(7):437-441.
34. Shehab N, Lovegrove MC, Geller AI, Rose KO, Weidle NJ, Budnitz DS. US Emergency Department visits for outpatient adverse drug events, 2013–2014. *JAMA*. 2016;316(20):2115-2125.
35. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative–A comprehensive approach to medical product surveillance. *Clin Pharmacol Ther*. 2016;99(3):265-268.
36. Girman CJ, Ritchey ME, Zhou W, Dreyer NA. Considerations in characterizing real-world data relevance and quality for regulatory purposes: a commentary. *Pharmacoepidemiol Drug Saf*. 2019;28(4):439-442.

37. Alzoubi H, Alzubi R, Ramzan N, West D, Al-Hadhrami T, Alazab M. A review of automatic phenotyping approaches using electronic health records. *Electronics-Switz*. 2019;8(11):1235.
38. Streeter AJ, Lin NX, Crathorne L, et al. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *J Clin Epidemiol*. 2017;87:23-34.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Chiang C-W, Zhang P, Donneyong M, Chen Y, Su Y, Li L. Random control selection for conducting high-throughput adverse drug events screening using large-scale longitudinal health data. *CPT Pharmacometrics Syst Pharmacol*. 2021;10:1032–1042. https://doi.org/10.1002/psp4.12673