SPECIAL ARTICLE

Recommendations for the Use of Social Media in Pharmacovigilance: Lessons from IMI WEB-RADR

John van Stekelenborg[1] · Johan Ellenius[2] · Simon Maskell[3,4] · Tomas Bergvall[2] · Ola Caster[2] · Nabarun Dasgupta[5] ·
Juergen Dietrich[6] · Sara Gama[7] · David Lewis[7,8] · Victoria Newbould[9] · Sabine Brosch[9] · Carrie E. Pierce[10] ·
Gregory Powell[11] · Alicia Ptaszyńska-Neophytou[12] · Antoni F. Z. Wiśniewski[13] · Phil Tregunno[12] · G. Niklas Norén[2] ·
Munir Pirmohamed[14,15]

Abstract

Over a period of 3 years, the European Union's Innovative Medicines Initiative WEB-RADR project has explored the value of social media (i.e., information exchanged through the internet, typically via online social networks) for identifying adverse events as well as for safety signal detection. Many patients and clinicians have taken to social media to discuss their positive and negative experiences of medications, creating a source of publicly available information that has the potential to provide insights into medicinal product safety concerns. The WEB-RADR project has developed a collaborative English language workspace for visualising and analysing social media data for a number of medicinal products. Further, novel text and data mining methods for social media analysis have been developed and evaluated. From this original research, several recommendations are presented with supporting rationale and consideration of the limitations. Recommendations for further research that extend beyond the scope of the current project are also presented.

Key Points

General social media, as exemplified by sample data from Facebook and Twitter, are not recommended for broad statistical signal detection.

Social media channels may provide a useful adjunct to pharmacovigilance activities in specific niche areas such as exposure during pregnancy and abuse/misuse of medicines.

Future enhancement of adverse event recognition algorithms may broaden the scope and utility of social media over time.

The views expressed in this paper are those of the authors only and not of their respective institution or company.

✉ John van Stekelenborg
   JVanstek@its.jnj.com

Extended author information available on the last page of the article

1 Introduction

The Innovative Medicines Initiative (IMI) WEB-RADR (Recognising Adverse Drug Reactions) consortium was a public–private partnership supported by the IMI Joint Undertaking (www.imi.europa.eu) under Grant Agreement no. 115632. Participating members were from European regulatory agencies, the pharmaceutical industry, academia, patient groups and other organisations with an interest in pharmacovigilance (PV). Full details of IMI WEB-RADR can be found at http://web-radr.eu.

The outputs from WEB-RADR arose from four work packages: two work packages undertook original research in social media and mobile application (app) technology; a third evaluated the scientific impact of the original research to determine where it had potential to add value to existing PV methodologies. A fourth work package addressed governance and policy, including personal data protection and ethical and societal considerations related to the use of mobile apps and social media for PV. This paper will focus on recommendations resulting from the research conducted using social media. Recommendations resulting from the work on the mobile apps is the subject of a separate publication [1].

The use of social media to monitor the safety profile of medicines is not a new concept [2–4]. Recent advances in information technology have, however, provided in-roads into addressing previous challenges posed by data volume and unstructured text, using tools like natural language processing and machine learning. At the same time, regulatory guidance for when exploitation of social media for safety purposes is appropriate or required has been limited by a lack of robust, evidence-driven methods to inform policy, and the implications of such guidance for patients, caregivers, industry, technology vendors and government. The WEB-RADR project set about in 2014 to address these specific issues from a European perspective, within a multinational consortium comprising key stakeholders [5].

Conventional methods of identifying safety concerns with licensed pharmaceutical products rely on patients and healthcare professionals to report suspected adverse drug reactions (ADRs) to regulatory agencies or industry intermediaries to collect and analyse. Significant under reporting of ADRs is well recognised despite decades-long efforts to address it [6]. Efforts to exploit other data sources to monitor pharmaceutical products for safety such as longitudinal observational health records have been met with modest success. Even in the early days of the internet, information on drugs has been exchanged [7] and patients and clinicians have increasingly taken to social media to discuss negative and positive experiences of medication use [8], in some cases creating a publicly available record that has the potential to provide new insights into a variety of medicinal product safety concerns.

Several online communities have evolved as robust discussion fora for health and medicine-related information exchange. Social media offers the opportunity to analyse patient perspectives of medication use that might not otherwise be communicated to healthcare professionals, as well as, at least in theory, the possibility to detect medicinal product safety concerns earlier than by traditional means. WEB-RADR intended to determine the possible added scientific and societal value of social media for safety surveillance and consider the consequential policy implications of this secondary data use.

The WEB-RADR consortium has considered the value and ethical and privacy concerns of public, patient-generated medicinal product safety information from social media platforms and patient discussion boards, including large datasets of posts from Facebook, Twitter, Reddit and online patient communities such as Inspire. As validation exercises conducted during the project revealed, each data source has unique properties with variable suitability for PV. The project has applied text and data mining algorithms for social media analysis and evaluated their utility in a PV setting within three themes: duplicate detection (record linkage), adverse event (AE) identification and signal detection (SD).

A significant challenge arising from the voluminous and constant stream of social media generated daily points to the need for a different approach to handling these data compared with a conservative but perhaps instinctive response to treat each social media post as an individual case safety report (ICSR). Visualisation and other analytic techniques are necessary to aggregate these data, although algorithmic tools are used on individual posts to prepare the data for analysis. To this end, one of the WEB-RADR work packages evaluated the principles of data display for dashboards and other outputs [9].

2 Summary of the Research and Recommendations

The main goal of the research presented here was to investigate the utility of social media for pharmacovigilance activities, investigate and improve analytical algorithms associated with the use of social media in PV, identify any areas of further improvement and provide recommendations in each of these areas.

The activities undertaken through IMI WEB-RADR and providing the evidence base for each set of recommendations are summarised here. For a full description of the studies, refer to the cited original publication or technical report. Implementation of these recommendations requires consideration of legislation and guidelines relevant to the particular locality in which they are implemented [10].

Separate working groups (Work Packages) in the WEB-RADR initiative undertook various studies.

Section 2.1 summarises the findings of the WEB-RADR analytic work packages and provides recommendations for the usage of social media in support of pharmacovigilance.

Sections 2.2 through 2.4 describe the three technical research areas underlying the recommendations, and provide additional technical recommendations specific to the scope of each of these areas.

Section 2.2 (signal detection) describes the investigation into the utility of social media in PV SD versus traditional sources.

Section 2.3 (adverse advent recognition) describes how to identify drug/AEs in social media.

Section 2.4 (duplicate detection) describes how to identify duplicate social media posts.

The recommendations in this paper are a reflection of the work done by each of the separate Work Packages.

2.1 Defining the Role of Social Media Data in Pharmacovigilance

The widespread use of social media by consumers and prescribers of medicines who share their experiences of those

Table 1 Proposed classification of social media data by potential value to pharmacovigilance

Area	Value proposition	Examples
Reporting and communication	Direct interaction between interested parties Increased awareness on part of the MAH, HA patient	Provides tools to report ADRs—company product websites, Medwatch, Yellow Card Sharing experiences and practices: communities of HCPs; communities of patients Two-way communication: risk communication; information sharing
Signal detection	Find rare events not often reported through spontaneous reporting to HAs and pharma companies Find medical side effects earlier than in other systems across a broad spectrum Alleviate underreporting known to occur in spontaneous systems	Primary signal detection tool alongside traditional (spontaneous) sources, across all products and events
Niche PV in pre-specified areas	Find new information in specific niche areas under-represented in current monitoring systems May be used as a primary tool for safety signal detection in certain pre-defined narrow areas (in contrast to broad-based safety monitoring across all products/events where social media are not value-added)	Exposure during pregnancy Abuse Misuse Low exposure, e.g., orphan drugs
Signal evaluation	Use for strengthening of hypotheses emerging from other systems Provide additional insight into safety issues identified through other means	Ad-hoc inspection of social media posts after a safety signal has been found in other sources
Quality of life	Find areas of patient and HCP concern that are not necessarily medically serious, but that have a significant impact on quality of life	Insomnia Stress Depressed mood

ADR adverse drug reaction, *HA* Health Authority, *HCP* health care provider, *MAH* Marketing Authorisation Holder, *PV* pharmacovigilance

medicines online make social media a potential source of PV data. An example is abuse potential with bupropion, a medicinal product approved for the treatment of depression and as an aid to smoking cessation. A study by Anderson et al. [11] showed that data from social media was more informative than spontaneous ADR reports, published literature and data from the Drug Abuse Warning Network regarding nonmedical use of non-controlled substances such as antidepressants, where data have been difficult to obtain. However, the value of such data needs to be established on a case-by-case basis; that is, each data source needs to be assessed in terms of product and safety topic 'richness'. Powell et al. [12] have described the variability of available data for a range of medicinal products in Facebook and Twitter. The role of social media in PV will ultimately be determined by the relative value it brings in uncovering new safety issues or new aspects of known safety issues.

When considering the role of any information source in PV, including social media, it can be helpful to broadly classify usage into five areas as shown in Table 1. The research presented aims to explore the potential uses of social media and highlight its applicability in each of these five areas.

Three sources of publicly available social media sources were included in the analyses presented in this paper:

1. Twitter ('short-form', i.e., limited to a number of characters per post): a fully public social media platform on which users can post ('tweet') short messages (< 280 characters) that can be responded to or repeated (reposted or 'retweeted').
2. Facebook ('long-form', i.e., not limited in length for a post): a social media platform that can be tailored in its accessibility settings (e.g., fully public, limited to certain individuals). Facebook allows sharing of long (i.e., unlimited in length) posts, photos and articles. Usage is primarily between groups of connected individuals. There are also focused discussion groups, company Facebook pages, and other (semi-)public Facebook sites.
3. Patient fora: social media sites dedicated to either a condition or disease, a drug or a patient population.
4. Reddit data were used for certain specific method development activities [13]. Reddit is a social media platform that contains so-called subreddits, each one of which is a public discussion forum focused on a specific topic.

Part of WEB-RADR's remit was to investigate the question of which social media platforms, if any, are of value in pharmacovigilance. The volume of the first two sources is inversely related to the average information content, with Twitter constituting the vast majority of posts, each

post containing a limited number of characters and consequently limited information.

The core SD analysis used a 'foreground' dataset of approximately 4.2 million tweets (65%) and Facebook (35%) posts collected during the period 1 March 2012 to 31 March 2015. In addition, 42,721 posts from 407 patient fora were collected, covering the same period. In addition, WEB-RADR developed a collaborative user interface to visualise and review these social media data, taken from a number of online sites over the period, and presented these data in a dashboard. The purpose of this project was to evaluate user experience and functionality of the interface [9].

WEB-RADR studied the value of social media (specifically Twitter, Facebook and some patient fora) as a primary broad-based statistical safety SD source (i.e. across all therapeutic areas, products and events) for uncovering new safety issues that would not have been detected by other means, or more timely discovery of safety issues compared with traditional data sources. The key conclusion from WEB-RADR is that, recognising the limitations of current technologies in identifying AEs in unstructured text, social media exemplified by Facebook and Twitter has very low value in the given context (see first recommendation in Table 2). Whilst this is the core result of the WEB-RADR analytics work package, there are a large number of health-related discussions on social media that might provide insights into safety issues for certain areas. Table 2 presents recommendations relating to the role of social media data in pharmacovigilance.

2.2 Signal Detection

One of the primary objectives of WEB-RADR was to determine whether the identification of safety signals in social media using aggregate statistical techniques could outperform or at least complement traditional aggregate methods used in spontaneous reporting systems. Specifically, the work presented here focused on the use of social media for aggregate statistical SD using spontaneous data as a comparator, namely VigiBase. Details of this study can be found in Caster et al. [15].

A summary of the investigation is shown graphically in Fig. 1.

In addition to the activities outlined in Fig. 1, a novel aggregate SD approach tailored toward social media data was developed and evaluated.

Two SD reference sets of positive and negative controls were used:

- Harpaz: the publicly available reference set by Harpaz et al. is based on US FDA labelling changes performed during the year 2013 [19]. The Harpaz reference contains 62 positive controls (i.e., labelling changes) on 55

medicinal products and 38 events. It also contains 75 negative controls.
- WEB-RADR SD reference set: a reference set not based on labelling changes, but on the concept of a 'validated safety signal', that is, a safety signal with some evidence suggestive of a causal medicinal product/event relationship beyond statistical disproportionality. The WEB-RADR SD reference set contains 200 positive controls and 5332 negative controls (Preferred Terms [PTs] that do not fall in any of the MedDRA® High Level Terms [HLTs] encompassing positive controls or listed/labelled PTs for the medicinal product) and covers 38 medicinal products.

Facebook, Twitter and patient fora posts were collected using a pre-existing Epidemico algorithm for classification described previously [12, 20, 21], mapping to medical products and MedDRA® event terms. These data were collected for a pre-specified set of 75 medicinal products that covered the two reference sets introduced above. Social media data collection was conducted for the period 1 March 2012 to 31 March 2015.

VigiBase reports were collected for the same period. These reports were taken from a frozen VigiBase version as of 16 October 2015 containing 14,897,935 reports in total.

The following four widely used statistical SD algorithms were used:

- $IC_{025} > 0$
- $PRR > 2$ and $N \geq 3$
- $PRR > 2$ and $N \geq 3$ and $\chi^2 \geq 4$
- $PRR_{025} > 1$ and $N \geq 3$

where IC_{025} is the lower bound of the 95% credibility interval of the Information Component, and PRR is the Proportional Reporting Ratio. Using receiver operating characteristics (ROC) curves to assess performance for both the Harpaz and WEB-RADR SD reference sets, the performance of all algorithms was poor in social media compared with VigiBase. For both reference sets, the area under the curve (AUC) for all methods was lower for social media than for VigiBase, indicating a uniformly lower performance in social media. The AUC ranged between 0.64 and 0.69 in VigiBase and was 0.55 or lower in all social media datasets using the WEB-RADR SD reference set. For Harpaz, the AUC values were 0.55–0.67 for VigiBase and 0.53 or lower for Twitter/Facebook.

Positive controls were analysed with respect to 'timeliness': the median time of first signal of disproportionate reporting (SDR) detection in both reference sets was generally after the index date for posts that were of 'better quality' (defined as posts with an indicator score threshold of 0.7 or

Table 2 Recommendations relating to the role of social media data in pharmacovigilance

Recommendation	Rationale
Social media should not be used as a source of ICSRs	With the exception of posts made by patients, carers and healthcare professionals on pharmaceutical company websites that make explicit mention of adverse events, the use of social media data for pharmacovigilance is secondary to the original intended use of these data [10, 14]. Although some posts may give detailed descriptions of an adverse event, the vast majority of posts lack the detail required for meaningful evaluation. Furthermore, large volumes of generally poor quality, non-informative data from social media should not be used to generate ICSRs since this has the potential to negatively impact signal detection systems [15]
Facebook and Twitter are currently not worthwhile to employ for the purpose of broad-ranging statistical signal detection at the expense of other pharmacovigilance activities	Applying disproportionality-based signal detection algorithms to automatically annotated Twitter/Facebook posts did not result in any predictive ability against two reference sets of signals and non-signals, in contrast to applying disproportionality analysis to VigiBase[a] cases. In addition, neither the first detected Twitter or Facebook posts nor the first occurrence of disproportionality in these sources would precede the actual time point of signalling, whereas in VigiBase this was more frequent, thereby negating any timing advantage of social media. This same lack of predictive ability was encountered with a relatively small sample of patient forum posts [15]
Future research should explore the value of social media as a source of information for additional cases in signal refinement/evaluation of ADRs that may significantly affect a patient's quality of life	Approximately 12% of posts inspected in WEB-RADR contained information relevant to quality-of-life issues, e.g. lack of sleep, anxiety etc. [15]. This was an average across 38 medical products; however, further analyses (unpublished) indicate that drugs in the neuropsychiatric area have much higher proportions of mentions with quality-of-life issues
If social media is considered for use in pharmacovigilance, it is recommended that a prior assessment of the absolute and relative number of available posts related to the drug and/or event of interest in different online sources is made in relation to its intended use	There is substantial variation across drugs and adverse event terms in the amount of information in social media as well as substantial variation across different social media sources. Of the 38 medicinal products included in the WEB-RADR signal detection reference set, the range of substance mentions was from five (ranibizumab) to approximately 24,000 (methylphenidate) over a 3-year period (1 March 2012 to 31 March 2015)—see Fig. 2
	Within the data collected prospectively for WEB-RADR (acquired from September 2014 through September 2017[b]), products with orphan or oncology-related indications were more likely to have higher volumes of posts describing potential AEs in patient fora than in Twitter (ruxolitinib had 3 × more posts describing potential AEs in fora, nilotinib had 8 ×, tobramycin 70 × and anastrozole 85 ×). Products with psychiatric indications were more likely to have a higher volume of posts in general, as well as a higher volume of posts describing potential AEs and mentions in Twitter than in patient fora (methylphenidate – 1.5 × more posts describing AEs in Twitter, zolpidem 7 ×) [16]
Further research should be carried out to determine whether there is value in social media data for niche areas of pharmacovigilance	WEB-RADR has demonstrated that there are niche areas of pharmacovigilance where social media data are more plentiful [17, 18] and can complement more traditional sources. For example, there is significant discussion about drug use in pregnancy [34] on social media to suggest that a combination of spontaneous reports and social media is likely to result in improved signal detection. However, the performance of this combined spontaneous/social media approach in specific areas is yet to be demonstrated as value-added relative to spontaneous reporting alone. In order to investigate this relative performance, additional work in algorithms and representative reference sets is needed
Consider using a predictive algorithm to identify and eliminate from the search query any medicinal product names with high levels of ambiguity to optimise time efficiency and, where applicable, cost effectiveness	The study by Hedfors et al. [13] showed that this could decrease the number of search terms by 67% and the number of extracted social media posts by 78%, with an associated increase in precision from 21.4% to 98.6% at a loss of only 0.9% of all relevant social media posts

ADRs adverse drug reactions, *AEs* adverse events, *ICSRs* individual case safety reports

[a]VigiBase is the World Health Organisation's (WHO) global database of ICSRs maintained by the Uppsala Monitoring Centre, Uppsala, Sweden. It is the largest database of its kind in the world, with over 19 million reports of suspected adverse effects of medicines submitted since 1968 by member countries of the WHO Programme for International Drug Monitoring

[b]In fact, data collection continued until December 2017; however, only data through September 2017 were included in the final report

higher), again showing poor performance of social media relative to VigiBase.

To evaluate this finding further, the content of 631 posts corresponding to 25 positive controls from the WEB-RADR SD reference set was assessed manually. Only 250 (39.6%) of the posts contained the correct medicinal product and medical event (ME) and the ME was an actual adverse experience (as opposed to a medical history term, for example). In only 33 of these 250 posts was there any mention of risk factors such as lifestyle, medical history, comorbidity,
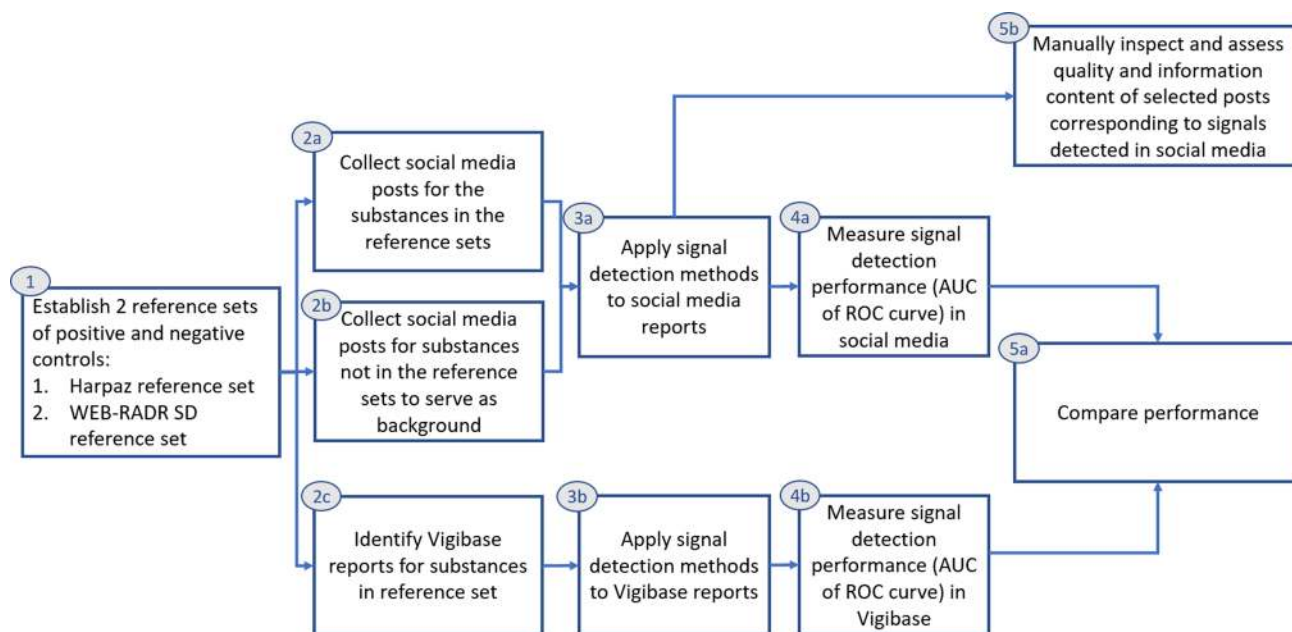
Fig. 1 Conceptual overview of the investigation of the utility of social media in safety signal detection. (*AUC* area under the curve, *ROC* receiver operating characteristics, *SD* signal detection)

indication or co-medication; and, in the opinion of the assessor, the contents of the posts would have strengthened the signal in three (12%) of the 25 signals, and in one of the signals the posts would actually have weakened the evidence.

Overall, the results provide very little support for the use of Twitter and Facebook as a broad-based, standalone data source for statistical SD in PV. Using two complementary SD reference sets, one containing validated safety signals and the other label changes, standard disproportionality analysis did not yield any predictive ability in a large dataset of combined Twitter and Facebook posts. Signal detection in VigiBase outperformed Twitter and Facebook for all algorithms. Importantly, there was no timing advantage observed either at the post level or SDR level for Facebook/ Twitter and patient fora.

The fact that Facebook and Twitter posts underlying 25 early signals of disproportionality only contained the correct medicinal product and the correct event as an adverse experience 40% of the time also diminished the value of first-line SD with social media, or signal strengthening for that matter. For most medicinal products, there simply does not seem to be much activity in social media. Future work should therefore focus on specific medicinal products, specific areas of PV interest and specific online patient groups or networks (taking into account access restrictions).

2.2.1 Exploratory Study into Alternative Approaches to SD

A novel SD algorithm based on Track Before Detect (TkBD) [22] was investigated for its applicability to social media. The new algorithm, named Signal Before Detect (SbD), considers the probability that each social media post is an actual AE (and conversely, a corresponding probability that it is not an actual AE).

The SbD technique employs methodology to allow use of all available posts, weighted according to their probabilities of being a true AE, here estimated via the indicator score. The method generates simulated datasets using a Monte Carlo-like approach described in the paper by Rutten et al. [22], albeit adapted to consider social media data. The $IC_{025} > 0$ algorithm was applied to the simulated datasets from Twitter/Facebook and evaluated against the WEB-RADR SD reference set. The results were no better than those achieved with fixed indicator score thresholds [23]. A possible reason for this lack of improved predictive ability lies in the lack of availability of a key component in a re-sampling technique such as SbD, namely the availability of the underlying probability distributions. In the context of SD, an AE probability distribution is required that should provide the probability for a given post with a given indicator score to contain an actual AE. While this distribution was established for the overall set of AEs, what is needed is a distribution for each individual AE, as there is large variation in the precision/recall of the AE recognition algorithm per AE. As these individual AE probability distributions

Table 3 Recommendations relating to the use of social media for signal detection

Recommendation	Observations
When evaluating signal detection algorithms for social media data, complement overall performance analyses (e.g. receiver operating characteristics) with manual post-level assessment as a sanity check: if a large proportion of the automatically identified posts do not contain the medicinal product or adverse event term indicated, overall results must be interpreted with considerable caution	During manual inspection of post text corresponding to a social media signal, only 39.6% of the posts contained the drug and medical event of interest as an actual adverse experience. In the subset of posts with indicator score of 0.7 or above, the corresponding result was 67.3% (72 of 107 posts)
In evaluation of signal detection methods, proprietary reference sets should be avoided if possible	Practically, working with our WEB-RADR SD reference set has been very cumbersome since all data extraction had to be performed locally at several different sites by those authorised to access the de-anonymised controls. Further, such a reference set cannot be critically inspected or re-used outside the specific study where it was used. Finally, certain types of analyses become impossible to perform, such as aggregation based on characteristics of the medicines or adverse event terms
If setting up a safety surveillance system based on social media today, it is more important to first improve and calibrate adverse event recognition than the algorithms for statistical signal detection	We have generally seen small differences between different algorithms and in our exploratory study, a more advance method like SbD provided no added benefit [15, 23]. No signal detection algorithm can extract information unless the data it depends on are of adequate quality and are well calibrated. In both the Caster and Dietrich studies, medical event recognition was a significant hurdle [15] [Dietrich *submitted 2019*]

SbD 'Signal Before Detect' algorithm, *SD* signal detection

were not available for the project, the main assumption for using the SbD approach was not met.

This is surprising since TkBD has previously been shown to offer improved performance across a wider range of diverse applications, such as radar, sonar and cameras. TkBD achieves this improvement by capitalising upon accurate probabilistic models for raw sensor outputs where accurate variants of these models have been extensively calibrated enabling modelling of the specific data applied to TkBD. The same is not true for the input to SbD; the mapping from classifier indicator score to the probability that a given post contains an actual AE is not yet well characterised, is poorly calibrated and lacks accuracy. While a probability distribution was established for the overall set of AEs, since there is large variation in the precision/recall of the AE recognition algorithm per AE, a distribution per individual AE is needed. As these individual AE probability distributions were not available at the time of the WEB-RADR studies, SbD has, as yet, been unable to use accurate, calibrated models and thus unable to demonstrate an improvement in SD performance. Furthermore, a lack of AE-specific calibration will ultimately degrade the performance of any SD algorithm that uses data obtained from a global (i.e., AE-agnostic) threshold based on indicator score (i.e., those considered in WEB-RADR's analysis of the utility of SD using social media). Table 3 presents recommendations relating to the use of social media for SD.

2.3 Adverse Event Recognition

The AE recognition in social media studies aimed to identify medicinal product and AE pairs in single posts where the medicinal product is expressed as a substance or trade name, this being a necessary step prior to statistical SD methods such as disproportionality analysis [Gattepaille LM, Vidlin SH, Bergvall T et al. Prospective evaluation of adverse event recognition systems in Twitter: results from the Web-RADR project. To be submitted].

A process for AE recognition in Twitter using a natural language processing workflow has been developed as part of WEB-RADR [24]. The predictive models used in the workflow were developed using cross-validation on a data set of tweets collected and gold standard classified by Epidemico. Evaluation of classification performance was done using a separate AE recognition reference set, a key deliverable of the WEB-RADR project, containing a total of 880 unique product names, each related to one of the substances zolpidem, insulin glargine, levetiracetam, methylphenidate, sorafenib and terbinafine [Dietrich 2019, submitted]. In all, 57,473 tweets were included, with 1396 medicinal product–AE pairs. The AE recognition reference set was then annotated by a team of certified MedDRA® coders that were not involved in the annotation of the data set used for training, thus making it appropriate to use for evaluating the transferability of the workflow to new contexts.

The first step of the processing workflow was a relevance filter based on Epidemico's indicator score that aims at

finding posts with a resemblance to AE posts. In the study by Powell et al. [12], a threshold of 0.7 was used to select posts, having a recall of 92% and precision of 50%. In this first step of the AE recognition workflow developed by WEB-RADR, the same threshold of 0.7 was used. The filter was very performant on the hold-out sample of the Epidemico dataset, discarding only 4.7% of the 3547 posts containing at least one drug–AE combination. However, a sizeable drop in performance was observed when applied to the WEB-RADR reference dataset: the filter discarded 37% of AE posts.

The second step performed Named Entity Recognition (NER) of medicinal product names in the subset of posts from the previous step. The method used a dictionary lookup of product names extracted from WHODrug Global (https://www.who-umc.org/whodrug/whodrug-portfolio/whodrug-global/, last accessed Jan 1 2015.). Product names with a high level of ambiguity were pruned out of the lookup to reduce noise, following a previously published method [13]. The product NER was able to recall 88% of the product annotations involved in the 3547 medicinal product–AE pairs found in the hold-out sample of the Epidemico dataset. Comparatively, the product NER was able to recall 90% of the product annotations involved in the 1396 medicinal product–AE pairs found in the WEB-RADR reference dataset.

The third step performed NER of medical events, a broader construct that includes AEs, indications and medical histories. Medical event recognition is a challenging problem since patients describe their feelings and experiences in diverse ways. It was built as a mix of three different components: a dictionary lookup based on MedDRA® lowest level terms, a dictionary lookup based on reported reactions extracted verbatim from VigiBase and, finally, a machine learning component where the words in the tweets are used as input for 124 independent logistic regression models, each trained to recognise a single distinct MedDRA® PT. The medical event NER was recognised as the main performance bottleneck of the workflow. Indeed, although it could recall 74% of the medical events annotations that were AEs in the hold-out sample of the Epidemico dataset, only 46% of the events were recalled in the WEB-RADR reference dataset. Using MedDRA® lowest level terms as the sole resource for the medical event NER gave a recall of 12% in both hold-out sample and reference dataset. Adding the expressions extracted from VigiBase led to a recall of 35% and 33% (almost triple the recall without), and further adding the machine learning-based NER algorithm gave final recalls of 74% and 46% in the two datasets, respectively.

In the fourth and last computational step of the workflow, a logistic regression classifier was developed to classify all combinations of medicinal products and medical events, correctly identified and coded by the NER modules, as well as involved in posts having passed the indicator score filter, as representing AE relationships or not. The classifier used several kinds of features as input: statistical, syntactic and semantic. On the hold-out sample of the Epidemico dataset, it was able to recall 81% of the medicinal product–AE pairs that survived thus far in the workflow, while it could recall 63% of the medicinal product–AE pairs of the WEB-RADR reference set.

The overall performance of the complete workflow for the recognition of the 1396 medicinal product–AE pairs in the WEB-RADR reference set gave a recall and precision equal to 20% and 38%, respectively, for an $F1$ score of 0.26. This corresponded to a drastic drop in performance when comparing with the results on the hold-out sample of the Epidemico dataset, which gave a precision of 53% and a recall of 52% ($F1$ score 0.52). In comparison, the drop in performance going from the training sample of the Epidemico dataset to the hold-out sample was much more moderate (performance on the training sample was 61% precision, 58% recall and 0.60 $F1$ score). Therefore, together with the drop in performance observed for the indicator score filter compared with published performance (see above), these results highlight the necessity of external validation of AE recognition workflows, as performance on hold-out samples might give poor estimates of performance on new independent data.

A large variation in performance was also observed across the different MedDRA® PTs observed in medicinal product–AE pairs. Among the top 10 most annotated PTs in the WEB-RADR reference set, F1 scores varied from 0 (Social problem, Adverse event) to 0.53 (Hallucination). Most PTs saw a drop in performance when going from the hold-out sample to the reference set, illustrating that the difference in performance cannot solely be explained by differences in the nature of the PTs present in the datasets. Table 4 presents recommendations in relation to AE recognition in social media data.

2.4 Duplicate Detection

Effective use of social media for safety SD is challenging and requires reliable data. Among the different factors that can affect data quality, duplication of posts needs to be addressed for several reasons. Most analyses rely on the assumption that no two records refer to the same event. The presence of duplicate records violates this assumption and may lead to an overestimation of the amount of evidence in support of a particular association. Furthermore, to the extent that duplication affects events differently (e.g., the same piece of news can be relayed in many posts while a personal experience might be described only once), bias may also be introduced. Further, social media monitoring has the potential to produce large volumes of data that can be a burden in storage, computation and review.

Table 4 Recommendations relating to adverse event recognition in social media data

Recommendation	Observations
For methods developed for AE recognition in social media, evaluate its performance on a standard reference data set such as that produced by WEB-RADR, to facilitate comparison of methods	We compared the classification performance of the NLP workflow for medicinal product–AE *pair* recognition, when evaluated on an independent sample from the same dataset that was used for training of the predictive models in the workflow with the performance on the AE recognition reference set. Recall dropped from 52% to 20%, and precision from 53% to 38%. Evaluation of a previously published method for detection of AE posts also displayed a drop in performance: from 50% precision and 92% recall (0.65 F1 score) in the publication to 37% precision and 63% recall (0.46 F1 score) on the WEB-RADR reference dataset. This illustrates the risk of overestimating the classification performance of a method if an independent dataset from another context is not used in the evaluation
Consider the use of machine learning technology to support the recognition of social media data relevant for pharmacovigilance	Less than 2% of tweets assessed in the development of the AE recognition reference set contain AE terms [Dietrich 2019, submitted]. A large proportion of irrelevant data will exist in any social media dataset. As such, employing automated processes may enable AE recognition whilst reducing the effort required for manual review
Human curation should be used in conjunction with automated processes aimed at identifying potential AEs from social media with methods available today	The NLP workflows for medicinal product–AE recognition and coding were evaluated to have a precision equal to 38%. This means that the majority of automatically recognised medicinal product–AE pairs are incorrectly classified. Human curation has the potential to detect and discard such pairs and thereby increase the precision. The content of social media posts underlying signals of disproportionate reporting (SDRs) in the signal detection study was inspected and found to be severely lacking in content and interpretability. In fact, of the posts inspected, only 40% of posts contained the correct drug and the correct event as an adverse experience, pointing to a significant issue with ADR recognition [15]
If available, use existing mappings between verbatim text and MedDRA® terms from spontaneous reporting systems to improve sensitivity in medical event recognition for social media	Our study showed that the inclusion of historical mappings from VigiBase verbatims to a dictionary of MedDRA® LLTs almost tripled the number of captured AEs. Generalisability beyond VigiBase as a source of mappings and Twitter as the domain of application is unknown
Consider incorporating information on medicinal product indications in automated AE recognition, thereby reducing the likelihood of falsely categorising an indication as an AE	In the AE recognition reference set, 18.5% of patients mention indications for use of a medication in conjunction with product names and AEs [Dietrich 2019 submitted]. In addition, patients may describe symptoms of their underlying conditions and AEs in the same post, making it difficult for automated processes to determine which medical conditions or symptoms are AEs versus related to a product's indication. Absolute removal of indication-related posts may not be beneficial or result in more accurate automated coding; for example, in a post where a patient mentions that a medical product aggravated the condition that the medicine is meant to be treating

ADRs adverse drug reactions, *AEs* adverse events, *LLTs* Lowest Level Terms, *NLP* natural language processing

Some types of duplicates can be difficult to identify: a user can refer to a single event in multiple posts; multiple users can refer to the same event in different ways, even across multiple platforms. Conversely, textually close descriptions are not necessarily duplicates but may represent distinct events of a similar nature.

The basis for duplicate detection in WEB-RADR was vigiMatch, a method that has been described extensively elsewhere [25, 26]. For WEB-RADR, we implemented vigiMatch to screen for duplicates based on comparing the textual content between posts as 'bags of words'.

For training and evaluation of the method, Twitter posts were selected for 23 active substances over 2 months, commencing 27 September 2016. The search terms used for data collection were based on a list of trade names expanded from WHODrug and were selected to give broad coverage of different medicinal products and vaccines. To reduce the number of irrelevant posts, trade names with a higher chance of being used in contexts other than referring to a medicinal product were excluded using the aforementioned method [27]. In total, 13,820 posts were collected, with substantial variation between substances (e.g. 23% of the posts were for the human papillomavirus (HPV) vaccine).

Table 5 Recommendations relating to duplicate record detection in social media data

Recommendation	Observations
Duplicate detection should be performed in preparing social media data for use in pharmacovigilance	Having first eliminated simple retweets etc., our study found 17% of the remaining posts to be suspected duplicates, with an algorithm that has an estimated precision of 99% [28]. In our signal detection study, several of the inspected series of posts contained large proportions of duplicates [15]
Probabilistic record linkage should be considered as a complement or alternative to rule-based methods for duplicate detection in social media data	Our study found 9% suspected duplicates in a set of Twitter posts that had already been deduplicated using a method based on rules and Bloom filters. A lower proportion of additional suspected duplicates were identified for posts related to adverse events (1.6%) [28]
Training data for duplicate detection in social media should be enriched with suspected duplicates ensuring that the method of enrichment is accounted for in the training and evaluation of the duplicate detection method; for example, through active learning	Our study showed that it was feasible to use active learning in training vigiMatch for duplicate detection in Twitter. Only 0.008% of all the possible pairs of tweets in our data were suspected duplicates, so a straight sample would include mostly non-duplicates. If training data are enriched with suspected duplicates and algorithms are trained and evaluated without considering the method of enrichment, then the method and their estimated performance will not generalise to the real-world setting
Future research should compare different approaches to improve computational efficiency such as blocking and locality-sensitive hashing	Computational efficiency is of great importance in duplicate detection and a comparison between different approaches was out of scope for the study at hand. In our study, a simple blocking scheme reduced the number of pairwise comparisons by 22% [28]

The collected tweets had not yet been annotated for duplicates and manual examination of a random selection of pairs would have yielded a very low proportion of true duplicates. Therefore, we applied active learning, a semi-supervised procedure in which the algorithm to be trained is used to select samples that are submitted for manual annotation. We then iteratively learned from the annotated samples. During each iteration of the active learning procedure, around 200 pairs were sampled and annotated from ten different ranges of vigiMatch scores. This enrichment of the training data with true duplicates was accounted for in the parameter estimates by creating a modified set of training data that included multiple copies of the annotated pairs in such proportions as to match the relative number of true duplicates in each range of vigiMatch scores.

To reduce the number of pairs to consider in applying vigiMatch, we applied a blocking scheme where only pairs with at least one word in common were considered. In this context, blocking refers to the process of removing posts from the analysis that do not conform to one or more predefined criteria. Blocking reduced the number of comparisons from around 95 million pairs to 74 million (a more substantial reduction might have resulted if blocking had been restricted to reasonably rare words).

After five iterations, enough to obtain stable parameters for vigiMatch and to compute a threshold for duplicate classification, the entire sample was screened and 17% of the 13,820 collected tweets were classified as suspected duplicates. Using the manually annotated pairs we could estimate the precision to 0.9999 and recall to 0.035 for the entire dataset. The precision obtained is very high, at the expense of the recall (i.e., nearly all the retrieved posts are duplicates). However, the proportion of the retrieved duplicates from those present in the dataset is very low. This is explained by a deliberate conservative standpoint: we only allow posts to be filtered out when there is strong evidence in favour of duplication, in order to not lose any relevant posts.

To examine the added value of probabilistic duplicate detection over existing methods, vigiMatch was also used on previously de-duplicated data provided by another WEB-RADR partner, Epidemico [20]. The entire dataset was analysed using vigiMatch with the threshold estimated from the training data. In total, about 14,000 out of 155,000 posts (9%) were classified as duplicates by our algorithm. We reviewed the 261 pairs that had at least one post annotated as a likely ADR and found eight false positive and 253 true positive duplicate pairs. Some of those false positives included responses to tweets where the user had quoted the original tweet and only added a short text such as "me too". Recommendations relating to the detection of duplicate records in social media data are presented in Table 5.

3 Discussion

The recommendations are presented with the intention of informing PV professionals, particularly those with an interest in the development of research methods and digital technologies, and are based directly on the outcomes of the research conducted under the auspices of IMI WEB-RADR.
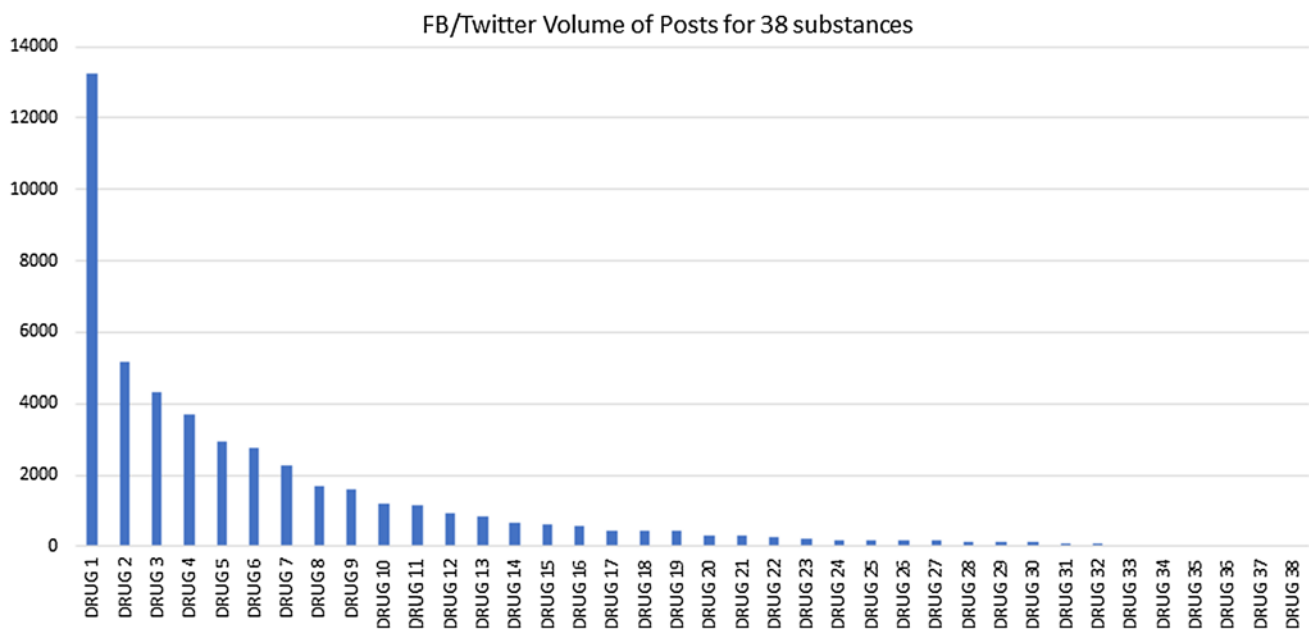
Fig. 2 Number of WEB-RADR substance mentions in Twitter/Facebook (FB) at an indicator score threshold of 0.7. Figure drawn using data from Caster et al. [15]

They should not be considered a comprehensive treatise on pharmacovigilance and social media, and they should be considered by readers within the context of the entire corpus of research and guidance documents that exist outside of IMI WEB-RADR.

Careful consideration should be given to the generalisability of the recommendations in circumstances substantially different from those described in the original research. The aim is to enable PV experts to start considering the impact of leveraging social media as part of a broader PV strategy. There is still a substantial amount of research that needs to be completed before the true strengths and weaknesses of this emerging data source can be adequately understood.

Based on the findings from the WEB-RADR project, there is little evidence that statistical safety SD in Twitter and Facebook adds any value beyond currently employed spontaneous data sources. Specifically, using these social media as a broad-based stand-alone data source for statistical SD in pharmacovigilance yielded no predictive ability in two complementary reference sets of signals and label changes. This is in contrast to VigiBase where significant predictive performance was established.

There are several possible reasons for this observed lack of predictive ability. Firstly, Fig. 2 suggests that there is a lack of discussion activity in social media for a large proportion of the medicinal products used in the WEB-RADR study and it is reasonable to assume this is generalisable to a broader range of products. Secondly, the SD algorithms investigated critically rely on the quality of the AE recognition algorithms applied to social media posts. If these algorithms are not sensitive enough, or if they surface high numbers of false positives, the downstream statistical algorithms will break down. Thirdly, from a methodological perspective, the SD reference sets may be somewhat biased to favour spontaneous sources. Specifically, to that point, it is possible that some of the signals used as positive controls were originally detected by the company in their spontaneous reporting system. Fourthly, the SD methods applied were of an aggregate nature (essentially counting product/event pairs), tailored toward spontaneous data sources. It is conceivable that other methods focussing on the very different structure and content of social media could be developed. Nevertheless, the dearth of social media posts of sufficient quality for many products makes it difficult to see how even radically different approaches could result in better performance for broad-based SD.

The estimated performance of the AE recognition algorithm using the WEB-RADR AE recognition reference set is substantially below previously published results [Dietrich 2019, submitted]. For example, the legacy AE recognition algorithm on which the WEB-RADR SD study relied had a recall of 86% and a precision of 72% in a previously published evaluation [21] compared with the recall of 32% and precision of 20% in the evaluation against the WEB-RADR AE recognition reference set. One contributing factor to such discrepancies is if performance is evaluated based on whether the algorithm successfully identifies any posts related to AEs or whether it also successfully annotates the putative medical product and AE. The former is easier and better performance would be anticipated; the WEB-RADR

AE recognition algorithm had a recall of 0.39 and a precision of 0.70 for the former and a recall of 0.20 and a precision of 0.38 for the latter.

Another source of variability is the differences in the enrichment of reference sets by positive controls and whether this is accounted for in the performance evaluation. Straight random samples of posts may not be feasible in constructing reference sets since the vast majority of posts do not relate to AEs. In the development of the WEB-RADR AE recognition reference set, < 2% of the posts matching one of our medicinal product search terms conveyed information related to personal experiences of AEs. However, if reference sets are enriched with positive controls, then nominal performance estimates will not generalise to the setting in which the methods will be applied. Specifically, the nominal precision of a predictive algorithm is heavily influenced by the prevalence of positive controls in the reference set—in effect it will be the precision expected of an algorithm with no predictive ability and thus the lower bound estimate for nominal performance. If the enrichment method employed results in the elimination of positive controls that are more difficult to detect, recall will be overestimated.

Inadequate protection against overfitting of algorithms to training data is another source of error. Cross-validation or evaluation against held-out data should be the starting point, but performance still may not generalise to new settings if, for example, the reference set is focused on a narrow set of drugs and MEs. It should be noted that the legacy AE recognition algorithm used in WEB-RADR covered only a limited set of MedDRA® PTs (based on colloquial phrases used by social media users), thereby restricting the types of signals that could be detected. As a further consideration, it is important that duplicate posts do not exist across training and validation data, otherwise optimistic performance estimates and overfitting of the predictive models will result. The latter may be avoided through duplicate detection and removal or by selecting validation data from different time periods than the training data.

Previous research has demonstrated that human curation can effectively eliminate noise and improve the precision of social media processing chains by removing false positives, ensuring correct coding of medical products and events in particularly nuanced patient narratives, and continuously improve automated classifiers by serving as a feedback loop and adding to training sets [12, 21].

Social media is a rapidly changing data source with an ever evolving language. New words, flexible definitions, abbreviations and slang terms [29] and the subjectivity of what is said versus what is actually meant (e.g. "the price of a new car almost gave me a heart attack" [30]) also change. There may be instances where topic-specific dictionaries are required to achieve an acceptable level of performance. In

a study that looked at abuse potential with bupropion [11], a custom dictionary had to be developed to reflect medicinal products of interest (e.g., 'vikes' = Vicodin), events of interest (e.g., 'trip' = altered state of consciousness), route of administration (e.g., 'nose candy' = snorting) and how the results are to be interpreted based on current regulatory guidelines (e.g., how to define drug abuse and misuse). Due to the evolving nature of language in social media, it is unlikely a static, automated approach will have consistent performance over time.

At the outset of the WEB-RADR project, it was anticipated that data from social media could provide promise in a number of areas important for PV [18]. It was recognised that there is very limited value in creating ICSRs, as the data quality and data privacy restrictions limit the value of these posts. WEB-RADR has shown that there is little or no activity within social media for certain medicinal products [Dietrich 2019, submitted]; it would be futile to conduct SD for those products. Admittedly, WEB-RADR's focus on Twitter and Facebook limits the generalisability of this conclusion and research conducted during the WEB-RADR project and elsewhere using patient fora have identified some data-rich areas. Although the WEB-RADR study of Caster et al. [15] did not show any benefit in SD performance using data from patient fora, evidence has been generated concerning methylphenidate and misuse [31, 32], and elsewhere research has shown that social media data can provide real-world use data and outcomes to inform safety decision making [11, 33, 34].

The current approach to SD is to use traditional data sources, such as spontaneous AE reports and published literature, independently. Based on the hypothesis that utilising and jointly analysing multiple data sources may lead to improved SD [35], WEB-RADR planned to assess if social media data can be used to improve SD performance (e.g., positive predictive value, time to detection). However, in view of the complete lack of predictive ability for Facebook/Twitter in our SD study, there was no prospect that it would inform such an ensemble method. If the social media data sources used for WEB-RADR improve in quality, or if other social media data sources are determined to be more appropriate (e.g., PatientsLikeMe, Inspire and HealthUnlocked) and/or methods can be improved (e.g., better AE recognition), this should be revisited. Additionally, it may be worth exploring the use of social media to enrich traditional SD activities based on the niche areas previously discussed (e.g. drug use in pregnancy, abuse/misuse).

Over the past several years, the PV focus has shifted from purely detecting and evaluating AEs to a more holistic benefit–risk evaluation. Although the results above demonstrate the lack of social media performance in identifying potential risks, there may be value in social media informing the benefits of medicinal products. Research presented in an earlier

paper by Powell et al. [12] highlighted that approximately 25% of social media posts that discuss a medicinal product will contain information relating to the benefits of its use. These include a continuum of topics ranging from the degree to which the product worked, to the duration of the benefit, to comparing benefits with other treatment options, to describing how its use has improved the patients' quality of life and/or average daily living.

The use of social media as a source of medical insights is still in its infancy. In order to fully appreciate the breadth and depth of what social media may offer, the strengths and weaknesses of each data source, how to maximise operational efficiency and to ensure appropriate governance and oversight, a coordinated effort across a range of stakeholders is warranted. For example, the WEB-RADR project demonstrated the challenges in AE recognition. If done independently, the time and effort required to create mappings from a variety of vernacular and colloquial language to the approximately 70,000 MedDRA® lowest level terms would be substantial. If these activities were coordinated across a range of stakeholders, it would reduce the burden on a single entity, shorten the time required to complete the task and would offer transparency into the process. The collaboration framework should focus on advancing the science around extracting medical insights from social media rather than generating profits, ideally using an 'honest broker'. Ultimately, a coordinated effort will lead to a more rapid maturation of this area as well as facilitate adoption and acceptance.

4 Conclusions

Over a period of 3 years, several IMI WEB-RADR work packages have addressed key research questions relevant to the use of social media for pharmacovigilance. WEB-RADR does not recommend the use of general social media, as exemplified by Facebook and Twitter, for broad statistical SD. However, there may be added value derived from social media channels for specific niche areas such as those seen in the case studies related to drug abuse and pregnancy-related outcomes. Subject to further research, primarily to enhance AE recognition algorithms, the scope and utility of social media may broaden over time.

Michael Goodman[4], Rajesh Gosh[5], Britta Anne Grum[6], Joanna Hajne[1]; Sara Hedfors Vidlin[3], Zeshan Iqbal[7], Letitia Jiri[8], Kristina Juhlin[3], Marie-Laure Kürzinger[1], Marina Lengsavath[1], Magnus Lerch[9], Julia Lien[10], Amy Purrington[11], Sue Rees[8], Harold Rodriguez, Daniele Sartori[3], Richard Sloane[1], Stéphanie Tcherny-Lessenot[2], Sara Hedfors Vidlin[3], Benoit Vroman[12] 1 University of Liverpool, Liverpool, UK. 2 Sanofi, Chilly-Mazarin, Cedex, France. 3 Uppsala Monitoring Centre, Uppsala, Sweden. 4 AstraZeneca, Gaithersburg, MD, USA. 5 Novartis, East Hanover, NJ, USA. 6 Bayer AG, Berlin, Germany. 7 Johnson & Johnson, High Wycombe, UK. 8 Amgen Limited, Cambridge, UK. 9 Lenolution GmbH, Berlin, Germany. 10 Booz Allen Hamilton, Boston, MA, USA. 11 Janssen R&D, Horsham, PA, USA. 12 UCB Pharma, Braine-l'Alleud, Belgium

Compliance with Ethical Standards

References

1. Pierce CE, de Vries ST, Bodin-Parssinen S, Harmark L, Tregunno P, Lewis DJ, et al. Recommendations on the use of mobile applications for the collection and communication of pharmaceutical product safety information: lessons from IMI WEB-RADR. Drug Saf. 2019;42(4):477–89. https://doi.org/10.1007/s40264-019-00813-6.

2. Zeng D, Chen H, Lusch R, Li SH. Social media analytics and intelligence. IEEE Intell Syst. 2010;25(6):13–6. https://doi.org/10.1109/MIS.2010.151.

3. Edwards IR, Lindquist M. Social media and networks in pharmacovigilance. Drug Saf. 2011;34(4):267–71. https://doi.org/10.2165/11590720-000000000-00000.

4. Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety signal detection. In: Proceedings of the 2012 international workshop on Smart health and wellbeing (SHB '12). 2012, pp 33–40. https://doi.org/10.1145/2389707.2389714.

5. Ghosh R, Lewis D. Aims and approaches of Web-RADR: a consortium ensuring reliable ADR reporting via mobile devices and new insights from social media. Expert Opin Drug Saf. 2015;14(12):1845–53. https://doi.org/10.1517/14740338.2015.1096342.

6. Hazell L, Shakir SA. Under-reporting of adverse drug reactions : a systematic review. Drug Saf. 2006;29(5):385–96. https://doi.org/10.2165/00002018-200629050-00003.

7. Cobert B, Silvey J. The internet and drug safety. Drug Saf. 1999;20(2):95–107. https://doi.org/10.2165/00002018-199920020-00001.

8. TenBarge AM, Riggins JL. Responding to unsolicited medical requests from health care professionals on pharmaceutical industry-owned social media sites: three pilot studies. J Med Internet Res. 2018;20(10):e285-e. https://doi.org/10.2196/jmir.9643.

9. (IMI) IMI. Work Package 2A. IMI WEB-RADR. 2019. https://web-radr.eu/outputs/. Accessed 21 Mar 2019.

10. Brosch S, de Ferran A-M, Newbould V, Farkas D, Lengsavath M, Tregunno P. Establishing a framework for the use of social media in pharmacovigilance in Europe. Drug Saf. 2019. https://doi.org/10.1007/s40264-019-00811-8.

11. Anderson SL, Bell GH, Gilbert M, Davidson EJ, Winter C, Barratt JM, et al. Using social listening data to monitor misuse and nonmedical use of Bupropion: a content analysis. JMIR Public Health Surveill. 2017;3(1):e6. https://doi.org/10.2196/publichealth.6174.

12. 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. https://doi.org/10.1007/s40264-015-0385-6.

13. Hedfors S, Bergvall T, Gilbert M, Pierce C, Dasgupta N, Ellenius J. Improving the yield of relevant data for pharmacovigilance analysis by reducing search term complexity—a study on reddit data. Pharmacoepidemiol Drug Saf. 2016;25:412–3. https://doi.org/10.1002/pds.4070.

14. Brosch S. Frameworks for use of social media in pharmacovigilance. WEB-RADR. 2017. https://webradr.files.wordpress.com/2017/08/web-radr-stakeholder-event_theme-1b-ppt.pdf. Accessed 13 Mar 2019.

15. Caster O, Dietrich J, Kurzinger ML, Lerch M, Maskell S, Noren GN, et al. Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: results from the WEB-RADR project. Drug Saf. 2018;41(12):1355–69. https://doi.org/10.1007/s40264-018-0699-2.

16. Pierce CE. WEB-RADR WP2A Final Report on Data Collection. 2017. https://webradr.files.wordpress.com/2019/02/wp2a-report-on-data-collection.pdf. Accessed 19 Mar 2019.

17. Maskell S, Heap J, Griffith E, Bollegala D, Sloane R, Jones A et al. Estimating the pertinent information present in social media and assessing where it can add value to pharmacovigilance. IMI WEB-RADR. 2018. https://webradr.files.wordpress.com/2019/02/wp4-estimating-the-pertinent-information-present-in-social-media-and-assessing-where-it-can-add-value-to-pharmacovigilance.pdf. Accessed 21 Mar 2019.

18. 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. https://doi.org/10.1111/bcp.12717.

19. Harpaz R, Odgers D, Gaskin G, DuMouchel W, Winnenburg R, Bodenreider O, et al. A time-indexed reference standard of adverse drug reactions. Sci Data. 2014;1:140043. https://doi.org/10.1038/sdata.2014.43.

20. Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, Van Le H, et al. Evaluation of facebook and twitter monitoring to detect safety signals for medical products: an analysis of recent FDA safety alerts. Drug Saf. 2017;40(4):317–31. https://doi.org/10.1007/s40264-016-0491-0.

21. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug Saf. 2014;37(5):343–50. https://doi.org/10.1007/s40264-014-0155-x.

22. Rutten MG, Gordon NJ, Maskell S. Recursive track-before-detect with target amplitude fluctuations. IEE Proc Radar Sonar Navig. 2005;152(5):345–52.

23. Caster O, Dietrich J, Kurzinger M-L, Lerch M, Maskell S, Norén GN et al. Technical report describing implementation and evaluation of safety signal detection in social media (D2B.3), IMI, 2018. https://webradr.files.wordpress.com/2019/03/web-radr-wp2b-technical-report-signal-detection.pdf. Accessed 19 Mar 2019.

24. Gattepaille L, Hedfors Vidlin S, Bergvall T, Ellenius J. Adverse event recognition in tweets: results from a WEB-RADAR project. Drug Saf. 2018;41(11):1160–1. https://doi.org/10.1007/s40264-018-0719-2.

25. Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov. 2007;14(3):305–28. https://doi.org/10.1007/s10618-006-0052-8.

26. Tregunno PM, Fink DB, Fernandez-Fernandez C, Lázaro-Bengoa E, Norén GN. Performance of probabilistic method to detect duplicate individual case safety reports. Drug Saf. 2014;37(4):249–58. https://doi.org/10.1007/s40264-014-0146-y.

27. Ellenius J, Bergvall T, Dasgupta N, Hedfors S, Pierce C, Norén GN. Medication name entity recognition in tweets using global dictionary lookup and word sense disambiguation. Pharmacoepidemiol Drug Saf. 2016;25(S3):414–5.

28. Bergvall T, Gattepaille L, Vidlin S, Norén GN. Probabilistic record linkage to detect duplicated content in twitter prior to pharmacovigilance analyses. Pharmacoepidemiol Drug Saf. 2018;27(S2):347.

29. Erowid. Drug slang & terminology vault. In: The vaults of erowid. https://erowid.org/psychoactives/slang/slang.shtml. Accessed 22 Mar 2019.

30. Donzanti BA. Social listening for cardiac safety research—a pilot project. https://cardiac-safety.org/wp-content/uploads/2016/06/S1_5_Donzati.pdf. Accessed 22 Mar 2019.

31. Ghosh R, Akhtar A. Insights from Twitter Proto-AE analysis for Methylphenidate. IMI WEB-RADR. 2016. https://webradr.files.wordpress.com/2019/02/wp4-ga-poster-3.pdf. Accessed 14 Apr 2019.

32. Chen X, Faviez C, Schuck S, Lillo-Le-Louët A, Texier N, Dahamna B, et al. Mining patients' narratives in social media for pharmacovigilance: adverse effects and misuse of methylphenidate. Front Pharmacol. 2018;9:541. https://doi.org/10.3389/fphar.2018.00541.

33. Bhattacharya M, Snyder S, Malin M, Truffa MM, Marinic S, Engelmann R, et al. Using social media data in routine pharmacovigilance: a pilot study to identify safety signals and patient perspectives. Pharm Med. 2017;31(3):167–74. https://doi.org/10.1007/s40290-017-0186-6.

34. Rezaallah B, Lewis DJ, Pierce C, Zeilhofer HF, Berg BI (2019) Social media surveillance of multiple sclerosis medications used during pregnancy and breastfeeding: content analysis. J Med Internet Res 21(8):e13003. https://doi.org/10.2196/13003

35. Harpaz R, DuMouchel W, Schuemie M, Bodenreider O, Friedman C, Horvitz E, et al. Toward multimodal signal detection of adverse drug reactions. J Biomed Inform. 2017;76:41–9. https://doi.org/10.1016/j.jbi.2017.10.013.

Affiliations

John van Stekelenborg[1] ⬤ · Johan Ellenius[2] · Simon Maskell[3,4] · Tomas Bergvall[2] · Ola Caster[2] · Nabarun Dasgupta[5] · Juergen Dietrich[6] · Sara Gama[7] · David Lewis[7,8] · Victoria Newbould[9] · Sabine Brosch[9] · Carrie E. Pierce[10] · Gregory Powell[11] · Alicia Ptaszyńska-Neophytou[12] · Antoni F. Z. Wiśniewski[13] · Phil Tregunno[12] · G. Niklas Norén[2] · Munir Pirmohamed[14,15]

[1] Janssen R&D, Horsham, PA, USA

[2] Uppsala Monitoring Centre, Uppsala, Sweden

[3] Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3GJ, UK

[4] Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK

[5] Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA

[6] Pharmacovigilance, Bayer AG, Berlin, Germany

[7] Chief Medical Office and Patient Safety, Novartis Global Drug Development, Novartis Pharma Basel, Basel, Switzerland

[8] Dept of Pharmacy, Pharmacology and Postgraduate Medicine, University of Hertfordshire, Hatfield, UK

[9] Pharmacovigilance Department, Inspections and Human Medicines Pharmacovigilance Division, European Medicines Agency (EMA), Amsterdam, The Netherlands

[10] Booz Allen Hamilton (formerly Epidemico, Inc.), Boston, MA, USA

[11] GlaxoSmithKline, Global Clinical Safety and Pharmacovigilance, RTP, Research Triangle Park, NC 27709, USA

[12] Vigilance, Intelligence and Research Group, Medicines and Healthcare products Regulatory Agency (MHRA), 10 South Colonnade, Canary Wharf, London E14 4PU, UK

[13] AstraZeneca, Patient Safety, Office of the Chief Medical Officer, Cambridge, UK, Granta Park, Cambridge CB21 6GH, UK

[14] Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool L69 3GL, UK

[15] Royal Liverpool and Broadgreen University Hospital NHS Trust, Liverpool L7 8XP, UK