scientific reports

OPEN

A tree-based scan statistic for zero-inflated count data in post-market drug safety surveillance

Goeun Park & Inkyung Jung

After new drugs enter the market, adverse events (AE) induced by their use must be tracked; rare AEs may not be detected during clinical trials. Some organizations have been collecting information on suspected drugs and AEs via a spontaneous reporting system to conduct post-market drug safety surveillance. These organizations use the information to detect a signal representing potential causality between drugs and AEs. The drug and AE data are often hierarchically structured. Accordingly, the tree-based scan statistic can be used as a statistical data mining method for signal detection. Most of the AE databases contain a large number of zero-count cells. Notably, not only an observational zero from the Poisson distribution, but also a true zero exists in zero-count cells. True zeros represent theoretically impossible observations or possible but unreported observations. The existing tree-based scan statistic assumes that all zeros are zero-valued observations from the Poisson distribution. Therefore, true zeros are not considered in the modeling, which can lead to bias in the inferences. In this study, we propose a tree-based scan statistic for zero-inflated count data in a hierarchical structure. According to our simulation study, in the presence of excess zeros, our proposed tree-based scan statistic provides better performance than the existing tree-based scan statistic. The two methods were illustrated using Korea Adverse Event Reporting System data from the Korea Institute of Drug Safety and Risk Management.

After new drugs enter the market, the adverse events (AE) induced by their use must be tracked because rare AEs may not be detected during clinical trials owing to short trial durations, limited sample sizes, or limited population representation. Once drugs are commercialized, they are used in different ways and by more people than those covered during clinical trials. Accordingly, drug safety must be monitored even after commercialization to identify AEs that may not have been identified previously[1–7].

Drug and vaccine safety monitoring systems have traditionally been based on spontaneous reporting systems, such as the US Food and Drug Administration's Adverse Event Reporting System (AERS), the US Vaccine Adverse Event Reporting System (VAERS), and VigiBase, the World Health Organization's (WHO) global Individual Case Safety Reports database. AERS is a large database supporting the US Food and Drug Administration's program for monitoring drug safety; VAERS helps monitor vaccine-related AEs and is maintained by the US Center for Disease Control and Prevention and the US Food and Drug Administration; and VigiBase is managed by the Uppsala Monitoring Centre (UMC) on behalf of the WHO. VigiBase receives individual case safety reports from 80 countries. In South Korea, the Korea Institute of Drug Safety and Risk Management provides information on AEs collected through the Korea Adverse Event Reporting System (KAERS) to the UMC. These spontaneous reporting systems play an important role in detecting AE signals in post-market drug safety surveillance[8,9].

Disproportionality data mining methods have been used to analyze these databases to identify signs that certain drugs may be posing unrecognized safety hazards. Frequentist methods, such as the proportional reporting ratio[10], relative odds ratio[11], Yule's test[12], chi-squared test[13], and likelihood ratio test (LRT)[14], and Bayesian methods, including the Bayesian confidence propagating neural network[15], multi-item gamma Poisson shrinker[16], and simplified Bayes (sB) methods[15–19] are often used to detect drugs with previously unrecognized AE[16,20–25].

In pharmacovigilance data, AE information uses adverse reaction terms, which have a hierarchical structure. For example, as shown in Fig. 1, the WHO Adverse Reaction Terminology (WHO-ART) developed for the WHO drug monitoring program has a four-level hierarchical structure. (https://www.who-umc.org) Owing to this type

Division of Biostatistics, Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea. ✉email: ijung@yuhs.ac

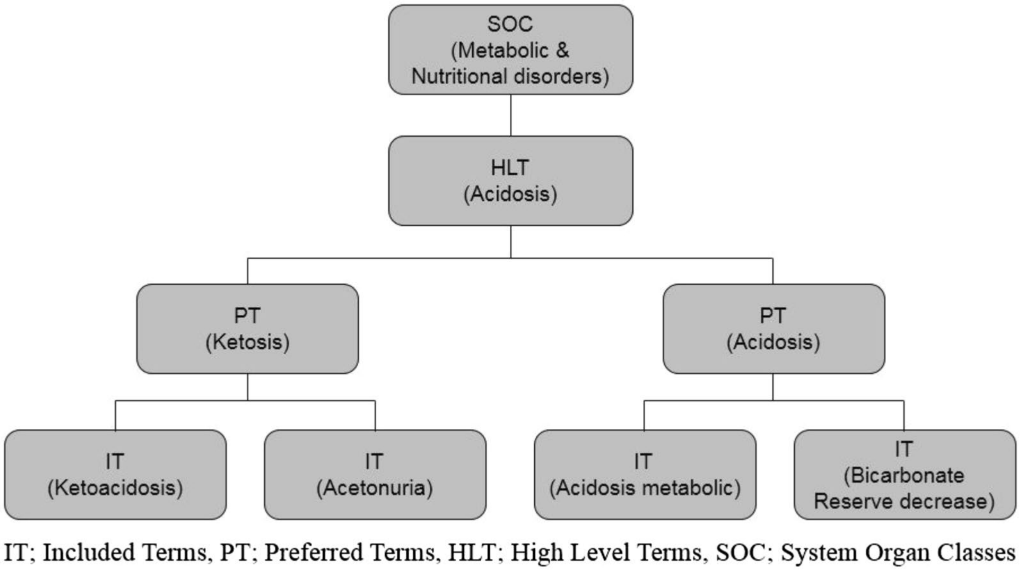
SOC
(Metabolic & Nutritional disorders)
HLT
(Acidosis)
PT
(Ketosis)
PT
(Acidosis)
IT
(Ketoacidosis)
IT
(Acetonuria)
IT
(Acidosis metabolic)
IT
(Bicarbonate Reserve decrease)
IT; Included Terms, PT; Preferred Terms, HLT; High Level Terms, SOC; System Organ Classes

Figure 1. WHO-ART structure.

of structure, it is difficult to determine the level of AE definition that should be used during data mining. To solve the problem, tree-based scan statistics, which find signals at each level of AEs in the form of a hierarchical tree, have been proposed by Kulldorff et al.[26] and have been recently used by some researchers to detect AE signals[27–29]. The tree-based scan statistic is distinct from most disproportionality methods; it is based on scan statistical theory and uses a hierarchical diagnosis tree to simultaneously assess risk at any level of granularity, adjusting for a multiple testing problem in several overlapping evaluated groups[7,26,30].

Most of these AE databases have large numbers of zero-count cells. For example, AERS data from 2006 to 2011 show that the percentage of zero-count cells by the drug ranges from 50 to 99.99%[31]. However, based on KAERS data from 2012 to 2016, the percentage of zero-count cells by the drug ranges from 75 to 100%. Zero-count cells may contain not only zero-valued observations from the Poisson distribution, but also true zeros, which represent theoretically impossible observations or possible but unreported observations. Data with a large number of zeros cannot be assumed to have a Poisson distribution as some zeros are true zeros. The distribution of such data is typically more dispersed than the Poisson distribution, resulting in equality between the variance and the mean of the distribution. To solve this problem, the zero-inflated Poisson (ZIP) model proposed by Lambert[32] can be used. Huang et al.[31,33] proposed a zero-inflated Poisson model based likelihood ratio test (ZIP-LRT) method as an extended version of LRT, a frequentist data mining method. Further, Hu et al.[24] developed the zero-inflated Poisson simplified Bayes method and the zero-inflated Poisson Dirichlet process method, which are Bayesian data mining methods.

The existing tree-based scan statistic assumes all zero values are zero-valued observations from the Poisson distribution. As a result, true zeros are not considered in the modeling, which can lead to bias in the inferences. Therefore, in this study, we proposed a new tree-based scan statistic using the ZIP model for data with excess zeros in a hierarchical structure.

In section "A tree-based scan statistic", we introduce the existing tree-based scan statistic. In section "A tree-based scan statistic for zero inflated count data", we propose a tree-based scan statistic for zero-inflated count data. In section "Simulation study", a simulation study to evaluate the performance of the proposed method is presented. In section "Real data", the two methods are compared through a real data example. Finally, in section "Conclusion and discussion", we summarize the results and conclude with our recommendations.

Hierarchical diagnosis tree. The tree-based scan statistic uses hierarchical classification systems to represent clinical concepts, such as drugs, procedures, or diagnoses[30]. To code adverse drug reactions in postmarket drug surveillance, medical terminologies, such as Medical Dictionary for Regulatory Activities (MedDRA) and WHO-ART, are used. In the KEARS data, WHO-ART is used to code the AEs.

WHO-ART is the terminology for coding clinical information related to pharmacotherapy and is commonly used for coding the AEs. When new drugs and new symptoms create new terms that incorporate them, the structure of the terms is updated to include the newly integrated terms while retaining their previous relationships and the existing structure of terms. WHO-ART has a four-level hierarchical structure, which consists of System Organ Class (SOC), High Level Terms (HLT), Preferred Terms (PT), and Included Terms (IT). The highest level, the SOC, corresponds to body systems and organs, which contain grouping terms. The HLT is used to group related or similar PTs, but all PTs are not grouped into the HLT. The PTs are principal terms used to describe AEs and the ITs are synonyms of the PTs, which help in the search for the PTs. An example of the WHO-ART is shown in Table 1.

SOC	PT	IT	Diagnosis	HLT
0800			Metabolic and nutritional disorders	
0800	0363		Acidosis	0363
0800	0363	003	Bicarbonate reserve decreased	0363
0800	0363	004	PH reduced	0363
0800	0363	005	Acidosis Metabolic	0363
0800	0363	006	Blood bicarbonate decreased	0363
0800	0363	007	Blood PH decreased	0363
0800	0363	008	Acidosis hyperchloraemic	0363
0800	0364		Acidosis lactic	0363
0800	0364	003	Lactate blood increase	0363
0800	0393		Ketosis	0363
0800	0393	003	Ketoacidosis	0363
0800	0393	004	Acetonuria	0363
0800	0393	005	Acetone breath	0363
0800	0393	006	Acetonaemia	0363
0800	0393	007	Diabetic ketoacidosis	0363
0800	1465		Acidosis respiratory	0363
0800	1465	002	Blood carbon dioxide increased	0363

Table 1. Example of WHO-ART.

A tree-based scan statistic

Review of a tree-based scan statistic.

The tree-based scan statistic is a statistical data mining method that has been used for signal detection in a hierarchically structured data, such as a classification system for coding AEs. This statistic searches signals at any level of AE definitions, called leaves. Each leaf contains information on the total number of patients with a specific AE and the number of patients with a specific AE from a certain drug. Mutually-related leaves are grouped into a higher level, called a node. Of note, a cut defines a branch of the tree where a node or a leaf may have more events than expected.

The tree-based scan statistic method considers all possible cuts. For each cut, the total number of AEs from all drugs and a certain drug are respectively calculated for the leaves within that cut. The test statistic is generated by a likelihood function in which risk is estimated separately for the leaves defined by the cut and those outside of the cut[26,34,35].

Let c_i be the observed number of patients with ith AE potentially caused by a certain drug in leaf i and n_i be the total observed number of patients with i th AE in leaf i. For a rare disease, with covariates ignored, c_i is approximately Poisson distributed with mean $n_i\lambda_i$, where λ_i is the probability that i th AE is caused by a certain drug. For all leaves on the tree, let $C = \sum_{i=1}^{I} c_i$ and $N = \sum_{i=1}^{I} n_i$ where I is the number of all leaves in the tree. For each cut G, a leaf or a group of related leaves, let $c_G = \sum_{i \in G} c_i$ and $n_G = \sum_{i \in G} n_i$. R is the rest of the leaves except those included in G. The following null hypothesis $H_0 : \lambda_G = \lambda_0$ and the alternative hypothesis $H_a : \lambda_G > \lambda_R$ are considered. The null hypothesis suggests that the probability that AEs in a cut G due to a certain drug are not lower or higher than that of all AEs. The alternative hypothesis is that at least one cut is defined by a set G such that $\lambda_G > \lambda_R$, where R is a group of the remaining leaves.

Of note, the analysis is only concerned with C, as the total number of AEs represented by the tree is not of interest. In fact, only the relative distribution between the different AEs is relevant. The likelihood can then be expressed as $L(\lambda, \boldsymbol{c}) = \prod_i \left(\frac{n_i\lambda_i}{\sum_i n_i\lambda_i}\right)^{c_i}$ using a multinomial distribution. As a maximum likelihood estimator (MLE) of λ_G/λ_R is $\frac{c_G/n_G}{(C-c_G)/(N-n_G)}$ given G, a likelihood ratio test statistic is $T = \frac{\max_{G,\lambda_G>\lambda_R} L(\lambda,\boldsymbol{c})}{\max_{\lambda_G=\lambda_R} L(\lambda,\boldsymbol{c})} = \left(\frac{N}{C}\right)^C \max_G \left(\frac{c_G}{n_G}\right)^{c_G} \left(\frac{C-c_G}{N-n_G}\right)^{C-c_G}$ when $\frac{c_G}{n_G} > \frac{C-c_G}{N-n_G}$; otherwise, the statistic is 1. The log-likelihood ratio-based test statistic is given by

$$\log T = \max_G \left\{ c_G log\left(\frac{c_G}{n_G}\right) + (C - c_G) log\left(\frac{C - c_G}{N - n_G}\right) \right\} \times I\left(\frac{c_G}{n_G} > \frac{C - c_G}{N - n_G}\right),$$

where $I()$ is the indicator function[26].

Hypothesis testing.

To calculate the test statistic T, the likelihood of each possible cut was determined. The cut, which is maximizing the likelihood ratio value, is defined as the most likely cut; the likelihood ratio value is defined as the test statistic T. As the null distribution of the test statistic is unknown, it is produced using the Monte Carlo simulation[36]. Given the total number of patients with AEs from a certain drug, a large number of random data sets was created under the null hypothesis, and the test statistics for each random data set and the real data were calculated. The obtained test statistics for random datasets were compared to the test statistic for the real data. The P-value was calculated using the equation: rank/(1 + B), where rank is the relative position of the test statistic for the real data among the test statistics for the random data sets and B is the number of Monte Carlo replications.

A tree-based scan statistic for zero-inflated count data

In the presence of excess zero, the Poisson model tends to underestimate the observed dispersion. In this case, the ZIP model can be employed as one of the approaches to resolve the problem as this model is more flexible than the Poisson model. If the number of ith AE with a certain drug C_i follows the ZIP model, with the probability p of a true zero and the average number of events $n_i\lambda_i$, $C_i \sim \text{ZIP}(p, n_i\lambda_i)$, the mean and variance can be expressed as $E(C_i|p, n_i\lambda_i) = (1-p)n_i\lambda_i$ and $V(C_i|p, n_i\lambda_i) = (1-p)n_i\lambda_i(1+pn_i\lambda_i)$. It can also be expressed as $V(C_i|p, n_i\lambda_i) = E(C_i|p, n_i\lambda_i)(1+pn_i\lambda_i)$; thus, $V(C_i|p, n_i\lambda_i) > E(C_i|p, n_i\lambda_i)$ when $p > 0$.

As the ZIP model has an additional parameter relative to the tree-based scan statistic, its mean is smaller than that of the Poisson model. Thus, the ZIP model correctly calculates a reduced number of ith AEs with a certain drug due to the presence of true zeros.

Given the parameters p and $n_i\lambda_i$, the probability of $C_i = c_i$ is described as follows:

$$P(C_i = c_i|p, n_i\lambda_i) = \begin{cases} p + (1-p)e^{-n_i\lambda_i} & , c_i = 0 \\ (1-p)\frac{e^{-n_i\lambda_i}(n_i\lambda_i)^{c_i}}{c_i!} & , c_i > 0. \end{cases}$$

For the tree-based ZIP scan statistic, the hypotheses of interest are the same as those in section "Review of a tree-based scan statistic". The zeros are assumed to be known, whether or not they are true zeros, as it is difficult to find a closed form of MLE when the nature of each zero is unknown. As tree-based scan statistics are based on scan statistic theory, the methodology of Cançado et al.[37], who proposed a spatial scan statistical method for zero-inflated Poisson processes, was employed.

We consider a vector $\delta = (\delta_1, \ldots, \delta_I)$ where $\delta_i = 1$ for a true zero in leaf i and $\delta_i = 0$ for an observational zero in leaf i. δ_i s are Bernoulli random variables with the probability p of a true zero. Given a set of observations $\delta = (\delta_1, \ldots, \delta_I)$ that are bivariate data such that (C_i, δ_i), $i = 1, \ldots, I$, the likelihood function for set G can be expressed as

$$L(p, \lambda_R, \lambda_G) = \left[\prod_{i \in G} p^{d_i}\left[(1-p)\frac{e^{-n_i\lambda_i}(n_i\lambda_G)^{c_i}}{c_i!}\right]^{(1-d_i)}\right]\left[\prod_{i \notin G} p^{d_i}\left[(1-p)\frac{e^{-n_i\lambda_i}(n_i\lambda_R)^{c_i}}{c_i!}\right]^{(1-d_i)}\right].$$

When δ_i s are known, the MLEs under the null hypothesis are $\widehat{\lambda}_0 = \frac{\sum_{i=i}^{I} c_i(1-d_i)}{\sum_{i=i}^{I} n_i(1-d_i)}$ and$\widehat{p}_0 = \frac{\sum_{i=i}^{I} d_i}{I}$. However, under the alternative hypothesis, the MLEs are $\widehat{\lambda}_G = \frac{\sum_{i \in G} c_i(1-d_i)}{\sum_{i \in G} n_i(1-d_i)}$, $\widehat{\lambda}_R = \frac{\sum_{i \notin G} c_i(1-d_i)}{\sum_{i \notin G} n_i(1-d_i)}$, and$\widehat{p} = \frac{\sum_{i=i}^{I} d_i}{I}$.

When δ_i s are unknown, an expectation–maximization (EM) algorithm is used to find the MLEs of $\lambda_0, \lambda_G, \lambda_R, p_0$ and p. In the expectation step (E-step), the expected value of δ_i, given C_i, is calculated using the following formula:

$$\widehat{\delta}_i^{(m)} = \frac{\widehat{p}^{(m)}}{\widehat{p}^{(m)} + (1 - \widehat{p}^{(m)})e^{-n_i\widehat{\lambda}_0^m}} I(c_i = 0), i = 1, \ldots, I.$$

Under H_a, $\widehat{\lambda}_0$ is considered $\widehat{\lambda}_G$ and $\widehat{\lambda}_R$ in each cut G and the remaining leaves R, respectively.

In the maximization step (M-step), the MLEs of $\lambda_0, \lambda_G, \lambda_R, p_0$, and p are updated via the equations with d_i replaced by $\widehat{\delta}_i^{(m)}$ when δ_i s are known. Until the maximum likelihood estimates for each possible cut G converge, the above E- and M-steps are performed repeatedly. To perform a faster calculation, we used the 'zeroinfl' function in the R package "pscl"[38]. For the possible candidate cuts, this process should be conducted and the most likely cut should be determined.

The likelihood ratio for cut G can be expressed as

$$LR_G = \frac{\left[\frac{\sum_{i \in G} c_i(1-d_i)}{\sum_{i \in G} n_i(1-d_i)}\right]^{\sum_{i \in G} c_i(1-d_i)} \left[\frac{\sum_{j \notin G} c_j(1-d_j)}{\sum_{j \notin G} n_j(1-d_j)}\right]^{\sum_{j \notin G} c_j(1-d_j)}}{\left[\frac{\sum_{i=1}^{I} c_i(1-d_i)}{\sum_{i=1}^{I} n_i(1-d_i)}\right]^{\sum_{i=1}^{I} c_i(1-d_i)}} \times I\left(\frac{\sum_{i \in G} c_i(1-d_i)}{\sum_{i \in G} n_i(1-d_i)} > \frac{\sum_{j \notin G} c_j(1-d_j)}{\sum_{j \notin G} n_j(1-d_j)}\right).$$

Thereafter, the maximum likelihood ratio is defined as the test statistic, $T = \max_G LR_G$.

As it is impossible to know the null distribution of the likelihood ratio test statistic T, Monte Carlo hypothesis testing was conducted to assess statistical significance[37].

Simulation study

Data generating process and performance assessment measures. We conducted a simulation study to assess the performance of the proposed tree-based scan statistic for zero-inflated count data (TreeScan-ZIP) and the existing tree-based scan statistic (TreeScan-Poisson). For the simulation study, datasets with the hierarchical structure where AEs can be expressed in terms of WHO-ART SOCs and PTs were generated. Only 105 of the 1292 AEs in the PT terms were considered to reduce computation time. Different artificial true signals and true zeros were generated using a tree with 105 leaves and 9 nodes. The total numbers of patients with each AE varied from 10 to 4670. The total number of patients in all leaves of the tree was 19,920 and the total number of patients with AEs from a certain drug was 640.

First, true zeros ($\delta_i = 1$) were randomly allocated using the Bernoulli distribution with the probability p, where p is the percentage of the true zero leaves. Thereafter, for each iteration, the total number of patients with AEs from a certain drug, that is $C = \sum_{i=1}^{I} c_i$, was randomly assigned to the leaves on the tree as multinomial, with probabilities proportional to the relative risk. The relative risk of ith leaf was computed as $\frac{c_i/n_i}{C/N}, i = 1, \ldots, I$. For true zero leaves, $(\delta_i = 1), c_i = 0$. If the ith leaf was not a true zero, the dataset was generated using a multinomial distribution. Under H_0, the vector $\boldsymbol{C} = (c_1, \ldots, c_I)$ follows a multinomial distribution with parameters $\boldsymbol{C}$ and $\boldsymbol{p}$, where $\boldsymbol{p} = \left(\frac{n_1}{N}, \ldots, \frac{n_I}{N}\right)$. Under H_a, $\boldsymbol{p} = \left(\frac{rr_1 \frac{n_1}{N}}{\sum_{i=1}^{I} rr_i \frac{n_i}{N}}, \ldots, \frac{rr_I \frac{n_I}{N}}{\sum_{i=1}^{I} rr_i \frac{n_i}{N}}\right)$, where $rr_1, \ldots, rr_I$ are the relative risks of all types of AEs. The relative risk of the randomly selected true signal leaves ranged from 3, 4, and 2 to 6; however, for the other leaves, except the true zero leaves, the relative risk was equal to 1.

Based on the total number of cases, $C = 640$, we considered 0, 10, 30, 50, and 70 for the number of true zero leaves, and 1%, 3%, 5%, and 10% for the true signal leaves with the relative risk (RR). All possible combinations were simulated.

To evaluate the performance of the two methods, we computed type I error, power, sensitivity, and positive predicted value (PPV). First, the critical value T^* was obtained from 10,000 random datasets under H_0 by the Monte Carlo replications for each scenario according to the number of true zeros (0, 10, 30, 50, 70). Thereafter, B random datasets were generated under H_0 and H_a to calculate type I error, power, sensitivity, and PPV. For each of the B random datasets, test statistic $T_k, k = 1, \ldots, B$, was calculated using both methods.

Thereafter, type I error and power were estimated using

$$\text{Type I error} = \frac{\sum_{k=1}^{B} I(T_k > T^* | H_0)}{B}$$

$$\text{Power} = \frac{\sum_{k=1}^{B} I(T_k > T^* | H_a)}{B}.$$

Sensitivity and PPV for each random datasets are expressed as

$$\text{Sensitivity} = \frac{\#\,\text{of}\,(\text{detected signal} \cap \text{true signal})}{\#\,\text{of}\,(\text{true signal})},$$

$$\text{PPV} = \frac{\#\,\text{of}\,(\text{detected signal} \cap \text{true signal})}{\#\,\text{of}\,(\text{detected signal})}.$$

Overall sensitivity and PPV were calculated as the average of sensitivity and PPV over B' random datasets, where $B' = \sum_{k=1}^{B} I(T_k > T^*)$.

Results. The results obtained using the simulated data are presented in Table 2. The type I errors for the TreeScan-Poisson and TreeScan-ZIP methods were close to 0.05, except when the data had a Poisson distribution. The type I error of the TreeScan-Poisson method was above the nominal significance level of 0.05, while the type I error of the TreeScan-ZIP method tended to be less than 0.05.

When the data did not include true zeros (i.e., the data were generated from the Poisson distribution), the TreeScan-Poisson and TreeScan-ZIP methods produced similar power, sensitivity, and PPV estimates.

The TreeScan-ZIP method was identified to produce higher power and sensitivity estimates than the TreeScan-Poisson method when the number of true zeros was greater than or equal to 10. In the presence of zero inflation, when the number of true signals was greater than or equal to 5 and the RR was high, the PPV of the TreeScan-Poisson method was 1.0. The TreeScan-Poisson method could detect highly significant cuts, resulting in a small number of detected signals, which indicated high PPV and low sensitivity.

The TreeScan-ZIP method performed better than the TreeScan-Poisson in every dataset with true zero. The estimated power was almost 1.0 and the PPV was greater than 0.98 when the number of true zeros was greater than or equal to 10 and the number of true signals was greater than or equal to 5. The TreeScan-ZIP method was more sensitive than the TreeScan-Poisson method. The sensitivity and PPV of the TreeScan-ZIP method became higher with higher RR. When two true signals existed, both methods had a relatively low power; however, the power of the TreeScan-ZIP method increased as the number of true zeros and RR increased.

The simulation study showed that in the presence of zero inflation, the TreeScan-ZIP method performed better than the TreeScan-Poisson method.

Real data

Korea adverse event reporting system data. KAERS is a spontaneous AE reporting system maintained by the Korea Institute of Drug Safety and Risk Management (https://www.drugsafe.or.kr). Consumers, Healthcare Professionals, Regional Pharmacovigilance Centers (RPVCs), and pharmaceutical companies can report suspected drug information and AE information using the KAERS. RPVCs evaluate causality between the suspected drug and AE and report them to KIDS. The information is then stored in the KAERS as an individual case safety report (ICSR), which contains information on suspected drug, AE, causal relationship, and demo-

True zero	True signal	RR	TreeScan-Poisson			TreeScan-ZIP		
			Power*	Sensitivity	PPV	Power*	Sensitivity	PPV
0	0		0.043			0.043		
	2	3	0.061	0.144	0.275	0.061	0.144	0.275
	2	4	0.086	0.260	0.490	0.086	0.261	0.491
	2	(3.8, 6)	0.125	0.343	0.654	0.126	0.339	0.648
	6	3	0.833	0.176	0.976	0.833	0.176	0.976
	6	4	0.988	0.192	0.984	0.989	0.192	0.983
	6	(3.8, 6)	1.000	0.209	0.987	1.000	0.209	0.986
	7	3	1.000	0.379	0.995	1.000	0.380	0.995
	7	4	1.000	0.429	0.997	1.000	0.429	0.997
	7	(3.8, 6)	1.000	0.451	0.997	1.000	0.451	0.997
	13	3	1.000	0.234	0.996	1.000	0.235	0.996
	13	4	1.000	0.297	0.998	1.000	0.299	0.998
	13	(3.8, 6)	1.000	0.386	0.999	1.000	0.389	0.999
10	0		0.052			0.046		
	2	3	0.047	0.000	0.000	0.070	0.173	0.338
	2	4	0.042	0.000	0.000	0.104	0.304	0.581
	2	(3.8, 6)	0.040	0.000	0.000	0.159	0.388	0.732
	6	3	0.002	0.069	0.417	0.952	0.192	0.983
	6	4	0.085	0.166	0.993	1.000	0.241	0.988
	5	(3.8, 6)	0.901	0.200	1.000	1.000	0.388	0.992
	7	3	1.000	0.286	1.000	1.000	0.398	0.998
	7	4	1.000	0.286	1.000	1.000	0.434	0.999
	8	(3.8, 6)	1.000	0.276	1.000	1.000	0.389	0.999
	13	3	0.996	0.153	1.000	1.000	0.250	0.998
	13	4	1.000	0.154	1.000	1.000	0.318	0.999
	13	(3.8, 6)	1.000	0.164	1.000	1.000	0.404	1.000
30	0		0.051			0.049		
	2	3	0.044	0.000	0.000	0.071	0.205	0.390
	2	4	0.037	0.000	0.000	0.115	0.338	0.629
	2	(3.8, 6)	0.035	0.000	0.000	0.180	0.404	0.742
	6	3	0.002	0.000	0.000	0.982	0.242	0.986
	6	4	0.000	0.000	0.000	1.000	0.339	0.991
	5	(3.8, 6)	0.174	0.200	1.000	1.000	0.452	0.993
	8	3	0.830	0.232	1.000	1.000	0.359	0.998
	8	4	1.000	0.250	1.000	1.000	0.388	0.999
	8	(3.8, 6)	1.000	0.250	1.000	1.000	0.421	0.999
	14	3	0.733	0.128	1.000	1.000	0.258	0.998
	14	4	1.000	0.143	1.000	1.000	0.326	0.999
	14	(3.8, 6)	1.000	0.143	1.000	1.000	0.407	1.000
50	0		0.052			0.051		
	2	3	0.040	0.000	0.000	0.103	0.298	0.543
	2	4	0.033	0.000	0.000	0.184	0.423	0.761
	2	(3.8, 6)	0.023	0.000	0.000	0.313	0.519	0.875
	6	3	0.001	0.000	0.000	0.998	0.349	0.990
	6	4	0.000	0.167	1.000	1.000	0.397	0.993
	6	(3.8, 6)	0.258	0.167	1.000	1.000	0.404	0.995
	8	3	0.969	0.247	1.000	1.000	0.409	0.998
	8	4	1.000	0.250	1.000	1.000	0.464	0.999
	8	(3.8, 6)	1.000	0.250	1.000	1.000	0.529	1.000
	14	3	0.908	0.138	1.000	1.000	0.287	1.000
	14	4	1.000	0.143	1.000	1.000	0.362	1.000
	14	(3.8, 6)	1.000	0.143	1.000	1.000	0.466	1.000

Continued

True zero	True signal	RR	TreeScan-Poisson			TreeScan-ZIP		
			Power*	Sensitivity	PPV	Power*	Sensitivity	PPV
70	0		0.050			0.049		
	2	3	0.040	0.000	0.000	0.226	0.499	0.826
	2	4	0.035	0.000	0.000	0.462	0.586	0.926
	2	(3.8, 6)	0.033	0.000	0.000	0.735	0.664	0.967
	6	3	0.000	–	–	1.000	0.417	0.996
	6	4	0.013	0.167	1.000	1.000	0.499	0.997
	6	(3.8, 6)	0.969	0.167	1.000	1.000	0.527	0.999
	7	3	1.000	0.286	1.000	1.000	0.429	0.999
	7	4	1.000	0.286	1.000	1.000	0.477	0.999
	7	(3.8, 6)	1.000	0.286	1.000	1.000	0.586	0.999
	13	3	1.000	0.154	1.000	1.000	0.249	0.999
	13	4	1.000	0.154	1.000	1.000	0.294	0.999
	13	(3.8, 6)	1.000	0.143	1.000	1.000	0.466	1.000

Table 2. Type I error, power, sensitivity and positive predictive value obtained by the two methods according to the number of true signals and relative risk. * Type I error when the number of true signals is 0.

graphic. The ICSRs are periodically summited to the WHO-UMC. Further, safety information obtained from KAERS data and signal analysis is periodically reported to the Ministry of Food and Drug Safety.

For the real data analysis, data cleansing was performed. Because a certain drug and AE information can be reported multiple times depending on the dose and time of administration, if the same drug and AE were reported twice or more, only the first report was used. In the causality, only drug–AE pairs that received ratings of possible or above were included in this study. There are 6 levels of causality: certain, probable, possible, unlikely, conditional, and unassessable[39,40]. In KAERS database, AEs are coded by the WHO-ART. As more than half of the reports included information down to the PT level, and HLT may not exist, this study used two levels of hierarchy, SOC and PT, with the exception of the HLT and IT level.

Data obtained between 2012 and 2016 from KAERS were used. During this period, 716,584 people reported experiencing AEs. There were 1.8 million drug reports on 1981 types of drugs and 1.1 million AE reports on 4078 types of AEs. Further, a total of 2.4 million unique drug-AE pairs were found. When removing pairs that had beneath the 'possible' threshold, the final dataset analyzed in this study included 1,077,060 drug-AE pairs representing 1292 types of AEs in PTs. Further, 1981 types of drugs were identified in 557,390 reports.

Paclitaxel and docetaxel.

The two proposed methods were applied to detect the AE signals to the drug–AE pairs data from KAERS. Paclitaxel and docetaxel, which have the highest sales among all anticancer drugs in the world, were selected[41]. Of note, these are representatives of the new class of taxane drugs, which have emerged as a fundamental treatment for breast cancer. Paclitaxel and docetaxel have similar main structures and mechanisms of action[42]. Paclitaxel is used to treat a number of cancer types, including Kaposi sarcoma, breast cancer, ovarian cancer, lung cancer, cervical cancer, and pancreatic cancer (https://www.ashp.org/). Docetaxel is also used as to treat several cancer types, including breast cancer, non-small cell lung cancer, prostate cancer, head and neck cancer, and stomach cancer (https://www.cancer.gov/). The most frequently reported AEs related to taxene from MICROMEDEX® include cardiovascular effects, dermatologic effects, endocrine/metabolic effects gastrointestinal effects, hematologic effects, hepatic effects, immunologic effects, musculoskeletal effects, neurologic effects, ophthalmic effects, otic effects, renal effects, respiratory effects, and others (https://www.who.int/).

Results.

Paclitaxel. Nine signals were identified by the TreeScan-Poisson method and 30 signals were detected by the TreeScan-ZIP method (Table 3). The nine signals detected by the TreeScan-Poisson method were also detected by the TreeScan-ZIP method. The AEs corresponding to the signals found by both methods were related to the following SOCs: central & peripheral nervous system disorders (0410), respiratory system disorders (1100), white cell and reticuloendothelial system disorders (1220), and body as a whole—general disorders (1810). Further, their PTs were paresthesia (0410.0137), neuropathy peripheral (0410.1313), dyspnea (1100.0514), granulocytopenia (1220.0572), leucopenia (1220.0908), chest pain (1810.0718), and temperature change sensations (1810.1705). The TreeScan-ZIP method detected signals related to 10 SOC terms. The nine signals detected by the two methods were included in the known AEs. However, some signals detected by TreeScan-ZIP alone were included in the known AEs.

Docetaxel. The TreeScan-Poisson and the TreeScan-ZIP methods identified 9 and 56 signals, respectively (Table 4). All signals detected by the TreeScan-Poisson method were also detected by the TreeScan-ZIP method. The AEs corresponding to the signals found by both methods were related to the following SOCs: skin and appendages disorders (0100), musculo-skeletal system disorders (0200), central & peripheral nervous system disorders (0410), red blood cell disorders (1210), white cell and reticulo-endothelial system (RES) disorders

					TreeScan-Poisson			TreeScan-ZIP		
SOC	PT	Diagnosis	Marginal total	Obs	Exp	O/E	p-value	Exp	O/E	p-value
0100		Skin and appendages disorders	352,949	1020	1176.5	0.9	1.000	868.3	1.2	1.000
	0002	ALOPECIA	13,193	94	44.0	2.1	0.929	32.5	2.9	0.001
	0043	SWEATING INCREASED	13,433	100	44.8	2.2	0.890	33.0	3.0	0.001
	0828	HYPOTRICHOSIS	234	13	0.8	16.7	0.897	0.6	22.6	0.001
0200		Musculo-skeletal system disorders	50,667	305	168.9	1.8	0.639	124.6	2.4	0.001
	0073	MYALGIA	24,679	244	82.3	3.0	0.119	60.7	4.0	0.001
0410		Central & peripheral nervous system disorders	233,135	781	777.1	1.0	1.000	573.5	1.4	1.000
	0117	HYPOAESTHESIA	2349	32	7.8	4.1	0.934	5.8	5.5	0.001
	0130	NEUROPATHY	4198	81	14.0	5.8	0.318	10.3	7.8	0.001
	0137	PARAESTHESIA	12,165	213	40.5	5.3	0.015	29.9	7.1	0.001
	1313	NEUROPATHY PERIPHERAL	5634	150	18.8	8.0	0.015	13.9	10.8	0.001
	2082	POLYNEUROPATHY	183	7	0.6	11.5	0.998	0.5	15.5	0.001
0800		Metabolic and nutritional disorders	68,669	92	228.9	0.4	1.000	168.9	0.5	1.000
	0368	CACHEXIA	3184	49	10.6	4.6	0.754	7.8	6.3	0.001
1030		Heart rate and rhythm disorders	22,346	206	74.5	2.8	0.270	55.0	3.7	0.001
	0221	PALPITATION	11,778	99	39.3	2.5	0.810	29.0	3.4	0.001
	0224	TACHYCARDIA	4569	94	15.2	6.2	0.177	11.2	8.4	0.001
1040		Vascular (extracardiac) disorders	13,671	100	45.6	2.2	0.897	33.6	3.0	0.001
	0207	FLUSHING	5334	91	17.8	5.1	0.312	13.1	6.9	0.001
1100		Respiratory system disorders	137,936	588	459.8	1.3	0.961	339.3	1.7	0.001
	0514	DYSPNOEA	36,735	410	122.4	3.3	0.010	90.4	4.5	0.001
	0537	RESPIRATORY INSUFFICIENCY	1292	18	4.3	4.2	0.996	3.2	5.7	0.001
1220		White cell and RES disorders	92,531	1085	308.4	3.5	0.001	227.6	4.8	0.001
	0570	AGRANULOCYTOSIS	8089	89	27.0	3.3	0.656	19.9	4.5	0.001
	0572	GRANULOCYTOPENIA	58,735	674	195.8	3.4	0.001	144.5	4.7	0.001
	0908	LEUCOPENIA	20,456	314	68.2	4.6	0.008	50.3	6.2	0.001
1700		Neoplasms	6336	25	21.1	1.2	1.000	15.6	1.6	0.996
	1345	NEOPLASM MALIGNANT	591	13	2.0	6.6	0.989	1.5	8.9	0.001
1810		Body as a whole—general disorders	220,690	1295	735.6	1.8	0.011	542.9	2.4	0.001
	0712	ALLERGIC REACTION	1394	18	4.6	3.9	0.998	3.4	5.2	0.001
	0718	CHEST PAIN	25,856	479	86.2	5.6	0.001	63.6	7.5	0.001
	0730	PAIN	7221	54	24.1	2.2	0.987	17.8	3.0	0.001
	1705	TEMPERATURE CHANGED SENSATION	9204	245	30.7	8.0	0.003	22.6	10.8	0.001
	2237	ANAPHYLACTIC REACTION	3680	44	12.3	3.6	0.895	9.1	4.9	0.001

Table 3. Results of signal detection of adverse events of paclitaxel by the two methods.

(1220). Their PTs were alopecia (0100.0002), nail disorder (0100.0020), myalgia (0200.0073), sensory disturbance (0410.0148), anemia (1210.0544), and granulocytopenia (1220.0572). The TreeScan-ZIP method detected signals related to 18 SOC terms. All signals detected by the two methods were included in the known AEs. A few signals that were not detected by TreeScan, but were detected by TreeScan-ZIP, were included in known AEs, such as vision disorders, gastro-intestinal system disorders, liver and biliary system disorders, urinary system disorders, etc.

Conclusion and discussion

This study sought to reveal how the tree-based scan statistic developed by Kulldorff et al.[26] can be extended for the zero-inflated count data. To consider a large number of zero cells, we proposed the TreeScan-ZIP method, which integrates a zero-inflated Poisson model into the TreeScan-Poisson method. Herein, a simulation study was conducted with different settings for the relative risk and the number of true zero leaves and true signal leaves. Based on the findings of the simulation study, the TreeScan-ZIP method performed better than the TreeScan-Poisson method in terms of power, sensitivity, and PPV, especially when the proportion of true zeros was high. The real data examples also supported the simulation results. The TreeScan-Poisson method may have missed many signals that were detected by the TreeScan-ZIP method in datasets with a large number of true zeros. If the TreeScan-ZIP method detects too many false positive signals, it may increase confusion in further investigation and utilize unnecessary energy. However, even the known AEs were not detected by the TreeScan-Poisson method. Although we do not know whether all signals detected by the TreeScan-ZIP method were true, it is safer to over-detect than to miss any signal in drug safety surveillance.

					TreeScan-Poisson			TreeScan-ZIP		
SOC	**PT**	**Diagnosis**	**Marginal total**	**Obs**	**Exp**	**O/E**	***p*-value**	**Exp**	**O/E**	***p*-value**
0100		Skin and appendages disorders	352,949	4420	5126.4	0.9	1.000	3467.1	1.3	1.000
	0002	ALOPECIA	13,193	2212	191.6	11.5	0.001	129.6	17.1	0.001
	0008	DERMATITIS EXFOLIATIVE	161	11	2.3	4.7	1.000	1.6	7.0	0.024
	0020	NAIL DISORDER	3248	760	47.2	16.1	0.005	31.9	23.8	0.001
	1199	SKIN EXFOLIATION	1673	54	24.3	2.2	1.000	16.4	3.3	0.001
	1634	NAIL DISCOLOURATION	429	91	6.2	14.6	0.927	4.2	21.6	0.001
0200		Musculo-skeletal system disorders	50,667	2387	735.9	3.2	0.011	497.7	4.8	0.001
	0063	ARTHRALGIA	8201	416	119.1	3.5	0.813	80.6	5.2	0.001
	0073	MYALGIA	24,679	1908	358.4	5.3	0.003	242.4	7.9	0.001
0410		Central & peripheral nervous system disorders	233,135	2081	3386.1	0.6	1.000	2290.1	0.9	1.000
	0148	SENSORY DISTURBANCE	2473	767	35.9	21.4	0.003	24.3	31.6	0.001
	1313	NEUROPATHY PERIPHERAL	5634	230	81.8	2.8	0.986	55.3	4.2	0.001
	1532	LOWER MOTOR NEURONE LESION	117	12	1.7	7.1	1.000	1.1	10.4	0.001
0431		Vision disorders	17,634	186	256.1	0.7	1.000	173.2	1.1	1.000
	1049	LACRIMATION ABNORMAL	647	116	9.4	12.3	0.879	6.4	18.3	0.001
	1462	EPIPHORA	151	22	2.2	10.0	1.000	1.5	14.8	0.001
0433		Special senses other, disorders	4692	563	68.1	8.3	0.121	46.1	12.2	0.001
	0267	TASTE PERVERSION	4195	555	60.9	9.1	0.103	41.2	13.5	0.001
0500		Psychiatric disorders	129,819	1261	1885.5	0.7	1.000	1275.2	1.0	1.000
	0165	ANOREXIA	36,109	690	524.5	1.3	1.000	354.7	1.9	0.001
0600		Gastro-intestinal system disorders	636,320	6813	9242.1	0.7	1.000	6250.7	1.1	1.000
	0204	CONSTIPATION	45,356	991	658.8	1.5	0.988	445.5	2.2	0.001
	0269	ANUS DISORDER	321	17	4.7	3.6	1.000	3.2	5.4	0.005
	0298	HAEMORRHOIDS	1442	45	20.9	2.1	1.000	14.2	3.2	0.005
	0321	PROCTITIS	91	24	1.3	18.2	0.999	0.9	26.8	0.001
	0327	STOMATITIS	10,870	256	157.9	1.6	1.000	106.8	2.4	0.001
	1014	HAEMORRHAGE RECTUM	655	33	9.5	3.5	1.000	6.4	5.1	0.001
	1083	GINGIVITIS	1353	133	19.7	6.8	0.949	13.3	10.0	0.001
	1351	MUCOSITIS NOS	4978	170	72.3	2.4	0.999	48.9	3.5	0.001
	1376	TOOTH ACHE	1032	42	15.0	2.8	1.000	10.1	4.1	0.001
0700		Liver and biliary system disorders	52,619	643	764.3	0.8	1.000	516.9	1.2	1.000
	0360	SGPT INCREASED	12,811	341	186.1	1.8	0.998	125.8	2.7	0.001
0800		Metabolic and nutritional disorders	68,669	714	997.4	0.7	1.000	674.6	1.1	1.000
	0381	HYPERCHOLESTEROLAEMIA	1982	139	28.8	4.8	0.978	19.5	7.1	0.001
	0387	HYPOCALCAEMIA	2433	112	35.3	3.2	0.998	23.9	4.7	0.001
1040		Vascular (extracardiac) disorders	13,671	255	198.6	1.3	1.000	134.3	1.9	0.001
	0207	FLUSHING	5334	216	77.5	2.8	0.986	52.4	4.1	0.001
	1413	ERYTHROMELALGIA	81	27	1.2	22.9	0.998	0.8	33.9	0.001
1100		Respiratory system disorders	137,936	1644	2003.4	0.8	1.000	1355.0	1.2	1.000
	0523	PHARYNGITIS	18,340	361	266.4	1.4	1.000	180.2	2.0	0.003
1210		Red blood cell disorders	30,116	1675	437.4	3.8	0.030	295.8	5.7	0.001
	0544	ANAEMIA	25,889	1668	376.0	4.4	0.011	254.3	6.6	0.001
1220		White cell and RES disorders	92,531	3969	1344.0	3.0	0.002	909.0	4.4	0.001
	0570	AGRANULOCYTOSIS	8089	375	117.5	3.2	0.899	79.5	4.7	0.001
	0572	GRANULOCYTOPENIA	58,735	2474	853.1	2.9	0.028	577.0	4.3	0.001
	0908	LEUCOPENIA	20,456	1091	297.1	3.7	0.167	200.9	5.4	0.001
1300		Urinary system disorders	49,509	301	719.1	0.4	1.000	486.3	0.6	1.000
	0621	RENAL PAIN	466	40	6.8	5.9	1.000	4.6	8.7	0.001
1420		Reproductive disorders, female	10,695	283	155.3	1.8	1.000	105.1	2.7	0.001
	0636	AMENORRHOEA	800	151	11.6	13.0	0.761	7.9	19.2	0.001
	0669	VAGINITIS	463	23	6.7	3.4	1.000	4.5	5.1	0.001
	1839	BREAST PAIN	500	38	7.3	5.2	1.000	4.9	7.7	0.001

Continued

SOC	PT	Diagnosis	Marginal total	Obs	TreeScan-Poisson			TreeScan-ZIP		
					Exp	O/E	*p*-value	Exp	O/E	*p*-value
1810		Body as a whole—general disorders	220,690	4995	3205.4	1.6	0.334	2167.9	2.3	0.001
	0401	OEDEMA PERIPHERAL	9444	607	137.2	4.4	0.398	92.8	6.5	0.001
	0716	ASTHENIA	24,301	456	353.0	1.3	1.000	238.7	1.9	0.006
	0717	BACK PAIN	9781	209	142.1	1.5	1.000	96.1	2.2	0.003
	0718	CHEST PAIN	25,856	928	375.5	2.5	0.681	254.0	3.7	0.001
	0724	FATIGUE	14,561	515	211.5	2.4	0.933	143.0	3.6	0.001
	1705	TEMPERATURE CHANGED SENSATION	9204	822	133.7	6.1	0.079	90.4	9.1	0.001
	1765	PALMAR-PLANTAR ERYTHRODYSAESTHESIA	7415	517	107.7	4.8	0.441	72.8	7.1	0.001
	2101	PAIN AXILLARY	159	20	2.3	8.7	1.000	1.6	12.8	0.001
1820		Application site disorders	25,336	150	368.0	0.4	1.000	248.9	0.6	1.000
	0058	INJECTION SITE REACTION	3385	106	49.2	2.2	1.000	33.3	3.2	0.001
2000		Secondary terms—events	12,322	92	179.0	0.5	1.000	121.0	0.8	0.001
	1813	SURGICAL SITE REACTION	290	68	4.2	16.1	0.964	2.8	23.9	0.001

Table 4. Results of signal detection of adverse events of docetaxel by the two methods.

The data used were extracted from spontaneous reporting systems, which is a limitation. As spontaneous reporting systems are based on self-reporting by people, such as consumers and healthcare professionals, underreporting or overreporting of AEs may easily occur. For example, only the number of cases reported can be known. Thus, whether the same AE occurred multiple times in the same person cannot be known. Cases of overreporting may thus lead to bias in the analysis.

In this study, the TreeScan-ZIP method and TreeScan-Poisson method identified signals of AEs for a particular drug, and could identify drugs that are more frequently reported to be related to a particular AE. Cuts were made either above or below nodes in this study; however, more elaborate cuts, such as the combinational cuts proposed by Kulldorff et al.[7] can also be made. In this study, we used a two-level structure; however, structures with more than two levels or other spontaneous reporting system data with more delicate levels can be employed. Further studies could use a zero-inflated double Poisson or zero-inflated negative binomial model to accommodate large numbers of true zeros and overdispersion[43]. When a priory level of AE definition cannot be determined in the tree structure and the data have a large number of zeros, the proposed tree-based scan statistic can serve as a very useful method for detecting signals in the post-market drug safety surveillance.

Data availability

The KARES database is provided via the Korea Institute of Drug Safety and Risk management webpage. (https://open.drugsafe.or.kr/original/invitation.jsp) upon request.

Received: 28 April 2022; Accepted: 7 September 2022
Published online: 29 September 2022

References

1. Baciu, A., Stratton, K., Burke, S.P. Committee on the Assessment of the US Drug Safety System. The Future of Drug Safety: Promoting and Protecting the Health of the Public (2006).
2. Platt, R. *et al.* The new Sentinel network—improving the evidence of medical-product safety. *N. Engl. J. Med.* **361**(7), 645–647 (2009).
3. Avorn, J. & Schneeweiss, S. Managing drug-risk information—what to do with all those new numbers. *N. Engl. J. Med.* **361**(7), 647–649 (2009).
4. Davis, R. L. *et al.* Active surveillance of vaccine safety: A system to detect early signs of adverse events. *Epidemiology* **16**(3), 336–341 (2005).
5. Platt, R. *et al.* Multicenter epidemiologic and health services research on therapeutics in the HMO research network center for education and research on Therapeutics. *Pharmacoepidemiol. Drug Saf.* **10**(5), 373–377 (2001).
6. Yih, W. K. *et al.* Active surveillance for adverse events: The experience of the vaccine safety Datalink project. *Pediatrics* **127**(Supplement 1), S54–S64 (2011).
7. Kulldorff, M. *et al.* Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiol. Drug Saf.* **22**(5), 517–523. https://doi.org/10.1002/pds.3423 (2013).
8. Singleton, J. A. *et al.* An overview of the vaccine adverse event reporting system (VAERS) as a surveillance system. *Vaccine.* **17**(22), 2908–2917 (1999).
9. Lindquist, M. VigiBase, the WHO global ICSR database system: Basic facts. *Drug Inf. J.* **42**(5), 409–419 (2008).
10. Evans, S. J., Waller, P. C. & Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol. Drug Saf.* **10**(6), 483–486. https://doi.org/10.1002/pds.677 (2001).
11. Rothman, K. J., Lanes, S. & Sacks, S. T. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol. Drug Saf.* **13**(8), 519–523. https://doi.org/10.1002/pds.1001 (2004).
12. Udny, Y., Kendall, M.G., et al. An introduction to the theory of statistics. An introduction to the theory of statistics (14th ed). (1950).
13. Greenwood, P. E. & Nikulin, M. S. *A guide to chi-squared testing* (Wiley, New York, 1996).
14. Huang, L., Zalkikar, J. & Tiwari, R. C. A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *J. Am. Stat. Assoc.* **106**(496), 1230–1241 (2011).

15. Bate, A. *et al.* A Bayesian neural network method for adverse drug reaction signal generation. *Eur. J. Clin. Pharmacol.* **54**(4), 315–321. https://doi.org/10.1007/s002280050466 (1998).
16. DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am. Stat.* **53**(3), 177–190. https://doi.org/10.1080/00031305.1999.10474456 (1999).
17. Huang, L., Zalkikar, J. & Tiwari, R. C. Likelihood ratio test-based method for signal detection in drug classes using FDA's AERS database. *J. Biopharm. Stat.* **23**(1), 178–200. https://doi.org/10.1080/10543406.2013.736810 (2013).
18. Noren, G. N., Bate, A., Orre, R. & Edwards, I. R. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat. Med.* **25**(21), 3740–3757. https://doi.org/10.1002/sim.2473 (2006).
19. DuMouchel, W., Pregibon, D. Empirical bayes screening for multi-item associations. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining; San Francisco, California: Association for Computing Machinery 67–76 (2001).
20. Cesana, M. *et al.* Bayesian data mining techniques: The evidence provided by signals detected in single-company spontaneous reports databases. *Drug Inf. J.* **41**(1), 11–21 (2007).
21. O'Neill, R. T. & Szarfman, A. Some US food and drug administration perspectives on data mining for pediatric safety assessment. *Curr. Ther. Res.* **62**(9), 650–663 (2001).
22. Szarfman, A., Machado, S. G. & O'Neill, R. T. 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. *Drug Saf.* **25**(6), 381–392 (2002).
23. Poluzzi, E., Raschi, E., Moretti, U. & De Ponti, F. Drug-induced torsades de pointes: data mining of the public version of the FDA Adverse Event Reporting System (AERS). *Pharmacoepidemiol. Drug Saf.* **18**(6), 512–518 (2009).
24. Hu, N., Huang, L. & Tiwari, R. C. Signal detection in FDA AERS database using Dirichlet process. *Stat. Med.* **34**(19), 2725–2742. https://doi.org/10.1002/sim.6510 (2015).
25. Wilson, A. M., Thabane, L. & Holbrook, A. Application of data mining techniques in pharmacovigilance. *Br. J. Clin. Pharmacol.* **57**(2), 127–134 (2004).
26. Kulldorff, M., Fang, Z. & Walsh, S. J. A tree-based scan statistic for database disease surveillance. *Biometrics* **59**(2), 323–331 (2003).
27. Lee, H., Kim, J. H., Choe, Y. J. & Shin, J.-Y. Safety surveillance of pneumococcal vaccine using three algorithms: Disproportionality methods, empirical bayes geometric mean, and tree-based scan statistic. *Vaccines.* **8**(2), 242 (2020).
28. Kim, J. H., Lee, H. & Shin, J.-Y. Bacillus Calmette-Guérin (BCG) vaccine safety surveillance in the Korea adverse event reporting system using the tree-based scan statistic and conventional disproportionality-based algorithms. *Vaccine.* **38**(21), 3702–3710 (2020).
29. Kim, S. *et al.* Data-mining for detecting signals of adverse drug reactions of fluoxetine using the Korea adverse event reporting system (KAERS) database. *Psychiatr. Res.* **256**, 237–42. https://doi.org/10.1016/j.psychres.2017.06.038 (2017).
30. Wang, S. V. *et al.* Data mining for adverse drug events with a propensity score-matched tree-based scan statistic. *Epidemiology* **29**(6), 895–903. https://doi.org/10.1097/ede.0000000000000907 (2018).
31. Huang, L., Zalkikar, J. & Tiwari, R. Likelihood ratio based tests for longitudinal drug safety data. *Stat. Med.* **33**(14), 2408–2424 (2014).
32. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**(1), 1–14 (1992).
33. Huang, L., Guo, T., Zalkikar, J. N. & Tiwari, R. C. A review of statistical methods for safety surveillance. *Ther. Innov. Regul. Sci.* **48**(1), 98–108. https://doi.org/10.1177/2168479013514236 (2014).
34. Loader, C. R. Large-deviation approximations to the distribution of scan statistics. *Adv. Appl. Probab.* **23**(4), 751–771 (1991).
35. Kulldorff, M. A spatial scan statistic. *Commun. Stat.-Theory Methods.* **26**(6), 1481–1496 (1997).
36. Dwass, M. Modified randomization tests for nonparametric hypotheses. *Ann. Math. Stat.* 181–7 (1957).
37. Cançado, A. L., Da-Silva, C. Q. & Da Silva, M. F. A spatial scan statistic for zero-inflated Poisson process. *Environ. Ecol. Stat.* **21**(4), 627–50 (2014).
38. Jackman, S. Package 'pscl'. Technical Report. (Stanford, CA: Political Science Computational Laboratory, Stanford University, 2006).
39. Edwards, I. R. & Biriell, C. Harmonisation in pharmacovigilance. *Drug Saf.* **10**(2), 93–102 (1994).
40. Edwards, I. R. & Aronson, J. K. Adverse drug reactions: Definitions, diagnosis, and management. *The Lancet.* **356**(9237), 1255–1259 (2000).
41. Wang, L., Du, G.-H. Paclitaxel. Natural Small Molecule Drugs from Plants. 537–43 (2018).
42. Verweij, J., Clavel, M. & Chevalier, B. Paclitaxel (Taxol) and docetaxel (Taxotere): Not simply two of a kind. *Ann Oncol.* **5**(6), 495–505. https://doi.org/10.1093/oxfordjournals.annonc.a058903 (1994).
43. de Lima, M.S., Duczmal, L.H., Neto, J.C., Pinto, L.P. Spatial scan statistics for models with overdispersion and inflated zeros. *Stat. Sinica.* 225–41 (2015).

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by G.P. The first draft of the manuscript was written by G.P. and I.J. commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1F1A1057182).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to I.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.