Pharmaceutical Medicine (2024) 38:321–329
https://doi.org/10.1007/s40290-024-00530-1

ORIGINAL RESEARCH ARTICLE

Statistical Signal Detection Algorithm in Safety Data: A Proprietary Method Compared to Industry Standard Methods

Eugenia Bastos[1] · Jeff K. Allen[2] · Jeff Philip[2]

Accepted: 18 June 2024 / Published online: 13 July 2024

Abstract

Introduction Several quantitative methods have been established, in pharmacovigilance, to detect signals of disproportionate reporting (SDRs) from databases containing reports of adverse drug reactions (ADRs). The signal detection algorithms (SDAs) and the source of the reporting per product vary, but it is unclear whether any algorithm can provide satisfactory performance using data with such large variance factors.

Objective Determine the appropriate SDA for Biogen's internal Global Safety Database (GSD) given the characteristics of the database including frequencies of events, data skewness, outliers, and missing information. Compare performance of standard approaches (EBGM, EB05, PRR, and ROR), well accepted by industry, to a Biogen-developed Machine Learning (ML) Regression Decision Tree (RDT) model, across several Biogen products, to determine a champion SDA.

Methods All data associated with seven marketed Biogen products were chosen and a historical subset of reported ADRs were considered. Six SDAs (five common industry disproportionality methods) and RDT were evaluated. The SDRs were calculated on training and test data composed of quarterly reporting intervals from 2004–2019. The performance measures used were sensitivity, precision, time to detect new events, and frequency of detected cases for each algorithm for each product. Outcomes in the test data are known a priori and easily compared to predicted outcomes. Validation was performed via rates of misclassification. This work solely represents Biogen's internal information, intentionally chosen to serve the performance review of its signal detection systems, and results will not necessarily be generalizable to other external sources.

Results Several algorithms performed differently among products, but no one method dominated any other. Performance was dependent on the thresholds used to define a signal according to different criteria. However, those different statistics subtly influenced the achievable performance. The relative performance of RDT and Medicines and Healthcare products Regulatory Agency (MHRA) algorithms were superior and paired across products. A reduction in precision for all methods spanning the products was present. Hence, companies evaluating signal detection approaches, search for innovative methods to minimize this effect.

Conclusions In designing signal detection systems, careful consideration should be given to the criteria that are used to define SDRs. The choice of disproportionality statistics does not affect the achievable range of signal detection performance. These choices should consider mainly ease of implementation and interpretation. The implementation of a method is specific to its accuracy. The RDT attempted to take advantage of known methods and compare results on a per-product basis. Many factors influencing ADRs may improve RDT in future efforts. In this experiment, RDT demonstrated superiority in terms of quickest time to detect and capturing of the highest number of ADRs. Next steps include expansion of data for products representing other indications and testing models in external databases to investigate generalizability of estimates when comparing SDAs.

Extended author information available on the last page of the article

Key Points

Machine learning (ML) algorithms demonstrate viability in detecting safety signals.

The RDT results, a bespoke Decision Tree analysis, detected records of 2 × 2 tables representing cases of highly disproportionality measures (outliers) while displaying high sensitivity and precision.

The RDT has the shortest times to detect (TTD) signals across products.

1 Introduction

The management and analysis of individual case safety reports (ICSRs) remains vital to detecting safety signals for marketed pharmaceutical products. Dependable ICSR-driven signal detection is a staged process combining both quantitative disproportionality analysis and the expertise of pharmacovigilance (PV) professionals providing ascertainment. With the advent of high-performance computing (HPC) environments and cloud-based platforms, novel mathematical approaches such as machine learning (ML) and artificial intelligence (AI), which are innately highly iterative and require an ever-expanding pool of source data to be effective, have become an industry passion. As these systems and methods are also well embedded in industry, it has become incumbent upon safety departments to explore the capacity of these systems/methods to predict/detect safety signals and subsequently perform time-to-event analysis.

Most pharmaceutical companies depend on safety data relevant to their own products drawn from a variety of sources: commercially available safety databases, direct customer and provider reports, regulatory authority databases, etc. These databases form the bedrock for disproportionality analysis. While clinical assessment is the most crucial and requisite stage of signal detection, disproportionality analysis is employed to generate hypotheses on potential relationships between pharmaceutical products and adverse effects; consistent with 2×2 table analysis, it is based on the contrast between observed and expected numbers of reports for any given combination of drug and adverse event [1].

2 Objectives

This paper focuses on the use of standard disproportionality measures in tandem with a proprietary ML approach, RDT (a bespoke decision tree analysis), to investigate the strengths and weaknesses of current standard algorithms when compared to and supplemented by an ML approach. The main measures we will use to draw our comparisons will be sensitivity, specificity, and accuracy.

3 Data and Methods

3.1 Data

All available data associated with seven marketed Biogen products were chosen in the form of a large historical and 'frozen' subset of all reported ADRs drawn from Biogen's Global Safety Database (GSD) database were considered. The data were then summarized at the MedDRA (Medical Dictionary for Regulatory Activities) Preferred Term (PT) and product level in discrete quarterly intervals of time (2014–2019). Counts were established in the fashion of a four-cell (A, B, C, D) two-by-two table.

Once the counts were established, from them, a series of ratios were then calculated. The resulting measures of disproportionalities used in this paper were accuracy, sensitivity, specificity, etc. The full data set, a total of 1193 table records, were split into a training set (80 %; $n = 954$ observations) and test set (20 %; $n = 239$ observations).

3.2 Signal Detection Algorithms

Six SDAs (five ratio-based commonly used disproportionality methods within from the industry based on disproportionality methods) inclusive of RDT, a proprietary method employing ML decision tree approach, were evaluated. The five ratio-based disproportionality methods include:

1. Medicines and Healthcare products Regulatory Agency (MHRA) Empirical Bayesian Geometric Mean (EBGM, EB05 [6]).
2. European Medicines Agency (EMA)-adapted Proportional Reporting Ratio (PRR).
3. BAYER-bespoke Proportional Reporting Ratio (PRR [5]).
4. ROCHE-bespoke Proportional Reporting Ratio (PRR).
5. Reporting Odds Ratio (ROR [5]).

Table 1 provides additional details on the calculation of each aforementioned method. The ratios (PRR, ROR) are

Table 1 Signal detection algorithm (SDA) method, identifier, and full value requirements

SDA	Identifier	Model (product average)
EBGM [3]	MHRA	EB05≥1.8, n ≥ 3, EBGM ≥ 2.5
PRR [3]	EMA	PRR LB 95 % CI ≥1, n ≥3
	BAYER	PRR ≥2; χ^2 ≥4, n ≥3
	ROCHE	PRR≥2, $p(\chi^2)$ ≤0.05, n≥3
ROR [3]	ROR	ROR LB 95 % CI≥1, n≥3

χ2 chi-squared, *CI* confidence interval, *EB05* lower limit of the 90 % one-sided CI of the EBGM, *EBGM* Empirical Bayesian Geometric Mean, *EMA* European Medicine Agency, *LB* lower limit of the 95 % two-sided CI, *MHRA* Medicines and Healthcare products Regulatory Agency; *n* number of adverse drug reports, *PRR* proportional reporting ratio, *ROR* reporting odds ratio

described in Table 2 as a function of a 2 × 2 contingency table. In Table 3, the letters A–D correspond to cells in Table 2.

Signals of disproportionate reporting (SDRs) were calculated on data composed of quarterly reporting intervals during 2004 to 2019. The performance measures used were cumulative sensitivity, cumulative precision (positive predictive value), time to detect new events, and frequency of detected cases for each algorithm for each product. Actual outcomes in the test data are known a priori and thus easily compared to predicted outcomes. Validation was performed using misclassification rates to determine measure accuracy.

The proprietary method, RDT, a newly proposed and tested classification method, Regression Decision Tree (RDT [sensitivity (s), precision (p)]), using a systematic approach of path optimization, was employed to predict cumulative sensitivity and cumulative precision utilizing independent disproportionality rates, leading to SDRs frequencies. Because rates are defined within the interval [0–1], regression was our algorithm of choice and RDT (s, p) was created. We combined a two-step model utilizing both outcomes: cumulative sensitivity and cumulative precision.

Step 1: Pred-Y1(Cumulative Sensitivity(a/a+c)) = X1(MHRA-EBGM) + X2(MHRA-EB005) + X3(EMA-PRR) + X4(BAYER-PRR) + X5(ROCHE-PRR) + X6(ROR) + error.

Step 2: Pred-Y2(Cumulative Precision(a/a+b)) = X1(MHRA-EBGM) + X2(MHRA-EB005) + X3(EMA-PRR) + X4(BAYER-PRR) + X5(ROCHE-PRR) + X6(ROR) + error.

Table 2 The 2×2 contingency table

2 × 2	Drug positive	Drug negative
AE Yes	A	B
AE No	C	D

AE adverse events

Table 3 Formulas for ratio calculations

EBGM as per RRR formula [6]	[A*(A+B+C+D)]/[(A+C)*(A+B)]
PRR	[A/(A+C)]/[B/(B+D)]
ROR	[(A/B)]/[(C/D)]

EBGM Empirical Bayesian Geometric Mean, *PRR* proportional reporting ratio, *ROR* reporting odds ratio

After executing both steps independently, using quarter periods for each Biogen Product, these resulting records were appended, from step 1 and step 2 and together produce the RDT results. We referred to this two-step model as RDT (s, p). The RDT results detected records of 2 × 2 tables representing cases of highly disproportionality measures.

4 Results

4.1 Comparisons of Measures, Median Sensitivity, and Median Precision, Across SDAs

The overall performance of six SDAs is listed in Table 4. The median sensitivity and median precision in this table are both calculated using the average across all seven Biogen products included in this work, summarized on SDA level. We looked for SDAs with the highest sensitivity (denominator contains related and unrelated events under drug) and highest precision (denominator contains all reported events, independent of drug exposure) medians. In this case, no method meets these exact criteria but the adoption of tradeoffs between both measures can maintain significance in the selection of signals. The methods with the highest sensitivities (ROR, EMA, ROCHE and BAYER) (cluster 1) yielded the lowest precisions. And the method with highest precision (MHRA) (cluster 2), yielded the lowest sensitivity. The RDT (s, p) falls in between those two clusters, with sensitivity and precision medians equal to (0.2781; 0.1585). RDT (s, p) sensitivity ranks higher than MHRA's sensitivity (0.2781 × 0.2449), its precision ranks the second highest, MHRA's precision (0.1585 × 0.3177). In summary, the first cluster will be dismissed due to low precision rates. The dispute between MHRA and RDT relies on a preference of either measure, while RDT (s, p) favors sensitivity versus MHRA's precision. A crucial point to account for method efficiency is to consider that both measures are needed at the same rate. Imprecise sensitivities or non-sensible precisions can be devoid of merit.

The next step was to inquire about which sources of variation played influential roles when creating these measures.

Table 4 Median sensitivity and median precision by each SDA

SDA	MODEL (product average)	Median sensitivity	Median precision
MHRA	EB05 ≥1.8, n ≥3, EBGM ≥2.5	0.2449	0.3177
RDT	RDT (s, p)	0.2781	0.1585
BAYER	PRR ≥2, χ^2 ≥4, n ≥3	0.4431	0.1338
ROCHE	PRR≥2, $p(\chi^2)\leq0.05$, n≥3	0.4435	0.1325
EMA	PRR LB 95 % CI≥1, n≥3	0.5245	0.1387
ROR	ROR LB 95 % CI≥1, n≥3	0.5245	0.1387

χ^2 chi-squared, *CI* confidence interval, *EB05* lower limit of the 90 % one-sided CI of the EBGM, *EBGM* Empirical Bayesian Geometric Mean, *EMA* European Medicine Agency, *LB* lower limit of the 95 % two-sided CI, *MHRA* Medicines and Healthcare products Regulatory Agency, *n* number of adverse drug reports, *PRR* proportional reporting ratio, *RDT* Regression Decision Tree, *ROR* reporting odds ratio, *SDA* signal detection algorithm, *s, p* sensitivity precision

When we analyzed median sensitivity and median precision, broken by seven products for each SDA (graph not shown), the behavior of such measures was characterized by spread. As each product was compared across SDAs, we observed a pairing behavior of BAYER and ROCHE; RDT pairing with MHRA, and EMA pairing with ROR. The pair RDT and MHRA displayed less spread across products, when compared with all product SDAs.

The work of RDT (s, p) in selecting records of relevance in signal detection benefits from the parsimony of two-bias influence (C and B from Table 2) equally: unrelated events under drug exposure (C), and at the same time, reported events unrelated to drug exposure (B), play a crucial role in allowing these possible error excesses being model adjusted. We believe that such equilibrium produces a precise, valid, and equanimous selection process of the observations under study.

4.2 Comparisons of Median Misclassification Rates Per Product, Across SDAs

Figure 1 displays misclassification (MC) rates [3] by each product per SDA, inclusive of comparisons to BAYER, EMA, MHRA, RDT, ROCHE and ROR, a measurement

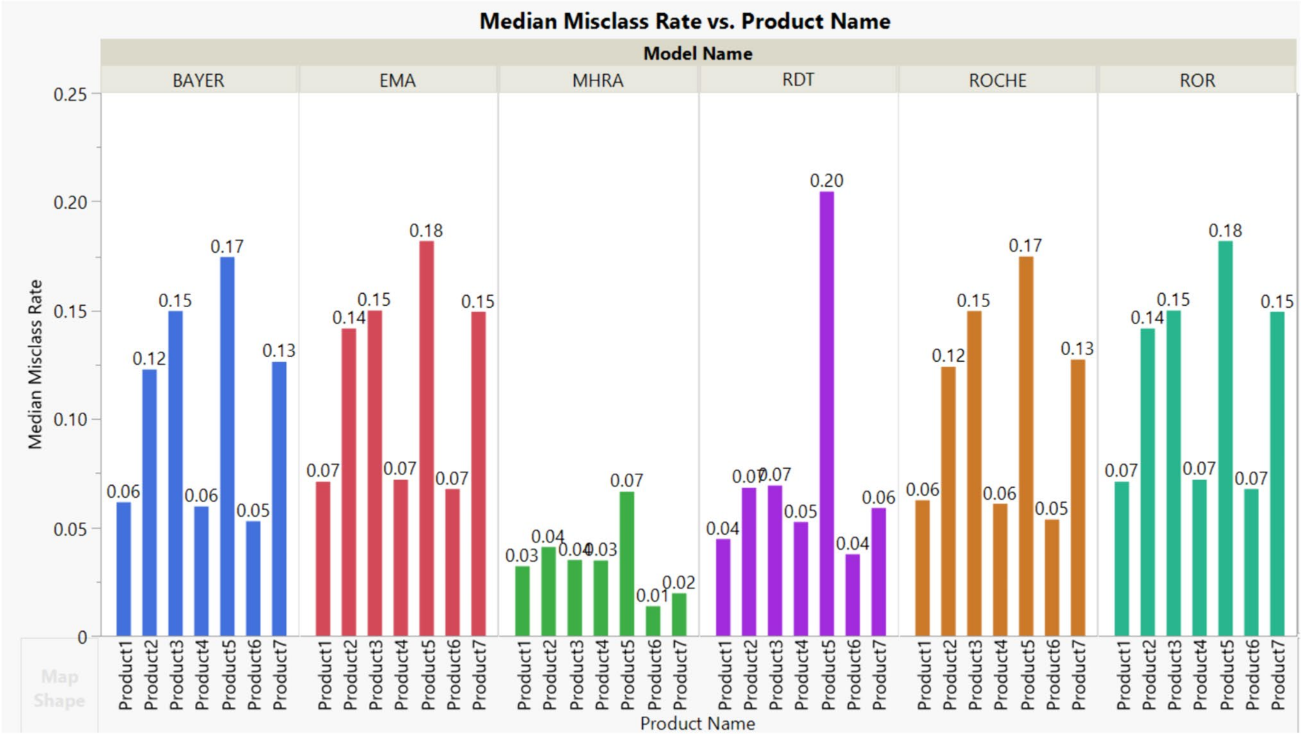

Fig. 1 Median Misclassification Rates by products and SDAs. *EB* Empirical Bayes, *EBGM* Empirical Bayes Geometric Mean, *PRR* proportional reporting ratio, *ROR* reporting odds ratio, *RDT* regression decision tree, *SDAs* signal detection algorithms

error that evaluates the accuracy of sensitivity and precision measures. The MHRA method displays the smallest MC rates (MC mean = 0.043), followed in descending order by RDT (MC mean = 0.100), BAYER (MC mean = 0.109), ROCHE (MC mean = 0.110), and EMA and ROR tied values (MC mean = 0.121). While all SDA methods have sudden increase and decrease changes in the MC rates, MHRA and RDT display stable small MC rates, except RDT, which points to a product 5 outlier. Of note, it is the MC rates of product 5, that display the highest rates across all methods. We observe mostly small (≥ 0.07) MC rates for all products, except for product 5, in MHRA and RDT. In summary, MHRA stands out with the smallest MC rates, followed by the RDT method. It is of interest to clarify that most products had multiple sclerosis (MS) indication; and one had indication for spinal muscular atrophy (SMA). Product 5 had MS indication specifically for patients with walking disability. All methodological procedures were equally applied to all products, and we are not sure why product 5 exhibited such a high MC rate. We know a priori that product 5 had the smallest sample size among all products, which could have driven spurious results.

4.3 Comparisons of Frequencies and Median Time to Detect the First Event, Across SDAs

Table 5 presents trends in frequencies and the median time to detect by number of observations, as the first signal associated with products compared across SDAs. When we consider the median time to detect and the frequencies of detected events, we desire to have the maximum number of events first detected in the shortest time spent by each method. In our data, these measures are approximately correlated. The RDT model had the quickest detection of new events by time (0.30) when compared to all other SDA methods, while detecting 118 observations. Of note, the absolute frequency of events is also valuable. If we drop the first cluster and compare RDT versus MHRA, RDT is the winner overall. We look forward to a method capable of a short and speedy time to detect events but also with a robust frequency of events, making both measures meaningful. The time unit was measured from the time a product enters the market until the last reported quarter of each product.

Table 5 Frequencies, median time to detect (mos.) and median time to detect (mos./obs.) by SDA

SDA	n	Median time to detect (mos.)	Median time to detect (mos.)/N
MHRA	64	25.94	0.41
RDT	118	34.95	0.30
BAYER	117	39.42	0.34
ROCHE	117	39.42	0.34
EMA	137	42.02	0.31
ROR	137	42.02	0.31

EMA European Medicine Agency, *MHRA* Medicines and Healthcare products Regulatory Agency, *n* number of adverse drug reports, *RDT* Regression Decision Tree, *ROR* reporting odds ratio, *mos.* months; *mos./obs.* months by observation, SDA signal detection algorithm

Next, we analyzed the median time to detect ADR events, broken down on product level. Medicines and Healthcare products Regulatory Agency had the highest time to detect an event for product 1. Except for product 7, RDT had the shortest times to detect signals across products (graph not shown).

4.4 Median Time to Detect the First Event by Products on Chronological Time, for MHRA and RDT

Figure 2 presents the median time to detect events by products, with chronological time on the x axis, comparing only the most efficient but similar methods, MHRA and RDT. The height of the small vertical lines on the graph represents the frequencies of events captured by each method, for each product, in each quarter. The blue color represents the frequencies of events captured by the MHRA method, and the red color represents the RDT method. The clouds with varying levels of red/blue intensity represent a smoothing effect intended to create a singular visual flow of event frequency over time where there are few-to-no events having occurred. Additionally, you can see products enter the market at various times where there is neither cloud nor line. When comparing both methods, the graph shows a higher prevalence of events across all quarters for RDT rather than MHRA, evidenced by the red color prevailing and dominating most areas of the graph in Fig. 2, especially for product 1, product 2 and product 7.

4.5 Comparisons of Frequencies of SDRs in Chronological Time, Across SDAs

Figure 3 displays a general comparison of frequencies of SDRs for all quarters since 2005, output by SDA methods, without product distinction. The colors are used to distinguish between the methods. The green color seen on the bottom of the graph is associated with the MHRA method, while the light purple indicates the dots associated with RDT method. These two methods, MHRA and RDT, detected most of the small quantities of SDRs, while the other SDR methods show a mix of colors with localized trends. Multiple occurrences of the same colors in a column (static point in time) are a function of multiple products experiencing an SDR. The overall view of the graph, drawn by the smooth line, shows that as time in the market for a product increases, the SDRs rates also increase.

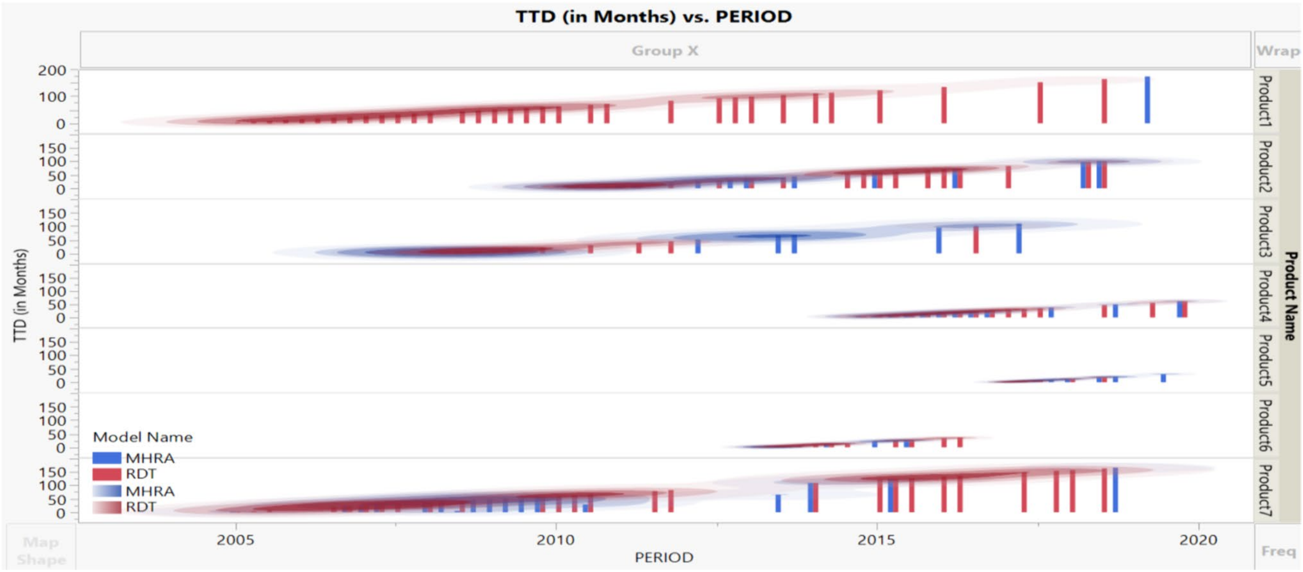

Fig. 2 Median time to detect in months by products comparing MHRA and RDT SDAs. *MHRA* Medicines and Health Care Products Regulatory Agency, *RDT* random decision tree, *SDAs* signal detection algorithms

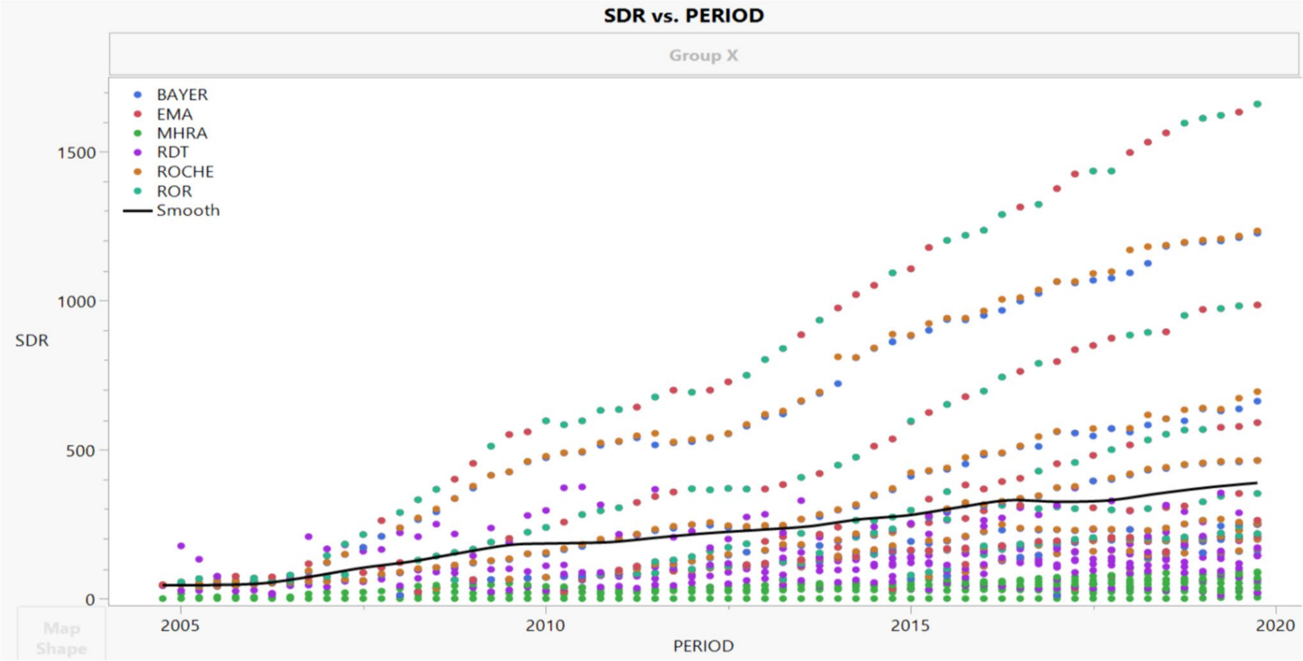

Fig. 3 Signals of disproportionate reporting by period (quarters) comparing all SDAs. *EMA* European Medicines Agency, *RDT* random decision tree, *ROR* reporting odds ratio, *SDAs* signal detection algorithms

4.6 Precision Rates in Chronological Time, Across SDAs

Next, as shown in Candore et al [2], the precision (or positive predictive value) over time can dictate future trends and the proportion of SDRs that turn into ADRs. We analyzed precision rates for all Biogen products by SDA methods during the totality of times of products in the market, some since 2005. The RDT and MHRA methods maintained stability in precision over increasing time periods, while the other methods have high volatility and decreasing trends (graph not shown). As mentioned in previous literature [8], a general reduction in precision during time in the market for most products is expected. Companies that monitor internal signal detection systems are compelled to explore advanced ML algorithms to compensate for decreasing precision.

5 Discussion

The proprietary decision tree approach we created for this experiment performed well in several key areas that are critical to detecting signals of disproportionate reporting. In terms of quantitative stability, RDT demonstrated the best trade-off between the study's key measures (s, p) among all methods compared. The RDT also proved superior in terms of time to detect a signal of disproportionality. However, the RDT did not perform as well as MHRA in misclassification rates across products. The RDT was also superior to MHRA in detecting SDRs, in total chronological time. The MHRA and RDT methods were parallel to one another in many of the results. This is quite encouraging as RDT is an ML approach using ratios as inputs while all other methods use count data.

We demonstrated a trend of stable and small MC rates for MHRA and RDT. After excluding product 5 outlier, we observed mostly small and acceptable ($\leq$0.07) MC rates for all products in MHRA and RDT. Although MHRA displays the smallest MC rates, followed by the RDT method, MHRA loses momentum when analyzing "rate of median time to detect events" by SDAs (MHRA 0.41; RDT 0.30).

The choice between granular, event-level ADR count data and aggregated, parameterized SDR measures (such as sensitivity, specificity, accuracy, etc.) as inputs for predictive models should generate a lively and healthy debate. As regulatory authorities continue to expect pro-active refinements and methodological breakthroughs in terms of predicting untoward drug events, a ML approach such as RDT, does represent a fresh foray in that spirit. Methods like RDT are worthy of consideration, at minimum, as a tool of governance. Current, well-accepted disproportion measurements depend directly on source data and formulaic simplicity. A decision tree approach allows for a top-down approach to (a) auto-generate common industry methods (ROR, PRR, EBGM, etc.) (b) control variances by generating a stable distribution (c) learn about proper thresholds and event count for detecting signals. Careful attention must be paid to likeness of drugs (drug classes) under consideration in models: indications/contra-indications, label warnings, intended consumers, route, etc. The period of data capture for the training data set and product market inception (most notably, time to event) may also affect interpretation of results.

Past literature has published increasing trends of SDRs across product time in the market [6] on a cross-sectional analysis of adverse events in the FDA Adverse Event Reporting System (FAERS) Database. Other works [7], which have analyzed SDRs trends, state the importance of periodicity (quarterly vs monthly), re-signaling, double detection of the same event in future time intervals, and threshold definitions when applying SDRs, in improving the quality of statistical signal detection or SDRs. These decisions influence the quality, efficiency, and workload of outputs for all SDRs methods.

Ideally, novel SDAs should demonstrate improvement in all outcomes over established SDAs. However, in this case, RDT championed the quickest time to detect ADRs and the highest number of detected known ADRs.

Sub-group analyses comparing SDA sensitivity, precision and SDR rates were calculated for events after demographic classification including age, gender and year groups. Small differences across classes of such variables were observed but no trend was statistically significantly observed. Tables are not shown.

For instance, the FAERS was used as a gold standard for ADR rates. Tables with model comparisons were created and results of Biogen GSD versus FAERS sensitivity and precision rates were not statistically significant. Tables are not shown. The establishment of a safety profile of products was not the goal of this work but rather a subproduct of such analysis.

The use of AUC measures [9] and F1-scores [10] would have been useful to summarize even more the performance of SDAs. However, we wanted to produce measures commonly used in literature to allow comparisons across different publications, therefore our choice for sensitivity and precision rates.

In order to address biases, we ran an equitable model, such that on step one, we predicted sensitivity, represented by the correlation between drug exposure with associated adverse events by all who took the drug, additionally, on step two, we predicted precision, represented by the correlation between drug exposure with associated adverse events by all who experienced adverse events. That way, we cautiously considered the two sources of bias reflected on the denominator of rates, from all drug exposure (a+c) and all product- and non–product-related adverse events (a+b). Our objective was to address bias coming from different sources of disproportionality rates and still obtain high rates for both sensitivity and precision.

In summary, RDT (s, p) methodological strengths are that the model is easily understandable and interpretable; the lack of non-linearity does not affect the model's performance; it is a non-parametric method and data distribution free, and robust against outliers and missing information. This model also maximizes the statistical effect on each observation while minimizing white noise and the number of hyper-parameters to be tuned. Its main disadvantage is over-fitting, which can be resolved using data splitting, cross-validation or comparisons with other ML algorithms like Random Forests, Neural Networks, etc.

6 Conclusions

In designing signal detection systems, careful consideration should be given to the criteria that are used to define SDRs. The choice of disproportionality statistics does not affect the achievable range of signal detection performance, but factors like periodicity and threshold definitions will influence the workload outputs of these SDRs. These choices should consider mainly ease of implementation, interpretation, and minimization of computing resources/platform dependencies. The changes in sensitivity and precision obtainable by replacing one algorithm with another are predictable. However, the absolute performance of a method is specific to its intended, particularized function. The RDT method in question here, attempted to take advantage of known industry-established methods and demonstrated the ability to customize SDAs on a per-product basis. Model refinements associated with internal and external factors influencing ADRs may improve ML methods, inclusive of RDT, in future developments. Other methods, such as Information Component (IC) have been discussed in other journals [4] and certainly showed strong results particularly when measured by AUC (i.e., sensitivity/specificity) against methods. While our experiment with RDT proved provocative, it is still not a thoroughly vetted or broadly understood approach and, at present, is not a tool employed in ongoing PV efforts.

As suggested here, there are many considerations when assessing the future of AI/ML and other new/emerging technologies in the realm of signal detection. The endeavor to further understand the safety of marketed drugs with these tools is not only a public 'consumer' good but an obligation when these tools are clearly available in the industry. Finally, the precision and reliability of such tools can be evolved if an industry/regulatory ecosystem allows for such experimentation.

Declarations

Funding The authors, all Biogen employees at time of writing, did not receive any direct compensation in the preparation (or publication) of this manuscript.

Conflicts of interest All authors and parties mentioned are/were full-time employees of Biogen during the process of executing, writing, and completing this work.

Availability of data and material The data cannot be made available at this time.

Ethics approval Not applicable

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability Code is not available for this work.

Authors' contributions All authors have read and approve the final version of the manuscript and agree to be accountable for the work. Dr. Eugenia Bastos created the RDT methodology and performed the comparative analysis. Additionally, Dr. Bastos wrote most of the text. Jeff K. Allen contributed text, edits, and tabular outputs. Jeff Philip's contributions include providing edits as well as research support for the paper.

References

1. Caster O, Aoki Y, Gattepaille LM, Grundmark B. Disproportionality analysis for pharmacovigilance signal detection in small databases or subsets: recommendations for limiting false positive associations. Drug Saf. 2023;2020(43):479–87. https://doi.org/10.1007/s40264-020-00911-w.Accessed06Jun.
2. Candore G, Juhlin K, Mahlik K, Thaklar B, Quarcoo N, Seabroke S, Wisniewski A, Slattery J. Comparison of statistical signal detection methods within and across spontaneous reporting databases. Drug Saf. 2015;38:577–87. https://doi.org/10.1007/s40264-015-0289-5.
3. Seabroke S. Statistical Signal Detection for Spontaneous Reports. Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium. Protect Symposium; 19–20 Feb 2015; Spontaneous Reporting Systems and Pharmacovigilance - WHO programme perspective (imi-protect.eu).
4. Ji X, Cui G, Xu CX, Jie H, Zhang Y, Ren Y. Combining a pharmacological network model with a Bayesian signal detection algorithm to improve the detection of adverse drug events. Front Pharmacol. 2022. https://doi.org/10.3389/fphar.2021.773135.
5. Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. Pharmacoepidemiol Drug Saf. 2004;13(8):519–23. https://doi.org/10.1002/pds.1001.
6. Zhou ZX, Yin XD, Zhang Y, Shao QH, Mao XY, Hu WJ, et al. Antifungal drugs and drug-induced liver injury: a real-world study leveraging the FDA adverse event reporting system database. Front Pharmacol. 2022;28(13): 891336. https://doi.org/10.3389/fphar.2022.891336.
7. Lerch M, Nowicki P, Manlik K, Wirsching G. Statistical signal detection as a routine pharmacovigilance practice: effects of periodicity and resignalling criteria on quality and workload. Drug Saf. 2015;38(12):1219–31. https://doi.org/10.1007/s40264-015-0345-1.
8. Wisniewski A, Bate A, Bousquet C. Good signal detection practices: evidence from IMI PROTECT. Drug Saf. 2016;39:469–90. https://doi.org/10.1007/s40264-016-0405-1.
9. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clin Pharmacol Ther. 2013;93(6):539–46. https://doi.org/10.1038/clpt.2013.24.
10. Yacouby R, Axman D. Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. 2020;10:79-91. https://doi.org/10.18653/v1/2020.eval4nlp-1.9.

Authors and Affiliations

Eugenia Bastos[1] · Jeff K. Allen[2] · Jeff Philip[2]

✉ Jeff K. Allen
jeff.allen@biogen.com

Eugenia Bastos
ebastos@bastosconsulting.com

Jeff Philip
Jeffrey.philip@biogen.com

[1] Cambridge, MA, USA

[2] Biogen, Safety and Benefit Risk Management,
Research Triangle Park, NC, USA