

Validation of New Signal Detection Methods for Web Query Log Data Compared to Signal Detection Algorithms Used With FAERS

Susan Colilla¹ · Elad Yom Tov² · Ling Zhang¹ · Marie-Laure Kurzinger³ ·
Stephanie Tcherny-Lessenot³ · Catherine Penfornis³ · Shang Jen⁴ ·
Danny S. Gonzalez⁵ · Patrick Caubel⁶ · Susan Welsh¹ · Juhaeri Juhaeri¹

© Springer International Publishing Switzerland 2017

Abstract

Introduction Post-marketing drug surveillance is largely based on signals found in spontaneous reports from patients and healthcare providers. Rare adverse drug reactions and adverse events (AEs) that may develop after long-term exposure to a drug or from drug interactions may be missed. The US FDA and others have proposed that web-based data could be mined as a resource to detect latent signals associated with adverse drug reactions.

Methods Recently, a web-based search query method called a query log reaction score (QLRS) was developed to detect whether AEs associated with certain drugs could be found from search engine query data. In this study, we compare the performance of two other algorithms, the proportional query ratio (PQR) and the proportional query

rate ratio (Q-PRR) against that of two reference signal-detection algorithms (SDAs) commonly used with the FDA AE Reporting System (FAERS) database.

Results In summary, the web query methods have moderate sensitivity (80%) in detecting signals in web query data compared with reference SDAs in FAERS when the web query data are filtered, but the query metrics generate many false-positives and have low specificity compared with reference SDAs in FAERS.

Conclusion Future research is needed to find better refinements of query data and/or the metrics to improve the specificity of these web query log algorithms.

Electronic supplementary material The online version of this article (doi:[10.1007/s40264-017-0507-4](https://doi.org/10.1007/s40264-017-0507-4)) contains supplementary material, which is available to authorized users.

✉ Susan Colilla
susan.colilla@sanofi.com

¹ Pharmacoepidemiology, Global Safety Sciences, Sanofi, 55 Corporate Dr., Bridgewater, NJ 08807, USA

² Microsoft Research, 13 Shenkar St., 4672513 Herzeliya, Israel

³ Pharmacoepidemiology & Signal Detection, Global Safety Sciences, Sanofi, 1 Avenue Pierre Brossolette, 91385 Chilly-Mazarin, France

⁴ Baxalta US, Inc., Global Drug Safety, 650 E. Kendall St., Cambridge, MA 02142, USA

⁵ US Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993, USA

⁶ Pfizer, Worldwide Safety, 235 E. 42nd St., New York, NY 10017, USA

Key Points

Internet search query methods provide an additional approach to examining a new and large data resource of internet search query data for signal detection of adverse events.

The best performance of the individual search query metrics was found with the Q-PRR and PQR metrics which had high sensitivity and moderate specificity to detect signals in web query data compared to metrics used for signal detection with FAERS database.

These new query metrics provide the ability to get faster results for the signal detection of adverse events than the current standard spontaneous adverse event reporting databases which have a long lag time in data releases.

1 Introduction

Multiple resources for signal detection exist. Commonly used spontaneous adverse event (AE) reporting systems such as the US FDA AE Reporting System (FAERS) and the World Health Organization's VigiBase include large amounts of data, but these systems have their limitations and may miss AEs [1]. It has been estimated that spontaneous AE reporting systems may under-report up to 98% of adverse drug events that occur [2], so alternative data resources have been considered [1, 3–6].

Signal-detection algorithms (SDAs), such as disproportionality and Empirical Bayes geometric mean (EBGM) metrics, have been used as the primary tools to detect signals in spontaneous AE reporting systems such as FAERS. However, these metrics were developed for use in spontaneous AE reporting systems and may not function appropriately with other data sources. Internet search engine logs and social media data have been proposed as supplemental data resources to mine for drug safety signals; however, the methods with which to process and analyze these large data sources are still under development [5, 7, 8]. Recently, the US FDA provided some draft guidance on how to utilize internet-based data sources for drug safety signal detection in the general population [1, 9, 10].

Internet search engine logs and social media data are new data sources that provide a faster detection system with a larger scope in which to capture health-related signals than traditional surveillance methods. While these internet 'big' data sources are plentiful, they also have limitations as the data are very noisy, they may produce many false-positives, and their performance in detecting signals has not been well validated. Some techniques that have been implemented have been noise-reduction techniques applied to the internet data (e.g., removing outliers, timing between drug and event, setting high thresholds) prior to analysis [5, 7] or restriction of data sources to datasets of known associations and medical professionals [11]. For example, a recent study by Odgers et al. [11] showed that search logs from healthcare professionals in a medical database can be used as a predictive data resource for detecting adverse drug reactions (ADRs) of well-established drug–event associations as well as a set of recently established ADRs. However, there is interest in examining all internet searches (as well as social media posts) to represent both healthcare provider and consumer input and capture a full picture of all potential ADRs.

The objective of this study was to assess the performance of newly proposed query metrics in detecting signals from web query log data compared with the signals detected with reference SDAs in the FAERS database.

2 Methods

2.1 Data Sources

We extracted all English language queries submitted to the Microsoft Bing search engine by users in the USA for the period 1 March 2013 to 30 September 2013. For each query, we extracted the query text, time, date, and an anonymized user identifier. To maintain user privacy, identifiers were first anonymized by hashing before the investigators gained access to them. They were then aggregated prior to analysis, and no individual-level user datum was examined by the experimenters. Data extraction and normalization processes were performed in a manner similar to that described in a previous publication [7].

Ten single-agent drugs marketed in the USA by Sanofi were identified and used in this analysis. The selected medications included drugs with different characteristics, including those from different drug classes, from different therapeutic areas (cardiovascular/diabetes, immunology, oncology, and central nervous system agents), and drugs with different administration routes (pills, injectables, etc.). Several drugs were also new to the market in the last few years, whereas others have been on the market for decades and are available as over-the-counter medications. These drugs were selected because Sanofi is the market authorization holder for these specific drugs and the authors were familiar with the drug's safety profile from multiple internal sources. The drugs' indications were widespread to ensure the metrics would capture signals over a widespread number of indications and events.

Health event-related terms were taken from a list of 195 symptoms from the *International Statistical Classification of Diseases and Related Health Problems* (ICD) and expanded with synonyms to capture nonmedical terminology used by patients in a web search (see Table S1 in the Electronic Supplementary Material [ESM]) as described previously [7]. These health-related terms were then mapped to ICD, Ninth Revision, Clinical Modification (ICD-9-CM) codes as described previously in a publication that introduced the query log reaction score (QLRS) method for signal detection. The ICD-9-CM terms were then mapped manually by investigators between the medical symptoms found in the query log data and the Medical Dictionary for Regulatory Activities (MedDRA[®], v.17) preferred terms (PTs) listed in FAERS. The first step was to group the medical terms or symptoms with the same meaning or those with typographical errors ('typos'), i.e., amnesia or can't remember. They were then standardized into the most relevant medical terms and encoded into the MedDRA[®] PTs.

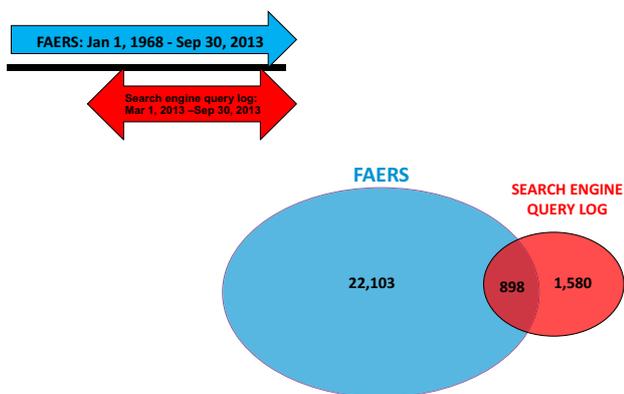


Fig. 1 Time periods covered by the US FDA adverse event reporting system (FAERS) database and search engine query log data and the amount of overlap

Web search queries were filtered to include only those that contained one of the above-mentioned health events and one of the ten medicines identified by brand name. The query data contained 1580 possible pairs. FAERS data contained 22,103 drug–event pairs, of which 898 drug–event pair combinations overlapped with the query data (see Fig. 1).

For validation purposes, data from the FDA spontaneous reporting database, FAERS, were used to assess whether the same signals could be detected in both data sources using SDAs. Empirica™ Signal 7.3 system (Oracle®) was used to extract and clean the raw safety data from the FAERS database. The AEs reported for the ten drugs of interest were extracted from the FAERS database within the time period 1 January 1968 to 30 September 2013; this search retrieved 22,103 records of drug–event pairs for the drugs of interest and reported health events coded by MedDRA® PTs and standardized MedDRA® narrow and algorithmic query terms. When compared with the 1580 drug–event pairs obtained from the query log data, some FAERS records matched multiple times. Thus, 22,224 records from FAERS were included in the database for the analysis. The MedDRA® terms captured in this database were then mapped to ICD-9-CM terminology via a MedDRA® to ICD-9-CM code dictionary.

Table 1 2 × 2 table for detecting disproportionate reporting of adverse events from spontaneous reporting database

	Event of interest	All other events	Total
Drug of interest	<i>a</i>	<i>b</i>	<i>a</i> + <i>b</i> = <i>M</i> ₁
All other drugs	<i>c</i>	<i>d</i>	<i>c</i> + <i>d</i> = <i>M</i> ₂
Total	<i>a</i> + <i>c</i> = <i>N</i> ₁	<i>b</i> + <i>d</i> = <i>N</i> ₂	<i>a</i> + <i>b</i> + <i>c</i> + <i>d</i> = <i>N</i>

Common signal-detection algorithms used to measure disproportionality:

Frequentist approach: proportional reporting ratio (PRR) = (*a*/*M*₁)/(*c*/*M*₂)

Bayesian method: Empirical Bayes Geometric Mean (EBGM) = models the distribution of observed/expected counts: (*a*/*M*₁)/(*N*₁/*N*)

2.2 Signal-Detection Algorithms

2.2.1 Metrics Applied to the US FDA Adverse Event Reporting System (FAERS) for Signal Detection

The current standard SDAs used for disproportionality analysis (DPA) in spontaneous AE reporting databases include the multi-item Gamma Poisson shrinker (MGPS) program providing the EBGM with the lower bound of the 90% confidence interval (CI) of EBGM (EB05), and the proportional reporting ratio (PRR) [12, 13].

Two commonly used measures for signal detection based on DPA in the FAERS database were selected as reference measures for our analysis (Table 1) [13–15]. First, a frequentist method commonly used is the PRR, which measures the proportion of reports of an event for drug A relative to that for all other drugs. For the PRR, a commonly used SDA consists of a $PRR \geq 2$, with the number of reports (*N*) ≥ 3 , and a $PRR \chi^2 \geq 4$ [15]. A Bayesian method, the MGPS program, uses a stratified full independence model and derives adjusted ratios of the observed versus expected counts. The MGPS program estimates an EBGM that approximates a χ^2 statistic: $EBGM = (a/M_1)/(N_1/N)$, and the most commonly used signal threshold for this statistic is $EB05 \geq 2$ [13]. Thresholds of $EBGM \geq 2$ and $EBGM \geq 4$ were also considered as evidence of a signal of disproportionate reporting in this analysis. These aforementioned SDAs and threshold cutoffs were considered the reference metrics for which to compare the performance of the query metrics as part of this validation exercise.

2.2.2 Metrics Applied to Search Engine Query Data

The search engine QLRs metric was the first metric compared with the reference SDA metrics.

The QLRs is computed as the χ^2 score of a two-by-two table (Table 2) comprising (1) the number of people who asked about an event before versus after they first asked about the drug and (2) the number of people who asked about the same event before versus after an arbitrary date. The latter are used as a normalization factor, e.g., for

Table 2 2×2 table for estimating disproportionate queries from web-based query log data

When user queried for the event	User queried for the drug?	
	No	Yes
Before day 0	N_{11}	N_{12}
Day 0 or after	N_{21}	N_{22}
Total	N_1	N_2

Query log-based metrics to measure disproportionality: query log reactions score (QLRS) = χ^2 of table above

Proportional query ratio (PQR) = $(N_{22}/N_2)/(N_{21}/N_1)$

Table 3 New 2×2 table for query proportional rate ratio (Q-PRR) metric for web-based query log data

Query for the event	Query for the drug	
	Drug A	Comparator: all other drugs
Event of interest	a	b
All others events	c	d
Total	$a + c = M_1$	$b + d = M_2$

Additional query metrics for disproportionality defined:

Query proportional rate ratio (Q-PRR) = $(a/M_1)/(b/M_2)$

Query odds ratio = $(a \times d)/(b \times c)$

This approach discards non-drug search queries and does not count events occurring prior to first day of drug name search

seasonal events, etc. Since there are no established thresholds for the QLRS metric, which equates a χ^2 statistic, this study examined two cutoffs: $QLRS \geq 5$ (median value for QLRS distribution) and $QLRS \geq 10$. The QLRS test measures the statistical probability of a drug and event association, but it does not measure the strength of an association between the event and a drug, so an alternative metric, called the proportional query ratio (PQR) was developed. The PQR represents the proportion of queries for an event on or after day 0 (the day when that user first searched for the drug) in users searching for a drug relative to the proportion of event queries in users not searching for the drug (Table 2). For users who did not search for the drug, day 0 is the midpoint between the first and last day of their queries.

When the query log data were considered in a manner similar to DPA PRR (see the 2×2 table in Table 3), a new approach was created and named the query proportional rate ratio (Q-PRR). The Q-PRR approach discards non-drug search queries and does not count events occurring before the first day of a drug name search. The Q-PRR examines the number of queries for the drug of interest and the number of queries for the event of interest compared with the number of queries for all other events. From Table 3, the Q-PRR = $(a/N_1)/(b/N_2)$.

Additional factors were examined (e.g., sample size [number of drug–event pairs], exclusions, and cutoff values) to see whether they would improve the metric’s performance, and a combination logistic model of all these factors was developed. Other measures examined included the number of drug–event pairs for QLRS metric (N-QLRS) and Q-PRR (N-Q-PRR), outliers, $PQR > 1$, and the χ^2 of Q-PRR. These measures were then used together in a composite model for testing. The composite model was created with a logistic regression model that included each query test metric and the additional factors listed above (e.g., outliers, sample sizes, etc.) for a total of 11 independent variables to predict overall outcome (FAERS metric outcome).

2.3 Statistical Analysis

The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the query metrics were calculated to validate the method against the reference SDAs used with the FAERS database. Receiver operating characteristic (ROC) curves [16] were performed to measure the performance of the query metrics, and the area under the curve (AUC) was estimated to show how well a test discriminates between positive and negative signals. C statistics were used to calculate an AUC that reflected the probability of concordance. The C statistic is a critical and equivalent parameter to measure the area under the ROC curve (AUC), which estimates the predictive accuracy of a logistic regression model [17].

Sensitivity testing regarding the query data was examined by excluding outliers and drug–event pairs with small sample sizes (where fewer than 100 Bing users searched for both drug and event) to see whether this modification would improve the performance of the query metrics. Outliers were defined, according to previous work by Yom-Tov and Gabrilovich [7], as the five symptoms found using a greedy selection process to most reduce the value of the correlation between QLRS and FAERS counts.

Next, a sensitivity analysis was also performed on variations of the PRR reference SDA by increasing the minimum number of drug–event pairs needed to be considered a signal to assess its concordance with the query metrics.

3 Results

Figure 1 shows the time periods covered by both databases and their period of overlap. There were 22,103 drug–event pairs in the FAERS database and 1580 drug–event pairs in the search engine query log database. The

898 drug–event pairs found in the FAERS and the search engine query log databases were used for comparison in this analysis. Distributions of the web query metrics are shown in Fig. S1 in the Electronic Supplementary Material.

3.1 Sensitivity/Specificity Performance

The sensitivity and specificity of the query metrics on all 898 drug–event pairs were estimated with the reference SDAs in FAERS data (Tables 4, 5, 6). When all drug–

Table 4 Query log reaction score metric sensitivity and specificity (%)

Reference SDA thresholds	Query log cutoff ^a	Sensitivity	False-positive rate (1–specificity)	PPV	NPV
Based on all drug–event pairs ($N = 898$ pairs)					
EB05 ≥ 2	QLRS ≥ 5.0	54.17	50.47	5.71	95.03
EBGM ≥ 2	QLRS ≥ 5.0	50.59	50.68	9.45	90.52
EBGM ≥ 4	QLRS ≥ 5.0	72.73	50.39	1.76	99.32
$N \geq 3$ and PRR ≥ 2 and PRR_CHISQ ≥ 4	QLRS ≥ 5.0	46.55	51.28	11.87	86.00
Excluding pairs with small sample size ($N = 606$ pairs)					
EB05 ≥ 2	QLRS ≥ 5.0	81.25	74.74	5.71	96.03
EBGM ≥ 2	QLRS ≥ 5.0	79.63	74.64	9.45	92.72
EBGM ≥ 4	QLRS ≥ 5.0	80.00	75.00	1.76	98.68
$N \geq 3$ and PRR ≥ 2 and PRR_CHISQ ≥ 4	QLRS ≥ 5.0	72.97	75.38	11.87	86.76
Excluding pairs with small sample size and outliers ($N = 571$ pairs)					
EB05 ≥ 2	QLRS ≥ 5.0	83.33	74.86	5.81	96.45
EBGM ≥ 2	QLRS ≥ 5.0	79.59	74.90	9.07	92.91
EBGM ≥ 4	QLRS ≥ 5.0	80.00	75.22	1.86	98.58
$N \geq 3$ and PRR ≥ 2 and PRR_CHISQ ≥ 4	QLRS ≥ 5.0	72.86	75.64	11.86	86.52

EBGM empirical Bayes geometric mean, NPV negative predictive value, PPV positive predictive value, PRR proportional reporting ratio, PRR_CHISQ proportional reporting ratio chi-square, QLRS query log reaction score, SDA signal-detection algorithm

^a Results were similar when using a QLRS cutoff threshold of 10

Table 5 Proportional query ratio metric sensitivity and specificity (%)

Reference SDA thresholds	Query log cutoff	Sensitivity	False-positive rate (1–specificity)	PPV	NPV
Based on all drug–event pairs ($N = 898$ pairs)					
EB05 ≥ 2	PQR ≥ 1.0	54.17	43.88	6.52	95.59
EBGM ≥ 2	PQR ≥ 1.0	47.06	44.16	10.03	90.98
EBGM ≥ 4	PQR ≥ 1.0	81.82	43.97	2.26	99.60
$N \geq 3$ and PRR ≥ 2 and PRR_CHISQ ≥ 4	PQR ≥ 1.0	47.41	43.99	13.78	87.78
Excluding small sample size ($N = 606$ pairs)					
EB05 ≥ 2	PQR ≥ 1.0	75.00	57.84	6.74	96.80
EBGM ≥ 2	PQR ≥ 1.0	64.81	58.15	9.83	92.40
EBGM ≥ 4	PQR ≥ 1.0	80.00	58.39	2.25	99.20
$N \geq 3$ and PRR ≥ 2 and PRR_CHISQ ≥ 4	PQR ≥ 1.0	64.86	57.89	13.48	89.60
Excluding small sample size and outliers ($N = 571$ pairs)					
EB05 ≥ 2	PQR ≥ 1.0	73.33	55.27	6.85	96.80
EBGM ≥ 2	PQR ≥ 1.0	61.22	55.75	9.35	92.40
EBGM ≥ 4	PQR ≥ 1.0	80.00	55.79	2.49	99.20
$N \geq 3$ and PRR ≥ 2 and PRR_CHISQ ≥ 4	PQR ≥ 1.0	62.86	55.29	13.71	89.60

EBGM empirical Bayes geometric mean, NPV negative predictive value, PPV positive predictive value, PQR proportional query ratio, PRR proportional reporting ratio, PRR_CHISQ proportional reporting ratio chi-square, SDA signal-detection algorithm

Table 6 Query proportional rate ratio metric sensitivity and specificity (%)

Reference SDA thresholds	Query log cutoff	Sensitivity	False-positive rate (1-specificity)	PPV	NPV
Based on all drug-event pairs ($N = 898$ pairs)					
$EB05 \geq 2$	$Q\text{-PRR} \geq 1.3$	43.75	33.41	6.89	95.45
$EBGM \geq 2$	$Q\text{-PRR} \geq 1.3$	47.06	32.59	13.11	92.41
$EBGM \geq 4$	$Q\text{-PRR} \geq 1.3$	54.55	33.71	1.97	99.16
$N \geq 3$ and $PRR \geq 2$ and $PRR_CHISQ \geq 4$	$Q\text{-PRR} \geq 1.3$	45.69	32.23	17.38	89.38
Excluding small sample size ($N = 606$ pairs)					
$EB05 \geq 2$	$Q\text{-PRR} \geq 1.3$	62.50	57.84	7.35	96.41
$EBGM \geq 2$	$Q\text{-PRR} \geq 1.3$	68.52	42.57	13.60	94.91
$EBGM \geq 4$	$Q\text{-PRR} \geq 1.3$	50.00	44.80	1.84	98.50
$N \geq 3$ and $PRR \geq 2$ and $PRR_CHISQ \geq 4$	$Q\text{-PRR} \geq 1.3$	66.22	41.92	18.01	92.52

EBGM empirical Bayes geometric mean, *NPV* negative predictive value, *PPV* positive predictive value, *PRR* proportional reporting ratio, *PRR_CHISQ* proportional reporting ratio chi-square, *Q-PRR* query proportional rate ratio, *SDA* signal-detection algorithm

event pairs were analyzed, the QLRs had a sensitivity ranging from 46.6 to 72.7% and a high false-positive rate (1-specificity), between 50.39 and 51.28% depending on which of the four SDAs were used in the FAERS database (Table 4). The sensitivity improved greatly for QLRs when the drug-event pairs with small sample sizes ($n = 292$ pairs) were excluded from analysis, but the false-positive rate increased dramatically to around 75% for the reference SDAs. The corresponding PPV was very poor, never achieving higher than 11.87% with any of these SDAs, but the NPV remained high, ranging from 86.00 to 99.32%. Removing outliers ($N = 35$) in addition to drug-event pairs with small sample sizes from the data did not measurably change the performance of the QLRs.

The sensitivity and false-positive rate were slightly better when using the PQR than with QLRs, around 47.06–81.82 and 43.88–44.16%, respectively, when tested using all 898 drug-event pairs (Table 5). Again, the sensitivity improved from 64.81 to 80.00% when drug-event pairs with small sample sizes were excluded; however, the false-positive rate remained moderately high (57.84–58.39%). The best PPV was found for PQR against the combined metric using PRR SDA, but the PPV was still quite low, below 14%. Removing outliers also did not change the overall performance of the PQR.

The Q-PRR metric showed a stronger specificity with the reference for all drug-event pairs (66.29–67.77%; false-positive rate 32.23–33.71%) but exhibited a lower sensitivity (43.75–54.55%) than PQR (Table 6). Excluding smaller sample sizes improved the sensitivity somewhat (50.00–68.52%) but also increased the false-positive rate (42.57–57.84%). Similar to PQR, the Q-PRR had the strongest PPV and NPV values when tested with the reference combined SDA using the PRR, N , and χ^2 metric. Overall, the best performance for the Q-PRR was found

with query data that excluded smaller samples where the NPV was estimated at 92.52% and PPV at 18.01% when tested against the reference PRR SDA.

When excluding the smaller sample sizes and outliers, the sensitivity of each of the metrics increased and, as expected, the false-positive rate of these tests also increased (specificity of tests decreased). In summary, of the query metrics tested against the reference SDAs, the Q-PRR metric had the lowest false-positive rate (32.2%; Table 6) and the QLRs had the highest false-positive rate (51.2%; Table 4) when compared with the reference PRR SDA based on all drug-event pairs.

3.2 Area Under the Curve (AUC)

The performance of each query metric compared with a reference measure was graphed and the AUC measured for all 898 drug-event pairs. Table 7 shows the AUC value for the performance of each of the query metrics against the reference metrics. For the QLRs metric, the best concordance was found with the reference of $EBGM \geq 2$, and this AUC estimate did not deviate much from random expectations at 0.536.

The PQR metric performed best with the $EBGM \geq 4$ SDA and resulted in a higher AUC estimate of 0.652 (Fig. 2). Lastly, the Q-PRR metric had an AUC value (0.652) similar to the PQR with the $EBGM \geq 4$ cutoff (Fig. 3). While there was no statistically significant difference between the PQR and Q-PRR AUCs, both of these metrics were significantly different from QLRs AUC ($p = 0.05$, two-tailed test) when compared with the signals found with $EBGM \geq 4$ SDA.

When determining the performance of the metrics using AUC curves, the Q-PRR (and PQR) had the highest value (0.65) when using the full dataset of all drug-event pairs

Table 7 Performance area under the curve of the query metrics and reference for adverse event reporting system

Reference SDA thresholds	Total number of positive cases for reference	AUC—metrics for query log data			
		QLRS	PQR	Q-PRR	Composite model of 11 metrics
$EB05 \geq 2$	48	0.513	0.497	0.526	0.6575
$EBGM \geq 2$	85	0.536	0.511	0.539	0.6162
$EBGM \geq 4$	11	0.413	0.652	0.652	0.8211
$N \geq 3$ and $PRR \geq 2$ and $\chi^2 \geq 4$	116	0.527	0.510	0.541	0.5583

AUC area under the curve, EBGM empirical Bayes geometric mean, PQR proportional query ratio, PRR proportional reporting ratio, QLRS query log reaction score, Q-PRR query proportional rate ratio, SDA signal-detection algorithm

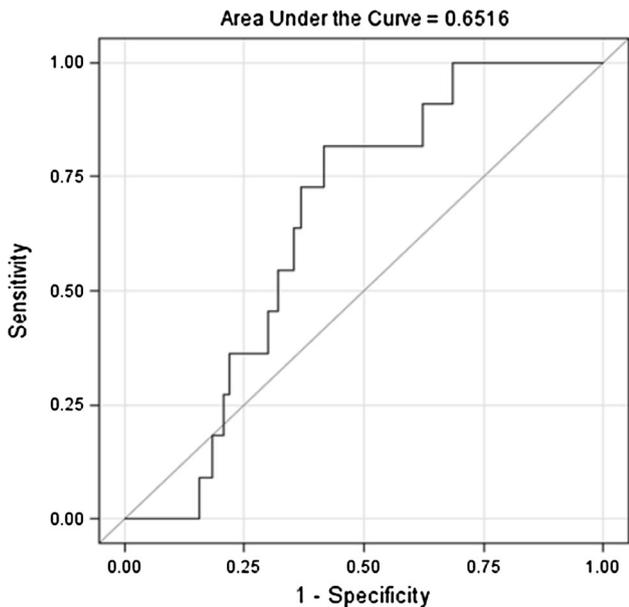


Fig. 2 Receiver operating characteristic curve for PQR metric with $EBGM \geq 4$ signal-detection algorithm

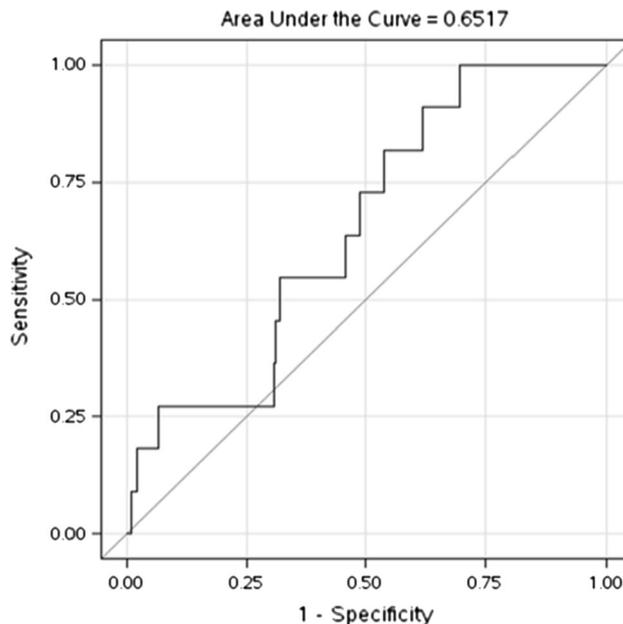


Fig. 3 Receiver operating characteristic curve for Q-PRR metric with $EBGM \geq 4$ signal-detection algorithm

($EBGM \geq 4$ cutoff). However, the metric with the highest PPV and NPV is the Q-PRR query metric when using the reference PRR SDA cutoff for FAERS data.

3.3 Additional Modifications to Maximize Performance

Sensitivity testing with the reference PRR SDA was also performed to assess its effect on the tested query metrics ROC AUC. Because the reference PRR SDA had the lowest false-positive rate for the query metrics, the AUC values were evaluated by increasing the value of N (number of reports for a drug–event pair), basically varying the threshold for the reference. The AUC results for the query metrics against different N values of the reference PRR

SDA are shown in Table 8. Increasing the size of N in the PRR gold standard variable increased the AUC value for some of the query metrics. In particular, the Q-PRR metric had the highest AUC value (0.614) when the PRR threshold required that N be ≥ 200 reports.

When all the metrics and measures were combined into a composite model of metrics (CMM) to measure performance, the AUC increased for most of the reference SDAs (Table 7). Again, the performance of the CMM had strongest agreement with the $EBGM \geq 4$ SDA, with an AUC of 0.821 (Fig. 4). The performance of the CMM improved from 0.558 to 0.758 when changing the various reference threshold cutoffs for the PRR metric (Table 8).

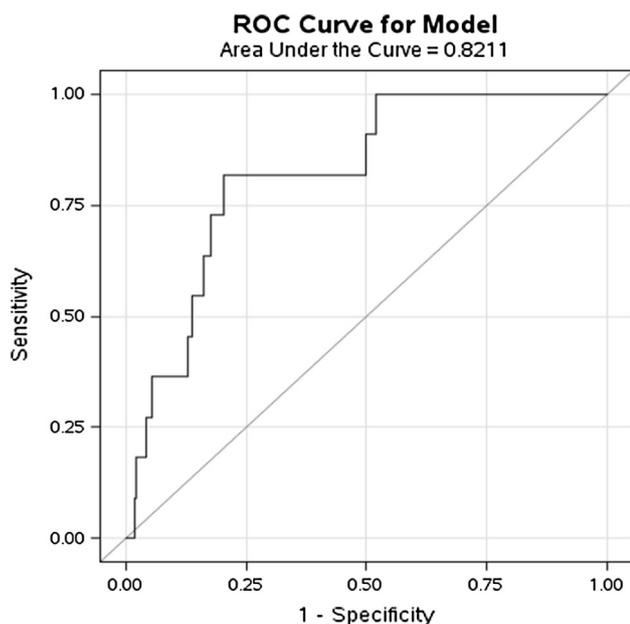


Fig. 4 Receiver operating characteristic curve for composite model of 11 metrics by $EBGM \geq 4$ signal-detection algorithm

4 Discussion

The best performance of the individual search query metrics was found with the Q-PRR and PQR metrics when testing against the $EBGM \geq 4$ as the SDA for FAERS. The PQR test had the highest sensitivity and highest AUC value for the internet search query metrics when using the $EBGM \geq 4$ as the reference. The Q-PRR had the highest specificity (lowest false-positive rate) as well as the highest AUC value for the search query metrics. However, these metrics were not very specific for any of the reference metrics, as they still had many false-positives.

When all the query metrics were put together into a composite model, the agreement with each of the reference SDAs improved. The composite model had the strongest agreement with the $EBGM \geq 4$ SDA, with an AUC of 0.82. This demonstrates that the addition of each metric to a composite model improved the predictability with the

reference over each metric individually. It also suggests that a surveillance system may require monitoring of multiple metrics for optimal detection.

The high false-positive rate may represent false signals (creating noise) or, alternatively, AEs that are under-reported (or unreported) in the FAERS database. Yom-Tov and Gabrilovich [7] suggested that under-reporting may be the more prominent of the two, as late-appearing less acute AEs can be detected using internet search query log data while being under-reported in FAERS. Therefore, additional investigation is needed to form more encompassing gold standards than FAERS.

Here, we focused on developing new measures to screen the web query data for potential signals using the FAERS database as the gold standard. While a spontaneous reporting system such as FAERS is not a true gold standard, it is a standard used by health authorities and has very practical applications in pharmacovigilance practices commonly used within industry.

The Q-PRR test had the lowest false-positive rate, at around 33%, across all the FAERS reference SDAs when tested among all the drug–event pairs. The lowest Q-PRR false-positive rate was obtained using the reference PRR SDA (32.2%); this may be because the Q-PRR has an algorithm design similar to that of the PRR measure used in this reference. The PQR test had a higher false-positive rate, at around 44%, suggesting that the best overall performing individual metric for any of the reference SDAs was the Q-PRR test. The CMM test exhibited the strongest agreement with the $EBGM$ metrics, particularly when $EBGM \geq 4$ was the reference.

These query log methods appear to capture signals similar to those captured by the reference PRR SDA and yet may also find additional statistical signals that may or may not be true safety signals. Limitations to these search query methods for signal detection include that a drug and event searched for by the same user does not necessarily equate to an AE occurring in an individual. Media attention, patients reading drug labels, and numerous other reasons may cause individuals to search a drug name and a symptom/AE at the same time. Despite these limitations, a

Table 8 Sensitivity analysis of increasing the sample size threshold for proportional reporting ratio reference metric

Threshold cutoffs for reference PRR metric	Total number of positive cases for reference	AUC—metrics for query log data			
		QLRS	PQR	Q-PRR	Composite model of 11 metrics
$N \geq 3$ and $PRR \geq 2$ and $\chi^2 \geq 4$	116	0.527	0.510	0.541	0.5583
$N \geq 50$ and $PRR \geq 2$ and $\chi^2 \geq 4$	59	0.527	0.519	0.507	0.6019
$N \geq 200$ and $PRR \geq 2$ and $\chi^2 \geq 4$	26	0.573	0.576	0.614	0.7583
$N \geq 300$ and $PRR \geq 2$ and $\chi^2 \geq 4$	21	0.556	0.555	0.586	0.7489

AUC area under the curve, PQR proportional query ratio, PRR proportional reporting ratio, QLRS query log reaction score, Q-PRR query proportional rate ratio

good measure of signal detection with known reliability and validity is needed to find true safety signals from search engine data because a high volume of internet data can be mined for safety signals when these data are used in the appropriate context [16]. This analysis is one way to enhance signal detection of AEs in a manner faster than is currently available with the spontaneous AE reporting system, FAERS. That said, query log data contain high levels of noise, and the best methods to reduce the noise and enhance searches for true safety signals are still being explored [5, 8, 18]. Any signal identified by SDAs using DPA requires further medical analysis for signal validation to determine the true safety signals warranting a thorough evaluation to assess whether they represent a risk for patients receiving the drug (e.g., qualitative analysis of case narratives). Further work is also needed to validate the predictive ability of internet search query log-derived signals for safety events [5].

At least one study that compared the performance of the reference SDAs in other healthcare databases with that in spontaneous-reporting databases found a similarly poor performance of the SDAs, suggesting these disproportionality methods did not discriminate true positives from true negatives in the observational healthcare databases and simulated datasets as they reportedly do with spontaneous-reporting data [19]. A recent study by Odgers et al. [11] showed that web search logs from healthcare professionals using a medical database can be used as a predictive data resource to detect ADRs with well-established drug–event associations as well as a set of recently established ADRs. This work utilized two different reference standards to validate the methods but was also limited to data provided by healthcare professionals rather than searches conducted by the general public.

Other limitations known to exist with spontaneous event-reporting databases such as the FAERS database are that these databases need extensive pre-processing prior to analysis to remove duplicate reports, correct terminology errors, collate event terms and drug names into relevant categories, and remove or adjudicate cases with missing data [20]. The performance of the current standard SDAs for FAERS and other spontaneous-reporting databases has been shown to vary by prevalence of AE, medical terminology classifications, how long the drug has been marketed, and the type of AE database being examined [12, 14, 21].

It has been suggested that spontaneous-reporting databases under-represent the true number of AEs that occur in the population [2]. The web query log data may capture additional AEs that are not reported to federal regulatory bodies, and this data source could be considered an augmented approach to FAERS database for signal detection. Testing the performance of any new method when

compared with spontaneous-reporting databases such as FAERS as the reference will remain a challenge. Alternate gold standards may need to be identified. While the generalizability of these query metrics is limited, we think this approach contributes significantly to the developing field of web-based signal detection where reliable methodologies have yet to be established.

5 Conclusions

Internet search query methods may provide an additional approach to examining a new and large data resource for signal detection of AEs. Of the three individual metrics analyzed, the PQR and Q-PRR metrics performed the best; however, they still generate a moderate level of false-positives compared with the reference SDAs. A composite model of the query metrics had the best performance overall, yet further testing needs to be considered prior to implementation in signal detection. These query-based metrics were developed primarily for hypothesis generation and provide a supplemental method to examine alternative data sources (e.g., web searches, social media) in pharmacovigilance. It must also be remembered that signals in social media and internet search data may not be true safety signals, being instead due to circumstantial events within the media (e.g., increased news media coverage of a health risk). Further enhancements will be needed to provide a validated and reliable tool for signal detection in the future.

Compliance with Ethical Standards

Funding This research was fully funded by Sanofi.

Conflict of interest Susan Colilla, Ling Zhang, Marie-Laure Kurzinger, Stephanie Tcherny-Lessenot, Catherine Penformis, Susan Welsh, and Juhaeri Juhaeri are all employed by Sanofi. Elad Yom Tov is employed by Microsoft Research and provided the data and methods expertise for this project. Dr. Patrick Caubel and Shang Jen were employed by Sanofi while working on this research. Dr. Danny S. Gonzalez also contributed to this article as a former employee of Sanofi Pharmaceuticals. The views expressed are his own and do not necessarily represent the views of his current employer, the US FDA, or the US Government.

References

1. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. *J Biomed Inform.* 2015;54:202–12.
2. Fletcher AP. An appraisal of spontaneous adverse event monitoring. *Adverse Drug React Toxicol Rev.* 1992;11(4):213–27.
3. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Saf.* 2014;37(10):777–90.

4. Morel P, Vandel B. Adverse drug reaction monitoring and the Internet: evaluation of the use of the Internet by French Pharmacovigilance Centres and a non-exhaustive survey of websites of interest for collecting information about adverse drug reaction. *Therapie*. 1999;54(5):525–32.
5. White RW, Harpaz R, Shah NH, DuMouchel W, Horvitz E. Toward enhanced pharmacovigilance using patient-generated data on the internet. *Clin Pharmacol Ther*. 2014;96(2):239–46.
6. Sloane R, Osanlou O, Lewis D, Bollegala D, Maskell S, Pirmohamed M. Social media and pharmacovigilance: a review of the opportunities and challenges. *Br J Clin Pharmacol*. 2015;80(4):910–20.
7. Yom-Tov E, Gabrilovich E. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *J Med Internet Res*. 2013;15(6):e124.
8. Yang M, Kiang M, Shang W. Filtering big data from social media—Building an early warning system for adverse drug reactions. *J Biomed Inform*. 2015;54:230–40.
9. Dal Pan G, editor. Adverse event reporting and medication safety considerations: a view from CDER's Office of Surveillance and Epidemiology. CHPA Regulatory, Scientific, and Quality Conference: Washington, DC; 2013.
10. US Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER). Center for Biologics Evaluation and Research (CBER). Guidance for industry. Postmarketing Safety Reporting for Human Drug and Biological Products Including Vaccines. Rockville: US Department of Health and Human Services; 2001.
11. Odgers DJ, Harpaz R, Callahan A, Stiglic G, Shah NH. Analyzing search behavior of healthcare professionals for drug safety surveillance. *Pac Symp Biocomput*. 2015:306–17.
12. Almenoff JS, LaCroix KK, Yuen NA, Fram D, DuMouchel W. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. *Drug Saf*. 2006;29(10):875–87.
13. 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. *Drug Saf*. 2002;25(6):381–92.
14. 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.
15. 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(6):483–6.
16. Powell GE, Seifert HA, Reblin T, Burstein PJ, Blowers J, Menius JA, et al. Social media listening for routine post-marketing safety surveillance. *Drug Saf*. 2016;39(5):443–54.
17. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12:82.
18. Sarntivijai S, Abernethy DR. Use of internet search logs to evaluate potential drug adverse events. *Clin Pharmacol Ther*. 2014;96(2):149–50.
19. DuMouchel W, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to healthcare databases. *Drug Saf*. 2013;36(Suppl 1):S123–32.
20. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combining signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *J Am Med Inform Assoc*. 2013;20(3):413–9.
21. Candore G, Juhlin K, Manlik K, Thakrar B, Quarcoo N, Seabroke S, et al. Comparison of statistical signal detection methods within and across spontaneous reporting databases. *Drug Saf*. 2015;38(6):577–87.