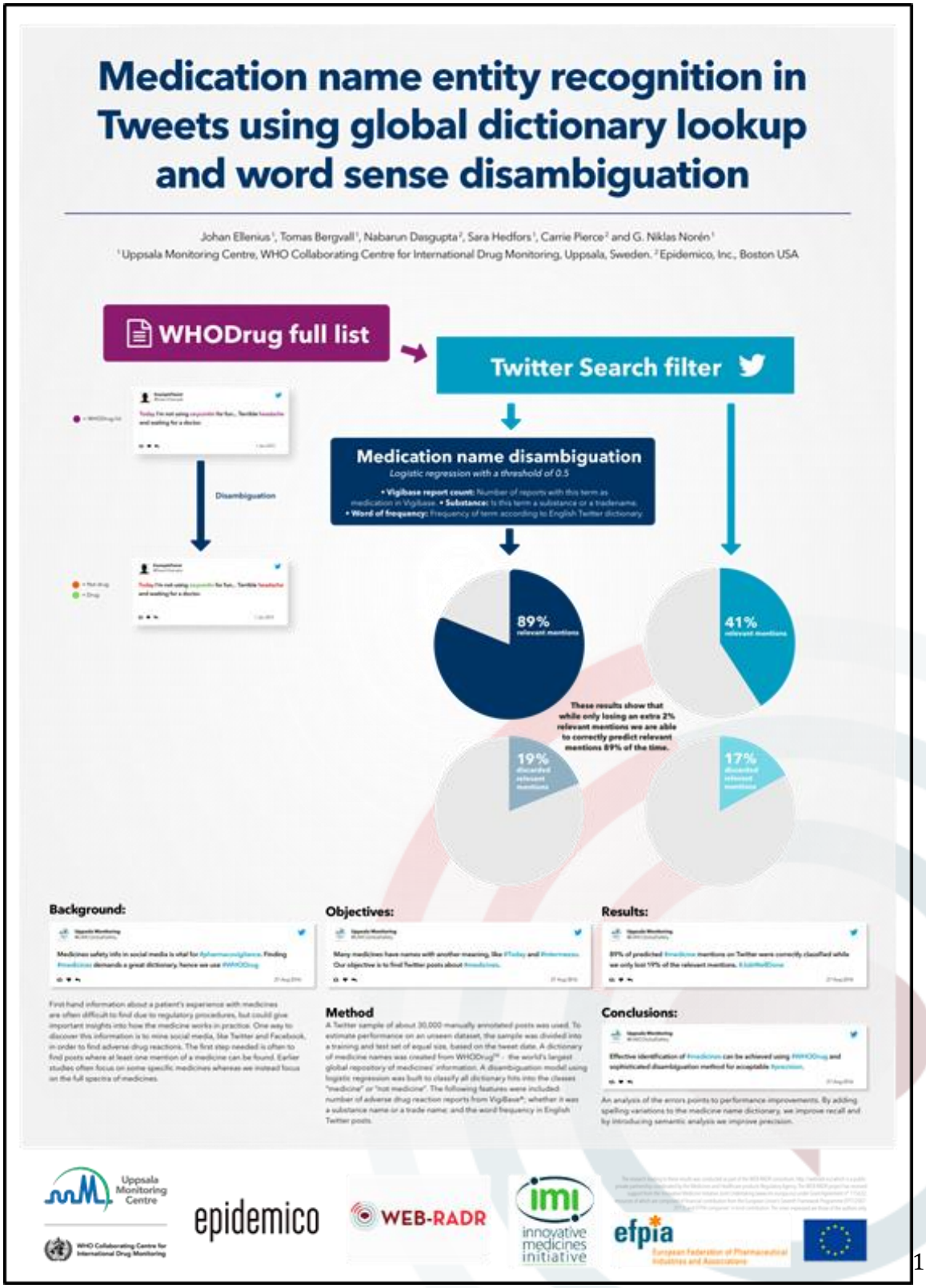
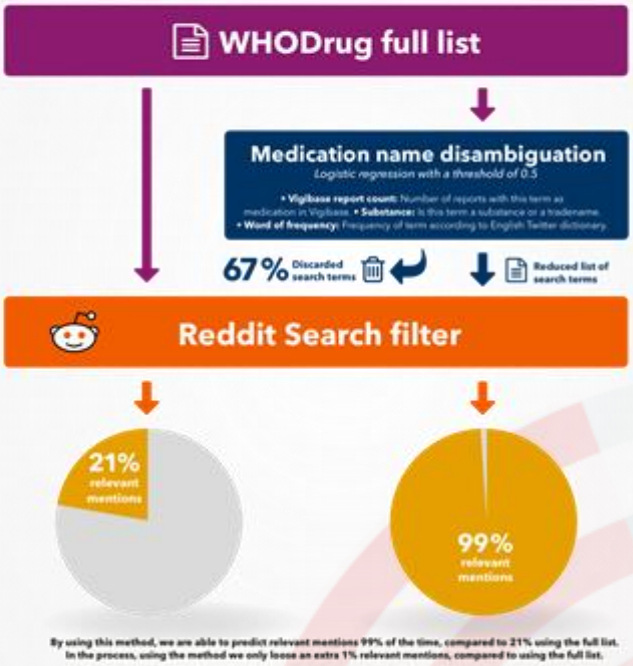
WP2B External Communication



Improving the yield of relevant data for pharmacovigilance analysis by reducing search term complexity

- a study on Reddit data

Sara Hedfors[1], Tomas Bergvall[1], Michael Gilbert[2], Carrie Pierce[2], Nabarun Dasgupta[2] and Johan Ellenius[1]
[1] Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Uppsala, Sweden. [2] Epidemico, Inc., Boston USA



By using this method, we are able to predict relevant mentions 99% of the time, compared to 21% using the full list. In the process, using the method we only loose an extra 1% relevant mentions, compared to using the full list.

Background

When monitoring social media, for example Reddit, you need **a list of search terms**. If the purpose is **pharmacovigilance**, the search term list may consist of **medicinal products.** However, some search terms will return mostly **non-relevant posts** due to **ambiguous meanings** (e.g. the product "Today"), whilst the aim is to obtain mostly relevant posts, i.e. those containing actual mentions of medicines.

Method

Five substances were investigated: levetiracetam, methylphenidate, terbinafine, zolpidem and insulin glargine. All the medical product names in the WHODrug™ dictionary containing these were extracted to a full list of search terms. With this list, a dataset of Reddit posts was extracted. We developed a **natural language processing** algorithm based on logistic regression. When applied to the list, the algorithm assigns a **probability prediction score** to each product name in the list, in order to identify terms that could be discarded.

Objectives

To present and evaluate a methodology to **improve cost efficiency** by **reducing the list** of search terms when retrieving posts, without significantly reducing the yield of **useful posts.**

Results

The algorithm predicts search term signal-to-noise ratio on a continuous scale, but the actual signal-to-noise ratio was found to be almost bisectional. The indication of a "mostly relevant" versus "mostly noisy" term is very well predicted by the algorithm, with just a few exceptions: if a threshold of 0.5 is chosen, 79 names are correctly classified (27 as mostly relevant and 52 as mostly noisy) and only 6 are erroneously classified. This corresponds to a precision of 0.96 and a recall of 0.84.

Selecting and using only product names as search terms with a probability prediction score exceeding 0.5 would yield the following benefits:

- Number of search terms decreased by 67%
- Number of extracted mentions decreased by 78%
- Proportion of relevant hits increased from 21% to 99%, losing only 1% relevant mentions

Conclusions

It is possible to substantially **increase the yield of relevant posts** by carefully selecting a reduced list of medication names used as search terms when retrieving social media data with **only a small loss of relevant posts.**

	Full list of search terms		Reduced list of search terms		
	# search terms	relevant mentions	# search terms	relevant mentions	relevant mentions lost
levetiracetam	12	7.1%	2	98.2%	1.4%
methylphenidate	19	90.1%	13	100%	0.3%
terbinafine	30	3.1%	4	100%	34.8%
zolpidem	22	5.9%	5	92.1%	0.4%
insulin glargine	2	100%	2	100%	0.0%
Total:	**85**	**21.4%**	**28**	**98.6%**	**0.9%**

Uppsala Monitoring Centre
WHO Collaborating Centre for International Drug Monitoring
epidemico
WEB-RADR
imi innovative medicines initiative
efpia European Federation of Pharmaceutical Industries and Associations

WEB-RADR

Performance of Disproportionality Analysis for Statistical Signal Detection in Social Media Data

Ola Caster[1,2], Magnus Lerch[3], Benoit Vroman[4], and John van Stekelenborg[5]

[1]Uppsala Monitoring Centre, Uppsala, Sweden. [2]Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden. [3]Lenolution GmbH, Berlin, Germany. [4]UCB Pharma, Braine l'Alleud, Belgium. [5]Janssen Research and Development, Horsham, PA, United States.

Background

Social media have emerged as a potentially useful complementary pharmacovigilance data source. While disproportionality analysis (DA) is the state of the art for statistical signal detection in spontaneous reporting data, its performance in social media data has not been systematically evaluated.

Objectives

To evaluate the performance of DA for statistical signal detection in social media data, and to compare that to spontaneous reporting data.

Method

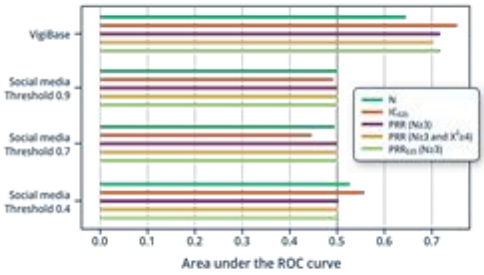
62 time-stamped US FDA label changes (positive controls) and 75 drug-event pairs without known association (negative controls) were used as reference[1]. About 2,025,000 posts dated between March 2012 and February 2015 (23% Facebook, 77% Twitter) on the 44 drugs from the reference set were combined with 3,710,000 background posts on other drugs. A trained classifier algorithm[2] identified posts suggestive of adverse events, at varying classification thresholds. For each threshold, monthly retrospective DA was performed, using IC_{025} (the lower IC 95% credibility limit) and three standard PRR variations. An analogous analysis was performed in VigiBase®, although with retrospective data collection extending back to the start of the database. Receiver operating characteristics (ROC) with respect to the reference set were computed as of December 2012, when all positive controls were yet unlabelled. Also, the retrospective time points at which the positive controls were statistically highlighted were computed and compared to the time-stamped label changes.

Conclusions

Disproportionality analysis on Facebook and Twitter data performed considerably worse than on global spontaneous reporting data, when benchmarked against historical label changes. This is likely due to limited prevalence or retrieval of the labelled events from social media, rather than the disproportionality methods themselves.

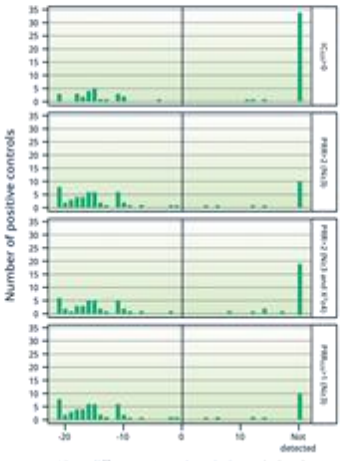
Overall performance results

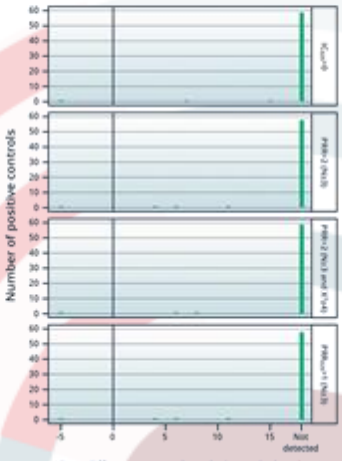
The below graph shows the performance of the four disproportionality analysis algorithms and simple report or post counting, for the various data sources investigated. Performance is measured as area under the respective ROC curves, which were generated by computing sensitivity and specificity of the various algorithms at all possible thresholds. An area of 0.5 is equivalent to random guessing. (N = number of reports / social media posts)



Timing of disproportionate reporting

These two graphs show at what time points the four disproportionality analysis algorithms highlighted the positive controls of the reference set. Times are presented as differences from the time-indexed label changes. The left graph is based on VigiBase, and the right is based on social media data using the recommended threshold 0.7. Not detected means that the threshold for disproportionate reporting was not met within the study period.





Disclosures

The research leading to these results was conducted as part of the WEB-RADR consortium, (http://web-radr.eu) which is a public private partnership coordinated by the Medicines and Healthcare products Regulatory Agency. The WEB-RADR project has received support from the Innovative Medicine Initiative Joint Undertaking (www.imi.europa.eu) under Grant Agreement n° 115632, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution. The views expressed are those of the authors only. Magnus Lerch provides scientific advice and support to Bayer Pharma AG within the WEB-RADR project.

References

1. Harpaz R, Odgers D, Gaskin G, DuMouchel W, Winnenburg R, Bodenreider O, Ripple A, Szarfman A, Sorbello A, Horvitz E, White RW, Shah NH. A Time-Indexed Reference Standard of Adverse Drug Reactions. Scientific Data, 2014; 1:140043.
2. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, Dasgupta N. Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter. Drug Safety, 2014; 37(5):343-350.

Uppsala Monitoring Centre

WHO Collaborating Centre for International Drug Monitoring

Uppsala Monitoring Centre (UMC)
Box 1051, SE-751 40 Uppsala, Sweden
Tel: +46 18-65 60 60 www.who-umc.org