A Drug Similarity-based Early Adverse Drug Event Detection Approach

Yi Shi,[1#] Yuedi Yang,[1#] Ruoqi Liu,[2#] Anna Sun,[1] Xueqiao Peng,[2] Lang Li,[3] Ping Zhang,[2,3*] and Pengyue Zhang[1*]

1. Department of Biostatistics and Health Data Science, Indiana University, Indianapolis, Indiana, USA
2. Department of Computer Science and Engineering, the Ohio State University, Columbus, Ohio, USA
3. Department of Biomedical Informatics, the Ohio State University, Columbus, Ohio, USA

# These authors contributed equally.

* To whom correspondence should be addressed:

Ping Zhang: ping.zhang@osumc.edu

Pengyue Zhang: zhangpe@iu.edu

Supplemental Material S1: Supplemental methods

S1.1. Reference methods used in FAERS data analysis

Let subscript i indicate the ith drug, subscript j indicate the jth adverse drug event (ADE), and subscript (k) indicate the kth time point. Let $X_{ij(k)}$ denote the observed frequency, $E_{ij(k)} = \left(\sum_i X_{ij(k)} \sum_j X_{ij(k)}\right) / \sum_i \sum_j X_{ij(k)}$ denote the expected frequency, and $\lambda_{ij(k)}$ denote unobserved relative risk. Additionally, let $\text{gamma}(\lambda; \alpha, \beta)$ denote the gamma distribution, $\Gamma(\cdot)$ denote the gamma function, and $\text{NB}(X, E; \alpha, \beta) = \frac{\Gamma(X+\alpha)}{X!\Gamma(\alpha)} \times \frac{E^X \beta^\alpha}{(E+\beta)^{X+\alpha}}$ denote the qausi-negative binomial distribution. The posterior distribution of $\lambda_{ij(k)}$s under multi-item gamma Poisson shrinker (MGPS) was defined as: $f\left(\lambda_{ij(k)} | X_{ij(k)}, E_{ij(k)}\right) = \hat{\pi}_{ij(k)} \times \text{gamma}\left(\lambda_{ij(k)}; \hat{\alpha}_1 + X_{ij(k)}, \hat{\beta}_1 + E_{ij(k)}\right) + \left(1 - \hat{\pi}_{ij(k)}\right) \times \text{gamma}\left(\lambda_{ij(k)}; \hat{\alpha}_2 + X_{ij(k)}, \hat{\beta}_2 + E_{ij(k)}\right)$. Under MGPS, posterior probability of null hypothesis (PP(MGPS)) and positive false discovery rate (pFDR(MGPS)) could be used to control false positive in high-throughput ADE mining [1, 2]. PP(MGPS) and pFDR were defined as:

$$\text{PP(MGPS)}_{ij(k)} = \text{Prob}\left(\lambda_{ij(k)} \leq 1 | X_{ij(k)}, E_{ij(k)}\right). \quad \text{(S1)}$$

$$\text{pFDR(MGPS)}_{ij(k)} = \frac{\hat{\pi} \times \sum_{X \geq X_{ij(k)}} \text{NB}\left(X, E_{ij(k)}; \hat{\alpha}_1, \hat{\beta}_1\right)}{\hat{\pi} \times \sum_{X \geq X_{ij(k)}} \text{NB}\left(X, E_{ij(k)}; \hat{\alpha}_1, \hat{\beta}_1\right) + (1-\hat{\pi}) \times \sum_{X \geq X_{ij(k)}} \text{NB}\left(X, E_{ij(k)}; \hat{\alpha}_2, \hat{\beta}_2\right)}. \quad \text{(S2)}$$

Additionally, empirical Bayesian geometric mean under MGPS (EBGM(MGPS)) could be used to prioritize signals without false positive control [3]. EBGM(MGPS) was defined as:

$$\text{EBGM(MGPS)}_{ij(k)} = 2^{\left[E\left(\log \lambda_{ij(k)} | X_{ij(k)}, E_{ij(k)}\right) / \log 2\right]}. \quad \text{(S3)}$$

Moreover, information component under Bayesian confidence propagation neural network (IC(BCPNN)) could be used to prioritize signals without false positive control [4]. Let $N_{i+(k)}$, $N_{j+(k)}$, and $N_{\text{Total}(k)}$ denote the marginal report frequency of drug i, marginal report frequency of ADE j, and total report frequency, respectively. IC(BCPNN) was defined as:

$$\text{IC(BCPNN)}_{ij(k)} = \log_2 \frac{\left(N_{ij(k)}+1\right)\left(N_{\text{Total}(k)}+2\right)^2}{\left(N_{\text{Total}(k)}+2\right)^2 + N_{\text{Total(k)}}\left(N_{i+(k)}+1\right)\left(N_{j+(k)}+1\right)} \quad \text{(S4)}$$

S1.2. Clinformatics® Data Mart Database analysis

We collected ~1.5 million potential ADE-induced emergency department (ED) visits (i.e., cases) from Optum's de-identified Clinformatics® Data Mart Database (years: 2007-2021). In short words, we used subject matter expert-derived diagnosis codes to identify potential ADEs including acute myocardial infraction, fall, gastrointestinal bleeding, and hypoglycemia. We collected the drug exposure immediately prior to the case (e.g., the case period) and distant to the case (e.g., the control period) under a case-crossover design (Figure S1). Additional details on data preparation were described in Shi et al. [5]. We included 358 relative newer drugs that had: A) frequencies ≤50 in January 2008, B) frequencies ≥500 in December 2021, and C) chemical similarity scores.

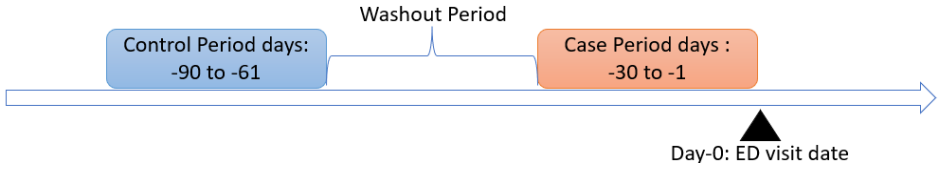

Figure S1. Illustration of the case-crossover design, case period and control period.

For a drug, let $N_{ij(k)}$ denote the total number of cases with discordant drug exposure status between case and control periods (Figure S1), and $X_{ij(k)}$ denote the total number of cases with drug exposure only in case period but not in control period (Figure S1). We computed $N_{ij(k)}$s and $X_{ij(k)}$s at each quarter. We assumed $X_{ij(k)}$ ~ Binomial ($N_{ij(k)}$,$p_{ij(k)}$) [5], where $p_{ij(k)}$ =0.5 under the null hypothesis (i.e. no association between drug and case).

For the proposed drug similarity-based method, we assumed $X_{ij(k)}$ ~ Binomial ($N_{ij(k)}$,$p_{ij(k)}$) and $p_{ij(k)}$ ~ Beta($p_{ij(k)}$;$\alpha_{ij(k)}$,$\beta_{ij(k)}$). We defined the posterior probability of null hypothesis as $\mathrm{P}\left(p_{ij(k)} \le 0.5 | X_{ij(k)}, N_{ij(k)}\right)$.

For reference methods, we used the precision mixture risk model (PMRM) [5]. Let $\theta_{ij(k)}$ denote the odds ratio of the drug-ADE pair, beta' denote the beta prime distribution, and BB denote the beta-binomial distribution. The posterior distribution of $\theta_{ij(k)}$ under PMRM was defined as: $f\left(\theta_{ij(k)} | X_{ij(k)}, N_{ij(k)}\right) = \hat{\pi}_{ij(k)} \times \mathrm{beta}'\left(\theta_{ij(k)}; \hat{\alpha}_1 + X_{ij(k)}, \hat{\beta}_1 + N_{ij(k)} - X_{ij(k)}\right) + \left(1 - \hat{\pi}_{ij(k)}\right) \times \mathrm{beta}'\left(\theta_{ij(k)}; \hat{\alpha}_2 + X_{ij(k)}, \hat{\beta}_2 + N_{ij(k)} - X_{ij(k)}\right)$. Under PMRM, posterior probability of null hypothesis (PP(PMRM)) and positive false discovery rate (pFDR(PMRM)) could be used to control false positive in high-throughput ADE mining [2, 5]. PP(PMRM) and pFDR(PMRM) were defined as:

$$\mathrm{PP(PMRM)}_{ij(k)} = \mathrm{P}\left(\theta_{ij(k)} \le 1 | X_{ij(k)}, N_{ij(k)}\right). \tag{S5}$$

$$\mathrm{pFDR(PMRM)}_{ij(k)} = \frac{\hat{\pi} \times \sum_{X \ge X_{ij(k)}} \mathrm{BB}\left(X, N_{ij(k)}; \hat{\alpha}_1, \hat{\beta}_1\right)}{\hat{\pi} \times \sum_{X \ge X_{ij(k)}} \mathrm{BB}\left(X, N_{ij(k)}; \hat{\alpha}_1, \hat{\beta}_1\right) + (1 - \hat{\pi}) \times \sum_{X \ge X_{ij(k)}} \mathrm{BB}\left(X, N_{ij(k)}; \hat{\alpha}_2, \hat{\beta}_2\right)}. \tag{S6}$$

S1.3. Simulation study

We used the marginal reporting frequencies and the estimated prior distributions of new drugs in FAERS data analysis to simulate data. Let subscript i indicate the ith drug, subscript j indicate the jth ADE, and subscript (k) indicate the kth time point. Let $\theta_{ij(k)}$ denote the simulated ADE reporting rate (i.e., true rate), $X_{ij(k)}$ denote the simulated frequency of drug-ADE pair, and $N_{ij(k)}$ denote the frequency of the drug in FAERS data. For one drug-ADE pair, the data were simulated as following: (A) to simulate $\theta_{ij(k)}$ from beta($\theta_{ij(k)}$; $\hat{\alpha}_{ij(k)}$, $\hat{\beta}_{ij(k)}$), (B) to simulate $X_{ij(k)}$ from binomial($X_{ij(k)}$; $N_{ij(k)}$, $\theta_{ij(k)}$); and (C) to simulate an indicator δ (i.e., δ from binary(probability =0.8)) to characterize mis-specification of prior distribution. Subsequently, we determined condition positive and condition negative by comparing $\theta_{ij(k)}$s (i.e., Prob[ADE | drug]) to their corresponding probabilities without drug exposure (e.g., $\theta^*_{ij(k)}$ = Prob[ADE | no drug]) in FAERS data. The tested posterior probability of null hypothesis was defined as:

$$\begin{aligned}\mathrm{P}\left(\theta_{ij(k)} \le \theta^*_{ij(k)} | X_{ij(k)}, N_{ij(k)}\right) = \delta \times \mathrm{IB}\left(\theta^*_{ij(k)}; \hat{\alpha}_{ij(k)} + X_{ij(k)}, \hat{\beta}_{ij(k)} + N_{ij(k)} - X_{ij(k)}\right) + \\ (1 - \delta) \times \mathrm{IB}\left(\theta^*_{ij(k)}; \check{\alpha}_{ij(k)} + X_{ij(k)}, \check{\beta}_{ij(k)} + N_{ij(k)} - X_{ij(k)}\right).\end{aligned} \tag{S7}$$

In equation S7, the two terms associated with δ and $1 - \delta$ represented posterior probability of null hypothesis under correctly specified and mis-specified prior distributions, where $\check{\alpha}_{ij(k)}$ and $\check{\beta}_{ij(k)}$ were randomly selected to account for model misspecification. We also used the marginal frequencies in FAERS data and the simulated $X_{ij(k)}$s to compute PP(MGPS) and pFDR(MGPS) (equations S1 and S2).

Supplemental Material S2: Supplemental results

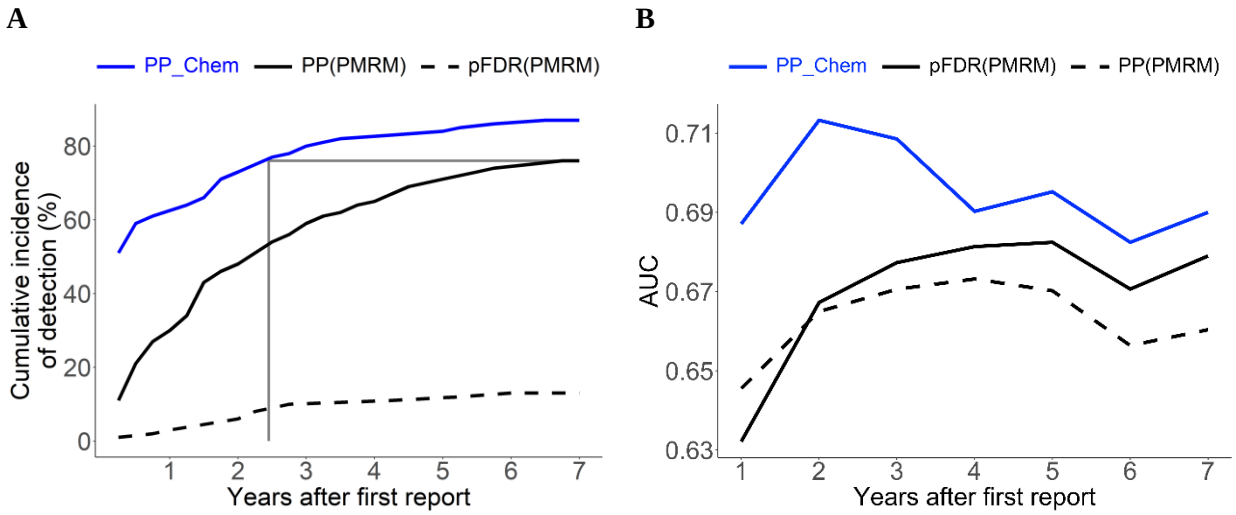

Figure S2. Results in Clinformatics® data analysis: **A.** Cumulative incidence of detection, **B.** AUC.

Supplemental Material S3: Model R codes

An example to estimate of drug similarity-based posterior probability.

```
###### setup
Count_similar_drug=25          ### number of similar drugs used in analysis
Alpha_similar_drug=2           ### alpha in the beta distribution
Beta_similar_drug=48           ### beta in the beta distribution
P_sim=rbeta(Count_similar_drug,Alpha_similar_drug,Beta_similar_drug)          ### ADE probabilities of similar drugs
N_sim=rpois(Count_similar_drug,500)                    ### total Ns for similar drugs
X_sim=rbinom(Count_similar_drug,N_sim,P_sim)        ### Ns drug-ADE pairs for similar drugs
N_new=50               ### N for new drug
X_new=2                ### N drug-ADE pairs for new drug
P_threshold= (X_new/N_new)/3          ### a threshold for hypothesis test (i.e. null hypothesis)
Box_low=log(c(1,1)/1000)              ### a lower bound of estimated parameter
Box_high=log(c(1,1)*100)              ### an upper bound of estimated parameter

###### log-likelihood function based on beta-binomial distribution for X_sim.
lc_Beta=function(fun_temp)    ### The input are parameters in the beta distribution (i.e., prior distribution)
{ fun_alpha=exp(fun_temp[1])
  fun_beta=exp(fun_temp[2])
  fun_ll=sum(lgamma(X_sim+fun_alpha)-lgamma(fun_alpha)+
          lgamma(N_sim-X_sim+fun_beta)-lgamma(fun_beta)+
          lgamma(fun_alpha+fun_beta)-lgamma(fun_alpha+fun_beta+N_sim))
  -fun_ll }       ### The output is the log-likelihood based on beta-binomial distribution

###### Estimation of drug similarity-based posterior probability
fit_Beta=nlminb(c(0,0),lc_Beta,lower=Box_low,upper=Box_high)          ### find parameter
beta_alpha=exp(fit_Beta$par[1])       ### alpha in the beta distribution
beta_beta=exp(fit_Beta$par[2])        ### beta in the beta distribution
Post_prob=pbeta(P_threshold,beta_alpha+X_new,beta_beta+N_new-X_new)    ### posterior probability
Post_prob     ### posterior probability
```

Sample code for matching drugs with and without a new drug based on co-medications.

```
###### setup
N_matcing_ratio=5     ### set a non-new drug observation to new drug observation ratio
Pseudo_case=vector("list",20)          ### create pseudo observations for new drug
names(Pseudo_case)=paste0("Case_",c(1:length(Pseudo_case)))        ### assign names to observations for new drug
for(i in 1:length(Pseudo_case))        ### assign drugs to observations for new drug
{ Pseudo_case[[i]]=c("New_drug",paste0("Drug_",sample(letters,2))) }
Pseudo_case[1:5]                       ### sample observations for new drug
Pseudo_control=vector("list",10000)    ### create pseudo observations for non-new drug
### assign names to observations for non-new drug
names(Pseudo_control)=paste0("Control_",c(1:length(Pseudo_control)))
for(i in 1:length(Pseudo_control))                ### assign drugs to observations for non-new drug
{ Pseudo_control[[i]]=c(paste0("Drug_",sample(letters,3))) }
Pseudo_control_original=Pseudo_control        ### keep a copy of observations for non new drug
Pseudo_control[1:5]                           ### sample observations for non-new drug
fun_match_score=function(fun_temp)            ### define a function to compute matching score
{ length(intersect(fun_temp,drug_to_match)) } ### matching score based on common co-medications

###### start to match
Matched_control=c()   ### matched observations will be saved in here
for(loop_i in 1:length(Pseudo_case))     ### loop over new drug observations
{ drug_to_match=Pseudo_case[[loop_i]]          ### concomitant drugs for the current case
  ### compute matching score between observations
  all_match_score=as.numeric(unlist(lapply(Pseudo_control,fun_match_score)))
  Control_ID_score=data.frame(names(Pseudo_control),all_match_score)          ### a dataset with matching score
  Control_ID_score=Control_ID_score[order(all_match_score,decreasing=TRUE),] ### order based on matching score
```

```
  Selected_control=Control_ID_score[1:N_matcing_ratio,1]                ### matching
  Matched_control_loop=data.frame(Selected_control)             ### matched observations
  Matched_control_loop$Case=names(Pseudo_case[loop_i])    ### asign names to link matched observations
  Matched_control=rbind(Matched_control,Matched_control_loop)       ### update matched observations
  ### update unmatched observations
  Pseudo_control=Pseudo_control[(names(Pseudo_control)%in%Matched_control[,1])==FALSE]
  cat("Iter = ",loop_i,"\n") }

###### Examples: a new drug observation and its matched non-new drug observations
Example_case=Pseudo_case["Case_3"]
Example_control=Pseudo_control_original[names(Pseudo_control_original)%in%Matched_control[Matched_control$C
ase=="Case_3",1]]
Example_case
Example_control
```

Supplemental Material S4: References

1. Ahmed I, Haramburu F, Fourrier-Reglat A, Thiessard F, Kreft-Jais C, Miremont-Salame G, et al. Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. Stat Med. 2009 Jun 15;28(13):1774-92.
2. Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat. 2003 Dec;31(6):2013-35.
3. DuMouchel W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. The American Statistician. 1999;53(3).
4. Bate A. Bayesian confidence propagation neural network. Drug Saf. 2007;30(7):623-5.
5. Shi Y, Eadon MT, Chen Y, Sun A, Yang Y, Chiang C, et al. A Precision Mixture Risk Model to Identify Adverse Drug Events in Subpopulations Using a Case-Crossover Design. Stat Med. 2024 Nov 30;43(27):5088-99.