Received: 9 April 2022 | Revised: 25 January 2023 | Accepted: 29 January 2023

DOI: 10.1002/sim.9684

RESEARCH ARTICLE

Statistics in Medicine WILEY

Mining adverse events in large frequency tables with ontology, with an application to the vaccine adverse event reporting system

Bangyao Zhao | **Lili Zhao**

Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Correspondence
Lili Zhao, Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA.
Email: lili.zhao@beaumont.org

Funding information
National Institute of Allergy and Infectious Diseases, Grant/Award Number: R01AI158543

Many statistical methods have been applied to VAERS (vaccine adverse event reporting system) database to study the safety of COVID-19 vaccines. However, none of these methods considered the adverse event (AE) ontology. The AE ontology contains important information about biological similarities between AEs. In this paper, we develop a model to estimate vaccine-AE associations while incorporating the AE ontology. We model a group of AEs using the zero-inflated negative binomial model and then estimate the vaccine-AE association using the empirical Bayes approach. This model handles the AE count data with excess zeros and allows borrowing information from related AEs. The proposed approach was evaluated by simulation studies and was further illustrated by an application to the Vaccine Adverse Event Reporting System (VAERS) dataset. The proposed method is implemented in an R package available at https://github.com/umich-biostatistics/zGPS.AO.

KEYWORDS

adverse event ontology, empirical Bayes, vaccine adverse event, VAERS, zero-inflated negative binomial distribution

1 | INTRODUCTION

The Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA) conduct post-licensure vaccine safety monitoring using the Vaccine Adverse Event Reporting System (VAERS).[1,2] VAERS accepts spontaneous reports of suspected vaccine adverse events after administration of any vaccine licensed in the United States from 1990 to the present. As a national public health surveillance resource, VAERS is a key component in ensuring the safety of vaccines. Numerous methods have been used to conduct safety studies with the VAERS database.[3-17] In these methods, a contingency table is generally created to display counts for all vaccine and adverse event pairs during a specified time period (see Table 1). In this table, there are I vaccines and J AEs. The cell count y_{ij} is the total number of reports mentioned both vaccine i and AE j in this time period, the column margin $y_{i.} = \sum_{j=1}^{J} y_{ij}$ is the total number of reports mentioned vaccine i, the row margin $y_{.j} = \sum_{i=1}^{I} y_{ij}$ is the total number of reports mentioned AE j, and $y_{..} = \sum_{ij} y_{ij}$ is the total number of reports in this time period.

The level of disproportional reporting of a vaccine-AE pair is commonly expressed as the ratio of the observed reporting frequency to the expected (or control) reporting frequency. The expected frequency for the i-jth vaccine-AE pair is defined as $M_{ij} = \frac{y_{i.} y_{.j}}{y_{..}}$, which is the frequency we would observe if the vaccine and the AE are independent.[3] The relative reporting

TABLE 1 An example of the contingency table.

Vaccine\AE	AE_1	AE_2	…	AE_J
VAX_1	y_{11}	y_{12}	…	y_{1J}
VAX_2	y_{21}	y_{22}	…	y_{2J}
…	…	…	…	…
VAX_I	y_{I1}	y_{I2}	…	y_{IJ}

rate (RR) is defined as $RR_{ij} = \frac{y_{ij}}{M_{ij}}$. If RR= 3.2 for a vaccine-AE pair, then this pair occurred in the data 3.2 times more frequently than expected under the assumption of no association between the vaccine and the AE. Each cell count is commonly assumed to have an independent Poisson distribution;[3] that is, $y_{ij} \sim Poisson(M_{ij}\lambda_{ij})$, where λ_{ij} is the parameter of interest representing the RR, with a larger value indicating the vaccine i - AE j pair is disproportionately reported in the dataset. A large RR might indicate a potential vaccine safety problem (called a safety "signal").

All the existing methods assume that AEs are independent; however, AEs are naturally related. For example, events of retching, dysphagia and reflux are all related to an abnormal digestive system. Explicitly bringing AE relationships into the estimation step allows information borrowing between similar AEs, thereby improving the accuracy of detecting true AE signals amid the noise while reducing false positives. In this paper, we use MedDRA (Medical Dictionary for Regulatory Activities) to group AEs. MedDRA is the largest resource for AE ontology, which defines disease relationships using a multi-level hierarchy.[18] VAERS uses the "Preferred Terms" (PT) level as a distinct descriptor for a symptom, sign, and disease. Related PTs are grouped into higher-level HLGT terms. We define AE groups using the "High Level Group Terms" (HLGT) level of MedDRA. MedDRA ontology data can be accessed through BioPortal https://bioportal.bioontology.org/ontologies/MEDDRA. We applied the web scraping technique to the HTML-formatted data in BioPortal and obtained the ontology for the adverse events.

One recent study performed AE enrichment analysis using the AE ontology in MedDRA.[19] Their method focuses on identifying enriched AE groups where AEs are more likely to be disproportionately reported than AEs in other groups. However, this enrichment analysis is done after RRs of individual AEs have been estimated. Therefore, their method cannot directly improve the accuracy of signal detection. Another challenge in analyzing VAERS data comes from excessive zero counts. MedDRA is a dictionary containing thousands of PT terms for various symptoms and diseases. Therefore, a large number of AEs in VAERS were never mentioned for many vaccines; For example, in VAERS data from 2002 to 2018, approximately 40% AEs were never mentioned with the "FLU4" vaccine, resulting in 40% AEs with a zero count.[19] In this paper, we propose a model allowing information sharing between AEs within the same group while accommodating zero counts.

To incorporate the grouping structure of AEs in the estimation of RRs while accommodating excessive zero counts in VAERS data, we consider a zero-inflated negative binomial distribution (ZINB). We assume that RR parameters, λ_{ij}'s, in the same AE group (defined by HLGT or SOC) are generated from a common gamma distribution. This common distribution allows information sharing between λ_{ij}'s in the same AE group. By parameterizing the negative binomial distribution through a gamma-Poisson mixture distribution, we can model the AE data on both the PT (child) and a higher level (parent), allowing us to simultaneously mine safety signals at both levels. As a simple example, suppose the respiratory system AE includes five child AEs: pneumonia, sinusitis, asthma, bronchitis, and rhinorrhea, all of which carry weak to moderate signals. We hypothesize that our model will not only flag each AE through information sharing between weak and moderate AEs, but will also flag the respiratory system as a whole because the five child AEs collectively suggest that the respiratory system might be the root of the adverse cause.

We adopt an empirical Bayes approach for parameter estimation. Our method is reproducible using the R package zGPS.AO available at https://github.com/umich-biostatistics/zGPS.AO.

2 | METHOD

2.1 | Model

In this section, we propose a model called the zero-inflated Gamma-Poisson shrinker with AE ontology (abbreviated as zGPS.AO). In zGPS.AO, we model each AE group separately. Suppose an AE group includes K AE terms. We assume the following hierarchical model for modelling the count y_{ij} ($i = 1, \dots, I$ and $j = 1, \dots, K$).

$$y_{ij} \sim Poi(M_{ij}\lambda_{ij}), \tag{1}$$

$$\lambda_{ij} \sim \begin{cases} 0 & \text{with probability } p_i \\ \Gamma(r, \mu_i/r) & \text{with probability } 1 - p_i \end{cases}. \tag{2}$$

The mean of the Poisson distribution is expressed as the product of the expected frequency M_{ij} times λ_{ij}, where λ_{ij} is the RR of AE j for vaccine i (the key parameter of interest to identify important AEs). Here, λ_{ij} is drawn from a mixture of 0 with probability p_i and a Gamma distribution with probability $1 - p_i$. $\Gamma(a, b)$ denotes a Gamma distribution with a shape parameter, a, and a scale parameter, b. In equation (2), μ_i is the mean of the Gamma distribution. All vaccines share a common dispersion parameter r, but they have different p_i's and μ_i's.

The group-level RR of vaccine i (denoted as s_i) is defined as the mean of λ_{ij}.

$$s_i = E(\lambda_{ij}) = (1 - p_i)\mu_i. \tag{3}$$

Equations (1) and (2) imply y_{ij}'s are marginally ZINB-distributed.

$$\text{ZINB}(y_{ij}|r, p_i, M_{ij}\mu_i) = \begin{cases} p_i + (1 - p_i)\text{NB}(0|r, M_{ij}\mu_i) & \text{if } y_{ij} = 0, \\ (1 - p_i)\text{NB}(y_{ij}|r, M_{ij}\mu_i) & \text{if } y_{ij} > 0, \end{cases} \tag{4}$$

where NB denotes a negative binomial (NB) distribution. When p_i is small and μ_i is large relative to a control value, s_i is large. That is, if the AE group has a small percentage of structural zeros and a large mean for the Poisson part, there is a high risk of the AE group associated with vaccine i (see the Appendix for mathematical details).

2.2 | AE-level estimation

We propose an Empirical Bayes approach to estimate the AE-level RRs (λ_{ij}'s). We first use the maximum likelihood estimator (MLE) to estimate the group-level parameters ($p_1, \dots, p_I, \mu_1, \dots, \mu_I, r$); see the next section on how to obtain these estimates. Then we plug the estimates into the model to obtain the posterior distribution of λ_{ij}, which is given by

$$\pi(\lambda_{ij}|y_{ij}) = \begin{cases} \hat{\pi}_{ij}\delta(\lambda_{ij}) + (1 - \hat{\pi}_{ij})\Gamma\left(\lambda_{ij}|r, \frac{\mu_i}{r+M_{ij}\mu_i}\right) & \text{if } y_{ij} = 0, \\ \Gamma\left(\lambda_{ij}|r + y_{ij}, \frac{\mu_i}{r+M_{ij}\mu_i}\right) & \text{if } y_{ij} > 0, \end{cases}$$
$$\text{where } \hat{\pi}_{ij} = \frac{p_i}{p_i + (1 - p_i)\left(\frac{r}{r+M_{ij}\mu_i}\right)^r}. \tag{5}$$

Here, $\delta(\cdot)$ is the Dirac delta function,[20] denoting the p.d.f. of the degenerated random variable at 0. If $y_{ij} = 0$, the posterior distribution of λ_{ij} is a mixture of 0 and a gamma distribution. The weight $\hat{\pi}_{ij}$ is the posterior probability that $y_{ij} = 0$ is from the structural zero part. The posterior mean is

$$\hat{\lambda}_{ij} = E(\lambda_{ij}|y_{ij}) = \begin{cases} (1 - \hat{\pi}_{ij})\frac{\mu_i r}{r+M_{ij}\mu_i} & \text{if } y_{ij} = 0, \\ \frac{\mu_i(r+y_{ij})}{r+M_{ij}\mu_i} & \text{if } y_{ij} > 0. \end{cases} \tag{6}$$

2.3 | Group-level estimation

In addition to estimating the RR for individual AEs, we can also estimate the RR for each AE group. The MLE of the group-level parameters ($p_1, \dots, p_I, \mu_1, \dots, \mu_I, r$) can be obtained by fitting a ZINB regression model.

Let $\boldsymbol{Y}$ represent a vector of stacked counts for all AE-vaccine pairs in a group. If there are K AEs in the group and I vaccines, the dimension of $\boldsymbol{Y}$ is $K \times I$. The corresponding vector of expected frequencies is denoted by $\boldsymbol{M}$. The mean vector $\boldsymbol{\eta}$ of the non-zero component of ZINB is expressed as

$$ln\ \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\varphi} + ln\ \boldsymbol{M}, \tag{7}$$

where $\boldsymbol{X}$ is a binary design matrix with $I \times K$ rows and I columns, indicating the vaccine product for each count, and $\boldsymbol{\varphi} = (ln\ \mu_1, \ldots, ln\ \mu_I)^T$ is a vector of regression coefficients for the I vaccines.

The vector of probabilities $\boldsymbol{p}$ for the structural zero part is expressed as:

$$logit\ \boldsymbol{p} = \boldsymbol{X}\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha} = (logit\ p_1, \ldots, logit\ p_I)^T$.

The ZINB regression method can be implemented using *zeroinfl* function in the *pscl* R package.[21] The function estimates $(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\alpha}}, \hat{r})$ using the Broyden–Fletcher–Goldfarb–Shanno algorithm. Due to the invariance property of MLE, we have $(\hat{\mu}_1, \ldots, \hat{\mu}_I)^T = exp(\hat{\boldsymbol{\varphi}})$ and $(\hat{p}_1, \ldots, \hat{p}_I)^T = exp(\hat{\boldsymbol{\alpha}})/(1 + exp(\hat{\boldsymbol{\alpha}}))$.

2.4 | Statistical significance

As we test associations for many pairs of vaccine and AE groups, it is important to control for multiple comparisons. Adjustment methods can be used to control for the false discovery rate based on the p-values generated from these association tests.[22] In this paper, we use the $maxS = \max_{i,l} s_{il}$ $(i = 1, \ldots, I; l = 1, \ldots, L)$ as the test statistics (the maximum is taking over all vaccines and all AE groups). We define the adjusted p-value (ie, q-value) of the RR for a particular vaccine-AE group combination to be quantile of the observed s in the null distribution of $maxS$. By using this maximum statistics, the method is conservative in detecting multiple signals and controls the overall Type I error at the pre-specified level.[23,24]

The distribution of $maxS$ under the null hypothesis (H_0: $\lambda_{ij} = 1$ for all vaccine - AE pairs) is not analytically tractable and is obtained using the permutation test. For the VAERS dataset, the permutation test needs to account for the correlation between AEs mentioned in the same report. For this reason, we consider AEs observed on the same report as a single set and reshuffle the AE sets in the permutation test. As a simple example with two reports, if one report mentions one vaccine denoted by V_1 and three AEs, a, b, c, and the other report mentions two vaccines, V_2 and V_3 and two AEs, e and f, by reshuffling the AE sets, (a, b, c) might be linked to vaccine $(V_1, V_2\ )$ and (e, f) linked to vaccine V_1. By permuting the AE sets, we can shuffle the data while maintaining the correlation between AEs on the same report.

For each permuted dataset, we compute a value of $maxS$. By generating N permuted datasets (N is generally large, say 5000), we obtain an empirical null distribution of $N + 1$ $maxS$ values (including the $maxS$ value from the observed dataset). Let R denote the rank of the observed s in the null distribution of $maxS$, then the adjusted p-value of that vaccine-AE group combination is $1 - R/(N + 1)$. A small adjusted p-value indicates a strong association between the vaccine and AE group. Similarly, we can obtain adjusted p-value of RR for all individual AEs, by using the maximum of λ (here, the maximum is taking over all vaccines and all AEs) as the test statistics.

3 | SIMULATION

3.1 | Simulation I

We first conducted small simulation studies to investigate the performance of the zGPS.AO model when the group size is different, the expected counts (M_{ij}'s) have different magnitudes, and the data is not over-dispersed. Specifically, we selected one large AE group from MedDRA and three vaccines with the most frequent report numbers from VAERS. M_{ij}'s were determined from the contingency table (ie, row total times column total and divided by the total count), and true group-level parameters were determined by fitting the ZINB model to 3 AE groups of the largest size in VAERS data. That is, $(\mu_1, \mu_2, \mu_3) = (0.569, 0.829, 0.482)$, $(p_1, p_2, p_3) = (0.067, 0.260, 0.207)$, and $r = 3.3$. In each simulation, we generated λ_{ij}'s for all individual AEs based on formula (2), and then generated y_{ij}'s based on formula (1). We named this simulation

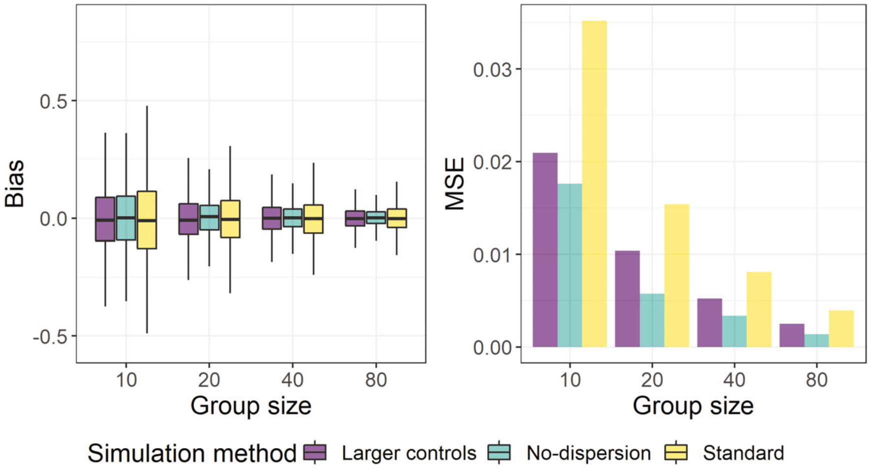

FIGURE 1 Bias (left) and MSE (right) of group-level RR estimates under three simulation setup (Standard, No-dispersion, and Large controls), each with 1000 simulated datasets.

setup as "Standard". Built on this standard setup, we considered two other scenarios: (1) multiply the expected counts by 5 (called this setup "Larger controls"), (2) generate count data from a zero-inflated Poisson (ZIP) model (called this setup as "No-dispersion"), which is equivalent to setting $r = \infty$ while keeping all other parameters the same. We varied the group size with $K = 10, 20, 40$, or 80 in each of the above three scenarios to test the impact of group size on the model performance.

To evaluate the model performance, we considered two commonly used metrics, bias (ie, the difference between the RR estimates and the true RR values) and the mean squared error (MSE) (ie, averaged squared difference between the RR estimates and the true RR values). These metrics were measured on both the group-level and the AE-level RRs. A good model should have a bias close to zero and a small MSE.

Figure 1 shows that our method accurately estimate the group-level RRs (s's). Both a larger group size and larger control counts reduced the bias and MSE, as a larger group allows more AEs to share information and larger control counts yield richer information available on each AE. Figure 2 shows the simulation results of the RR estimates for individual AEs. The accuracy increases when the group size increases from 10 to 20, as we have more information to share between λs. When the size reaches 20, we do not see further improvement in accuracy, which is likely due to diminishing returns. Furthermore, both Figure 1 and 2 demonstrate that our method is able to handle data without over-dispersion. In this case, ZINB automatically reduces to ZIP by a large $\hat{r}$.

3.2 | Simulation II

We conducted an extensive simulation study to mimic the real data in section 4. We generated the count data for a 10×1477 contingency table (ie, 10 vaccines and 1477 AEs); see details in Study I. The expected counts M_{ij}'s were calculated as row total times the column total and divided by the total counts (see Introduction). In this simulation, we compared zGPS.AO with an existing method called the Gamma-Poisson Shrinker (GPS) model.[3,4] GPS assumes that cell counts in the contingency table follow Poisson distributions with rate parameters (RR) representing AE-vaccine associations. The model assumes that RR are drawn from a mixture of two gamma distributions. One gamma distribution has values of RRs clustered at or below one (non-signals), and the other gamma distribution have values of RRs clustered at a rate above one (signals). GPS allows information borrowing between AEs with similar values of RRs rather than with similar disease biology (defined by MedDRA) as in zGPS.AO.

We designed three scenarios. In the **first scenario**, 1477 AEs were mapped to 42 AE groups based on the HLGT terms in MedDRA. Then λ_{ij}'s were generated from equation (2) with $p_i \sim Unif(0, 0.2)$, $\mu_i \sim \Gamma(5, 0.4)$, and $r = 5$. Finally, y_{ij}'s were simulated from equation (1). Parameters in these distributions were determined by applying the ZINB model to the VAERS data, and the produced count data was similar to the real data with approximately 26% zero counts. We applied two zGPS.AO models, one used HLGT terms in MedDRA to define the groups (AE group structure was correctly specified), and one used a randomly generated group structure (AE group structure was mis-specified; denoted as zGPS.AO*).

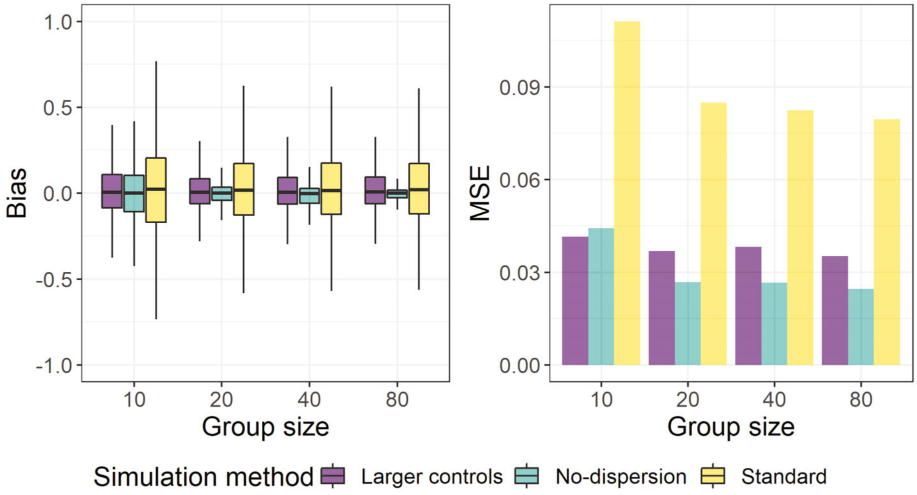

FIGURE 2 Bias (left) and MSE (right) of AE-level RR estimates in three simulation studies (Standard, No-dispersion, and Large controls), each with 1000 simulated datasets.

In the **second scenario**, each AE was simulated independently from a different gamma distribution (parameters were the same as in the first scenario); therefore, no group structure was imposed on the data. We used the HLGT terms to define AE groups in zGPS.AO. Therefore, the group structure was also mis-specified.

In the **third scenario**, λ_{ij}s were stimulated from a two-component gamma mixture model: $\Gamma(5, 0.2)$ with probability 0.95 and $\Gamma(5, 0.6)$ with probability 0.05. Under this setup, 95% AEs have RRs centered around 1 and 5% centered around 3 (signal AEs). In this scenario, the data generating model is the same as the GPS model.

We compared zGPS.AO with GPS. GPS enhances the simple use of the separate Poisson model by allowing "shrinkage" of similar λ_{ij}s towards each other. Compared to zGPS.AO, GPS does not use the AE ontology. We applied both models to each simulated data, and simulation results were averaged over 100 simulated datasets. Since GPS cannot estimate group-level parameters, we only compared the RR estimates, $\hat{\lambda}_{ij}$'s, for individual AEs. We calculated the mean square error (MSE) and the area under the curve (AUC), to evaluate the accuracy of the parameter estimates and the signal detection, respectively. The MSE was defined as the squared difference of the estimator $\hat{\lambda}_{ij}$ and the truth λ_{ij} averaged over all vaccine-AE pairs. For the AUC, we defined an AE j to be a true signal for vaccine i if $\lambda_{ij} > 2$, and we used $\hat{\lambda}_{ij}$ to construct the ROC curve.

Figure 3 shows the simulation results under the three scenarios. As shown in this figure, zGPS.AO performed significantly better than GPS in the scenario I and II, as demonstrated by the smaller MSE and larger AUC. The improved performance was likely due to information borrowing between similar AEs and appropriately handling the zero counts. When the ontology structure was mis-specified, the performance of zGPS.AO was reduced (zGPS.AO versus zGPS.AO*), but it was still better than GPS (zGPS.AO* versus GPS). The poor performance of GPS was likely due to its ignorance of the excessive zero counts. This finding highlights the importance of accommodating excessive zero counts in the model. In the last scenario, as expected, the GPS performed better than zGPS.AO with regard to MSE. However, to our surprise, the zGPS.AO and GPS demonstrated a very similar ability for signal detection, as evidenced by comparable AUCs.

4 | ANALYSIS OF VAERS DATASET

Study I. We used reports received from year 2005 to 2018 and restricted the age of the vaccine recipients between 2 to 49. We investigated AEs for 10 vaccines of interest, including FLU (inactivated influenza vaccine; trivalent or quadrivalent), FLUN (live attenuated influenza; trivalent or quadrivalent), HEP (Hepatitis B vaccines), HEPA (Hepatitis A vaccines), HEPAB (Hepatitis A + Hepatitis B), HPV4 (human papillomavirus 4-valent vaccine), HPV9 (human papillomavirus 9-valent vaccine), MMR (measles, mumps and rubella virus vaccine, live), TDAP (tetanus toxoid, reduced diphtheria toxoid and acellular pertussis vaccine, adsorbed), and VARCEL (Varivax-Varicella Virus, live). Those vaccines were selected from 84 types of vaccines based on their high report frequencies in the dataset and high level of public attention. All AEs were mapped to the "Preferred Terms" (PT) level of MedDRA, and we used the High Level Group Terms (HLGT) level of

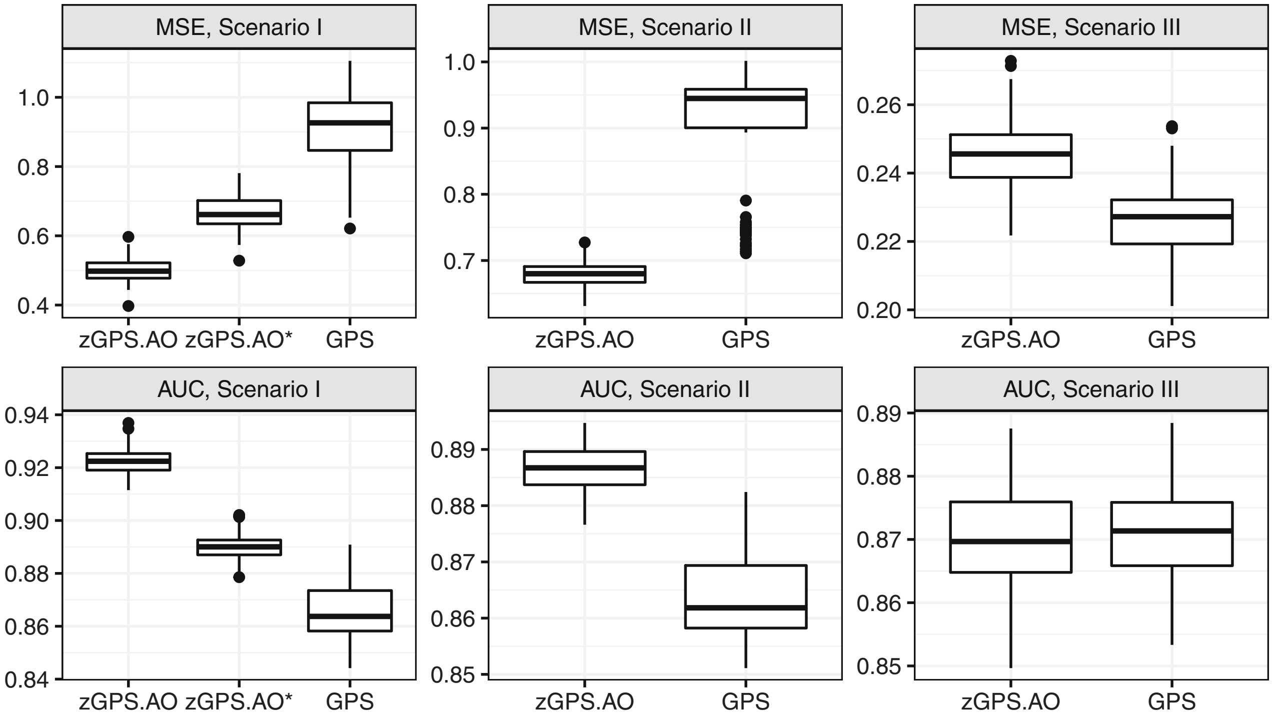

FIGURE 3 Simulation results of MSE and AUC from the three scenarios.

MedDRA to define AE groups. We removed AEs with a total frequency less than 20. If they were mentioned fewer than 20 times in the VARES database in 14 years, they were unlikely to be related to any vaccine. We also removed AE groups containing less than 15 AEs. Finally, we had 169,538 reports, 1477 AEs, and 42 AE groups.

The next step was to compute the count for each vaccine-AE pair. A traditional way to create the data pair is to get all vaccine-AE combinations in each report, regardless of the number of vaccines. If a report mentions two vaccines, A and B, and an AE of fever, the current strategy creates two pairs of data: A-fever and B-fever. However, with the presence of vaccine B, the link might not exist between vaccine A and fever, and likewise for the link between vaccine B and fever. We adopted a weighting strategy to handle the report with multiple vaccines.[25] Specifically, we assign a weight to each vaccine-AE pair. If there was a single vaccine in the report, the weight is one. If there were multiple vaccines, we weighted each vaccine-AE pair by the inverse of the number of vaccines mentioned in the report, assuming that the AE is linked to each vaccine with equal probability. To compute the count for a vaccine-AE pair, we summed over the weights and then rounded it to the nearest integer.

We applied both zGPS.AO and GPS to the final dataset. The RR estimates from both methods are highly correlated, and the Pearson correlation coefficient was 0.91. Next, we compared detected AE signals from both methods. In zGPS.AO, we defined an AE as a safety signal if the RR estimate was larger than 3 and q-value < 0.01, while in GPS, we required the lower bound of the 99% confidence interval to be larger than 3 (GPS doesn't provide a p-value or q-value). Based on the above decision rules, zGPS.AO and GPS detected 281 and 194 AE signals, respectively, among the 10 studied vaccines. Of them, 167 AE signals were detected by both methods. The full list of these AEs and their RR estimates can be found in the Supplementary Material.

Figure 4 shows RR estimates from zGPS.AO for all vaccine—AE group combinations. It visually helps us to identify headline-grabbing vaccine safety issues. For example, it seems that the HEPAB vaccine is associated with more Haematology investigations, and FLUN vaccine is associated with more respiratory system disorders. To further quantify the association, we defined an AE group having a safety problem if q-value < 0.01 and RR > 3. Based on this criteria, we detected three AE groups associated with FLUN (see Table 2). All the three groups are related to the respiratory system, while none of these groups are associated with FLU. Therefore, the nasal spray vaccine, relative to injection vaccine, is associated with an increased risk of respiratory system disorders. The individual AEs, Rhinitis, Nasal congestion, Sinus disorder, have been reported before,[26-28] whereas Epistaxis and Croup infectious might be new signals that need attention and validation in large healthcare databases. We also compared the AE profile with the combination of hepatitis A

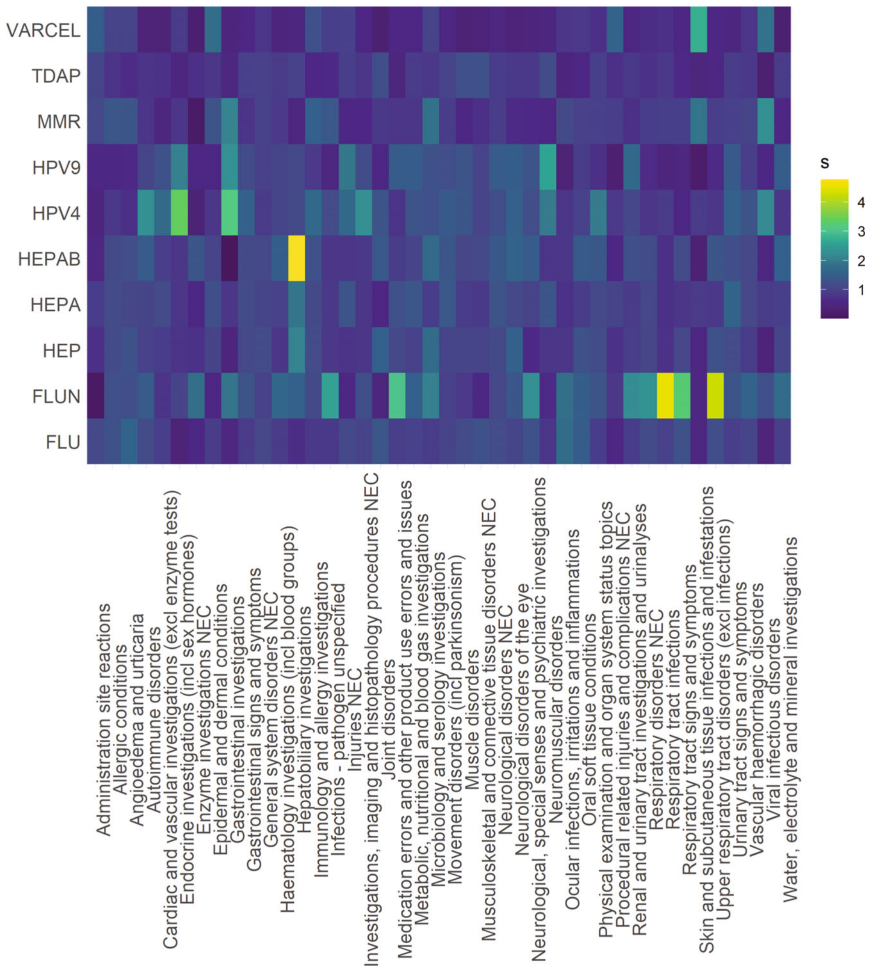

FIGURE 4 The RR heatmap of study I. Rows represent vaccines and columns represent AE groups. A brighter color represents a larger RR (ie, a stronger association between the vaccine and the AE group).

and B vaccine (HEPAB) to the monovalent hepatitis A (HEPA) and B vaccine (HEP) alone. Our method detected one signaled AE group (ie., Hepatobiliary investigations) with the combination vaccine (see Table 3), whereas this AE group was not associated with the HEPA or HEP alone. This finding might indicate an increased risk when Hepatitis A and B vaccines are combined, which needs further attention and validation in large healthcare databases. We have developed an interactive web app Rshiny to visualize results from study I.

Study II. We did another study comparing the risk of COVID-19 vaccines with FLU and FLUN. The purpose of this study is the compare the three types of COVID-19 vaccines, the BNT162b2 (Pfizer–BioNTech), the mRNA-1273 (Moderna), and the Ad26.COV2.S (Johnson & Johnson–Janssen, abbreviated as J & J), to influenza vaccines (FLU and FLUN). We collected VAERS reports received from January 1, 2019 to March 15, 2022, with recipients aged above 18. We removed AEs with a frequency of less than 20 and groups containing less than 15 AEs. The final dataset for analysis has a total of 714,330 reports and 3209 AEs, which were classified into 80 AE groups. We computed the counts for all vaccine-AE pairs using the same strategy as described in study I, and applied our zGPS.AO model to the finial dataset.

We applied both zGPS.AO and GPS to the final dataset. The RR estimates from both methods were highly correlated, and the Pearson correlation coefficient was 0.85. Based on the decision rules described in Study I, zGPS.AO and GPS detected 116 and 156 AE signals, respectively, among the three types of COVID-19 vaccines. Of them, 111 AE signals were detected by both methods. The full list of these AEs and their RR estimates can be found in the Supplementary Material.

Figure 5 shows the group-level RR's for study II. With the criteria of RR > 2 and q-value < 0.01, six AE groups were associated with FLU and FLUN, and one AE group was associated with COVID-19 vaccines. The AE group of Embolism and thrombosis (RR $= 2.14$, q-value < 0.001) is associated with the Johnson & Johnson–Janssen vaccine. This group of AEs has already caught public attention and official scrutiny.[29,30] We did not find an AE group associated with Pfizer–BioNTech and Moderna vaccines.

TABLE 2 AE groups associated with FLUN, along with top 5 significant AEs within each group.

AE group	RR (q-value)	Top 5 significant AEs	RR (q-value)
Respiratory tract infections	4.56 (<0.001)	Croup infectious	9.52 (<0.001)
		Influenza	8.95 (<0.001)
		Rhinitis	6.41 (<0.001)
		Pneumonia	6.33 (<0.001)
		Atypical pneumonia	6.10 (<0.001)
Upper respiratory tract disorders (excl infections)	4.26 (0.002)	Epistaxis	8.44 (<0.001)
		Nasal congestion	7.10 (<0.001)
		Nasal oedema	6.19 (<0.001)
		Paranasal sinus	5.53 (<0.001)
		Sinus disorder	5.22 (<0.001)
Respiratory tract signs and symptoms	3.26 (0.008)	Nasal discomfort	9.24 (<0.001)
		Rhinorrhoea	7.29 (<0.001)
		Sneezing	6.48 (<0.001)
		Sinus headache	5.65 (<0.001)
		Rhinalgia	5.01 (<0.001)

TABLE 3 AE groups associated with HEPAB, along with top 5 significant AEs within each group.

AE group	s (q-value)	Top 5 significant AEs	RR (q-value)
Hepatobiliary investigations	4.76 (0.001)	Aspartate aminotransferase increased	4.81 (<0.001)
		Hepatic enzyme increased	4.81 (<0.001)
		Liver function test abnormal	4.79 (<0.001)
		Alanine aminotransferase increased	4.78 (<0.001)
		Bilirubin urine	4.78 (<0.001)

Computation time. Our method has been implemented to take advantage of parallel processing. It took approximately 2 hours (1 minute for the parameter estimation and two hours for bootstrapping 1000 times) for studies I and II, running on a six-core AMD Ryzen 5 3600X 3.80 GHz 16 GB RAM x64 computer.

5 | DISCUSSION

VAERS has many limitations including (i) reporting bias, (ii) inconsistent data quality and completeness, (iii) lack of an unvaccinated comparison group, and (iv) the inability to assess if a vaccine caused an AE. However, due to its national scope, VAERS continues to serve as the nation's frontline post-licensure vaccine safety monitoring system.

In this article, we have developed a method to detect vaccine-associated AE signals while incorporating the AE ontology and accommodating excess zeroes (commonly seen in VAERS data). Our simulation studies have shown that zGPS.AO improves the accuracy of parameter estimation and signal detection. The AE ontology defines the similarity between AEs based on disease biology, allowing information borrowing between similar AEs. In this paper, we used MeddRA to define AE ontology. MedDRA is the largest dictionary for disease and symptoms, covering a large number of potential adverse events. A different AE ontology, containing different AE terms and different AE relationships, can be used when available. The ZINB distribution was used to model AE count data with excess zeros, which significantly improved the model fitting. Intuitively, the zero counts have two sources (i) structural zeros for the vaccine-AE pair that can never occur (eg, an AE at the injection site for a vaccine given orally), and (ii) zeros from the random sampling of the Poisson distribution for the vaccine-AE pair that is possible but has not been reported yet.

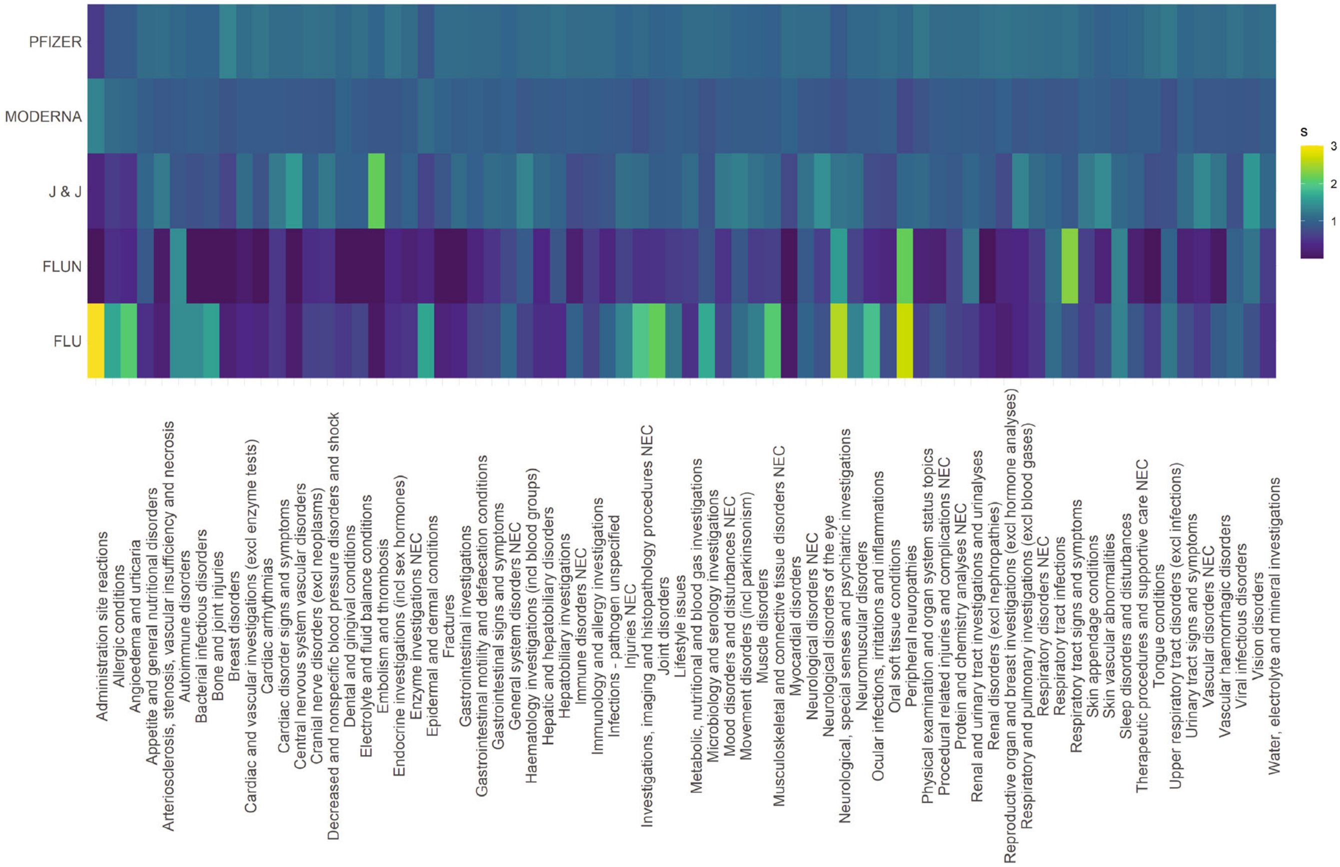

FIGURE 5 The RR heatmap of study II. Rows represent vaccines and columns represent AE groups. A brighter color represents a larger RR (ie, a stronger association between the vaccine and the AE group).

Given the large number of AEs in the passive reporting database, such as VAERS, existing methods of performing hypothesis testing for thousands of AEs are likely to identify hundreds of AE signals, which are hard to interpret in a biological context. Our method not only detects safety issues of individual AEs but also identifies AE groups of concern, which can provide one way to understand the mechanisms behind occurrences of AEs.

Our method has some limitations. First, zGPS.AO uses a two-level AE hierarchical structure in MedDRA. Further work will utilize a three-level hierarchy in MedDRA. To include the PT-HLGT-SOC hierarchy in the model, we can create an interconnected network for the HLGT AEs based on their relationships with the SOC AEs; that is, two HLGT terms are connected if they have the same SOC AEs. We can convert the relationship network into a graphical prior and assign it to the parameters in the negative binomial distributions. Using this graph prior, we can incorporate the three-level ontology structure into the signal detection method. Second, a single gamma distribution may not be sufficient to model AEs in a large group. In this article, we defined AE groups using AE terms on the HLGT level, typically containing 20-40 AEs per group. For studies using the SOC-level to define groups, typically containing hundreds of AEs, a more flexible distribution, such as the mixture Gamma distribution,[3] can be considered. Thirdly, in reprocessing the VAERS data, we removed AEs which were mentioned less than 20 times in 14 years as we believe that these AEs are likely reporting errors. The use of 20 is subjective. A slightly smaller or larger threshold can be used, although it is unlike to change the results.

Although this paper focuses on the VAERS database, the proposed methods generally apply to other databases which rely on passive reporting, such as FDA Adverse Events Reporting System (FAERS) and the Adverse Drug Reactions (ADR) database for conducting post-marketing drug safety surveillance.

ACKNOWLEDGEMENTS

Thanks to Dr. Kirsten Herold for proofreading the article and Michael Kleinsasser for assistance on developing the R package.

FUNDING INFORMATION

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R01AI158543. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The authors confirm that the data supporting the findings are publicly available. VAERS data is publicly available at https://vaers.hhs.gov/data.html. MedDRA data is publicly available at https://www.meddra.org/. The data is also available from the corresponding author upon reasonable request.

ORCID

Bangyao Zhao https://orcid.org/0000-0003-2482-391X
Lili Zhao https://orcid.org/0000-0002-6366-8206

REFERENCES

1. Varricchio F, Iskander J, DeStefano F, et al. Understanding vaccine safety information from the vaccine adverse event reporting system. *Pediatr Infect Dis J*. 2004;23:287-294.
2. Shimabukuro TT, Nguyen M, Martin D, DeStefano F. Safety monitoring in the vaccine adverse event reporting system (VAERS). *Vaccine*. 2015;33(36):4398-4405.
3. 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.
4. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA; 2001:67-76.
5. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf*. 2001;10:483-486.
6. Puijenbroek vEP, Bate A, Leufkens HG, Lindquist M, RO R, Egberts AC. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol Drug Saf*. 2002;11:3-10.
7. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol*. 1998;54:315-321.
8. Lansner ROA, Bate A, IRE ML, et al. Bayesian neural networks with confidence estimations applied to data mining. *Comput Stat Data Anal*. 2000;34:473-493.
9. Nóren GN, Bate A, Orre R, Edwards IR. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *BCPNN*. 2006;25:3740-3757.
10. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *PLoS ONE*. 2002;25:381-392.
11. Kulldorff M, Davis RL, Kolczak M, Lewis E, Lieu T, Platt R. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Anal*. 2011;30:58-78.
12. Davis RL, Kolczak M, Lewis E, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology*. 2005;16:336-341.
13. Li L, Kulldorff M. A conditional maximized sequential probability ratio test for pharmacovigilance. *Stat Med*. 2009;29:284-295.
14. Li R, Stewart B, Weintraub E, McNeil MM. Continuous sequential boundaries for vaccine safety surveillance. *Stat Med*. 2014;33:3387-3397.
15. Kulldorff M, Dashevsky I, Avery TR, et al. Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiol Drug Saf*. 2013;22:517-523.
16. Huang J, Cai Y, Du J, et al. Monitoring vaccine safety by studying temporal variation of adverse events using vaccine adverse event reporting system. *Anna Appl Stat*. 2021;15(1):252-269. doi:10.1214/20&hyphen;AOAS1393
17. Cai Y, Du J, Huang J, et al. A signal detection method for temporal variation of adverse effect with vaccine adverse event reporting system data. *BMC Med Inform Decis Mak*. 2017;17(2):93-100.
18. Mozzicato P. MedDRA: an overview of the medical dictionary for regulatory activities. *Pharm Med*. 2009;23:65-75.
19. Li S, Zhao L. Vaccine adverse event enrichment tests. *Stat Med*. 2021;40(19):4269-4278.
20. Khuri AI. Applications of Dirac's delta function in statistics. *Int J Math Edu Sci Technol*. 2004;35(2):185-195. doi:10.1080/00207390310001638313
21. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. *J Stat Software*. 2008;27(8):1-25. doi:10.18637/jss.v027.i08
22. Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *J Edu Behavior Stat*. 2002;27(1):77-83.

23. Huang L, Zalkikar J, Tiwari RC. A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *J Am Stat Assoc*. 2011;106(496):1230-1241. doi:10.1198/jasa.2011.ap10243
24. Ding Y, Markatou M, Ball R. An evaluation of statistical approaches to postmarketing surveillance. *Stat Med*. 2020;39(7):845-874. doi:10.1002/sim.8447
25. Zhao L, Lee S, Li R, Ong E, He Y, Freed G. Improvement in the analysis of vaccine adverse event reporting system database. *Stat Biopharm Res*. 2020;12:1-14. doi:10.1080/19466315.2020.1764862
26. Baxtera R, Eatona A, Hansena J, Aukesa L, Caspardb H, Amboseba CS. Safety of quadrivalent live attenuated influenza vaccinein subjects aged 2–49 years. *Vaccine*. 2017;35:1254-1258.
27. Haber P, Moro PL, Cano M, Lewis P, Stewart B, Shimabukuro T. Post-licensure surveillance of quadrivalent live attenuated influenza vaccine United States, Vaccine adverse event reporting system(VAERS), July 2013–June 2014. *Vaccine*. 2015;33:1987-1992.
28. Lambkin-Williams R, Gelder C, Broughton R, et al. An intranasal proteosome-adjuvanted trivalent influenza vaccine is safe, immunogenic and efficacious in the human viral influenza challenge model. Serum IgG and mucosal IgA are important correlates of protection against illness associated with infection. *PLoS one*. 2016;11:e0163089.
29. Shimabukuro TT, Cole M, Su JR. Reports of anaphylaxis after receipt of mRNA COVID-19 vaccines in the US—December 14, 2020-January 18, 2021. *Jama*. 2021;325(11):1101-1102.
30. Woo EJ, Mba-Jonas A, Dimova RB, Alimchandani M, Zinderman CE, Nair N. Association of Receipt of the Ad26. COV2. S COVID-19 vaccine with presumptive Guillain-Barré syndrome, February-July 2021. *Jama*. 2021;326(16):1606-1613.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhao B, Zhao L. Mining adverse events in large frequency tables with ontology, with an application to the vaccine adverse event reporting system. *Statistics in Medicine*. 2023;42(10):1512-1524. doi: 10.1002/sim.9684

APPENDIX A. MATHEMATICAL DETAILS

A.1 Model setup

The NB distribution as a Poisson-gamma mixture,

$$\begin{aligned} y &\sim Pois(\lambda), \\ \lambda &\sim \Gamma(r, \mu/r), \end{aligned} \tag{A1}$$

where the parameters are r (the dispersion parameter) and μ (the mean parameter), and the λ is a the latent random variable.

The mean and variance of NB distribution is given by,

$$E(y) = E(E(y|\lambda)) = E(\lambda) = \mu,$$

$$\begin{aligned} Var(y) &= Var(E(y|\lambda)) + E(Var(y|\lambda)) \\ &= Var(\lambda) + E(\lambda) \\ &= \mu + \mu^2/r. \end{aligned}$$

The marginal p.d.f of y can be obtained by integrating out λ in the joint p.d.f of (y, λ):

$$\begin{aligned} f(y|r,\mu) &= \int_0^\infty f(y,\lambda|r,\mu)d\lambda \\ &= \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!}\frac{(r/\mu)^r}{\Gamma(r)}\lambda^{r-1}exp(-r\lambda/\mu)d\lambda \\ &= \frac{\Gamma(r+y)}{y!\Gamma(r)}\left(\frac{r}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y \\ &:= NB(y|r,\mu). \end{aligned}$$

To extend A1 to ZINB distribution, we add the zero part probability p to the model, and also an offset M. (The offset is treated as a constant)

$$\begin{aligned} &y \sim Pois(M\lambda), \\ &\lambda \sim \begin{cases} 0 & \text{with probability } p, \\ \Gamma(r, \mu/r) & \text{with probability } 1-p. \end{cases} \end{aligned} \tag{A2}$$

Under this formulation, the marginal p.m.f of y is

$$\begin{aligned} f(y|r,p,\mu) &= \begin{cases} p+(1-p)NB(0|r,M\mu) & \text{if } y=0, \\ (1-p)NB(y|r,M\mu) & \text{if } y>0, \end{cases} \\ &:= ZINB(y|r,p,M\mu). \end{aligned}$$

A.2 Appendix: Derive the posterior distribution of AE-level RRs

We can write the p.d.f of λ as the following.

$$\pi(\lambda) = p\delta(\lambda) + (1-p)\Gamma(\lambda|r,\mu/r). \tag{A3}$$

If $y > 0$, the λ is from the non-zero part, and the posterior distribution of λ is a Gamma distribution.

$$\pi(\lambda|y) \propto \lambda^{r+y-1} exp\left(\frac{-\lambda}{\frac{\mu}{r+M\mu}}\right)$$

which is,

$$\lambda|y \sim \Gamma\left(r+y, \frac{\mu}{r+M\mu}\right).$$

Therefore,

$$\hat{\lambda} = E(\lambda|y) = \frac{\mu(r+y)}{r+M\mu}.$$

In the case of $y = 0$, the posterior distribution is

$$\begin{aligned} \pi(\lambda|y=0) &\propto f(y=0|\lambda)\pi(\lambda) \\ &= \hat{\pi}\delta(\lambda) + (1-\hat{\pi})\Gamma\left(\lambda|r, \frac{\mu}{r+M\mu}\right), \end{aligned}$$

where

$$\hat{\pi} = \frac{p}{p+(1-p)\left(\frac{r}{r+M\mu}\right)^r}$$

is the posterior probability that $y = 0$ comes from the zero component. This shows $\lambda|y = 0$ is a gamma—zero mixture, and the posterior mean is as the following.

$$\hat{\lambda} = E(\lambda|y=0) = (1-\hat{\pi})\frac{\mu r}{r+M\mu}.$$