the domains of individual, commercial, public good, law enforcement, and national security interests. In these five domains, existing legislation and practices have balanced competing interests.

Disruptive technologies force reexamination of existing tradeoffs, and rightly so. Should the National Security Administration's interests be curtailed in favor of individuals' interests to exchange information, build, and nurture social media networks? Should marketers be able to use Facebook social media data and integrate this with offline data to offer more targeted advertisements? Should a liver transplant team—as recently reported by Art Caplan[8]—be allowed to use Instagram pictures of transplant candidates drinking alcohol to reject a patient's eligibility on the waiting list?

Tomorrow, should private firms be allowed to use social media data to detect disability insurance fraud, marijuana consumption, or means-tested ineligibility for entitlement programs on behalf of the government?

Rather than try and solve these individual cases, I believe that society-wide discussion of the tradeoffs between patient privacy and caring better for patients, or consumer diversity of opinion and enhancing the credibility of social media data is crucial. How we make the best use of novel social media data in smart health, and how we respect the rights of the ultimate sources of such data is an open question whose solutions are likely to dramatically impact the trajectory of innovation and health system performance.

Acknowledgments

The views expressed in this article are mine alone, and not those of my colleagues, co-authors, affiliated institutions, or funders. No endorsement of these views is implied; none should be inferred. I disclose payments received from Parkland Center for Clinical Innovation for consulting on a risk-prediction project, as well as grant support from the Agency for Healthcare Research and Quality for an exploratory social media project (R21 HS021868-01; PI: Huesch), from Precision Health Economics for consulting with pharmaceutical industry clients, and from Lockheed Martin for an exploratory project on alternative business models in healthcare.

References

1. J. Clabby, "An Update on Seton Healthcare Family and Watson," *Wintergreen Research*, blog, 2012; http://wintergreenresearch.com/blog/?p=802.
2. M.D. Huesch, M.K. Ong, and G.C. Fonarow, "Measuring 30 Day Readmission: Rethinking the Quality of the Outcome Quality Measures," *Am. Heart J.*, vol. 166, no. 4, 2013, pp. 605–610.
3. A.M. Hersh, F.A. Masoudi, and L.A. Allen, "Post-Discharge Environment Following Heart Failure Hospitalization: Expanding the View of Hospital Readmission," *J. Am. Heart Assoc.*, 2013; doi:10.1161/JAHA.113.000116.
4. R. Amarasingham et al., "An Automated Model to Identify Heart Failure Patients at Risk for 30-Day Readmission or Death Using Electronic Medical Record Data," *Medical Care*, 2010, vol. 48, no. 11, pp. 981–988.
5. M.D. Huesch, G. Ver Steeg, and A. Galstyan, "Vaccination (Anti-) Campaigns in Social Media," *Proc. Assoc. Advancement of Artificial Intelligence*, 2013; www.aaai.org/ocs/index.php/WS/AAAIW13/paper/viewFile/7094/6502.
6. G. Ver Steeg and A. Galstyan, "Information Transfer in Social Media," *Proc. Int'l Conf. World Wide Web*, 2012, pp. 509–518.
7. Centers for Disease Control and Prevention, *Vaccine Safety: Human Papillomavirus (HPV) Vaccine*, 2014; www.cdc.gov/vaccinesafety/vaccines/HPV/Index.html.
8. A. Caplan, "Is Your Doctor Spying on Your Tweets? Social Media Raises Medical Privacy Questions," *NBC News*, 2014; www.nbcnews.com/health/health-news/your-doctor-spying-your-tweets-social-media-raises-medical-privacy-f8C11427782.

Marco D. Huesch is an assistant professor at the University of Southern California's Price School of Public Policy, at Duke University's School of Medicine in the Department of Community and Family Medicine, and at the Fuqua School of Business's Health Sector Management Area. Contact him at mhuesch@healthpolicy.usc.edu.

Signal Fusion for Social Media Analysis of Adverse Drug Events

Donald Adjeroh and Richard Beal, *West Virginia University*
Ahmed Abbasi, *University of Virginia*
Wanhong Zheng and Marie Abate, *West Virginia University*
Arun Ross, *Michigan State University*

Rapid detection of adverse reactions to a drug is essential in limiting the potential harm to patients taking the drug. Detecting adverse drug reactions (ADRs) is an evolving science that has traditionally relied on research in biochemistry, genetics, and pharmacokinetics. In contrast, we attempt to exploit online data sources to analyze and predict these reactions.

A number of online information sources have spawned over the past decade: Web forums, chat rooms, blogs, social networking sites, news websites, personal webpages, and so on. We hypothesize that these online sources can assist in the study of drug adverse events. Currently, information about drugs and drug-related problems can be gleaned from online and open data sources, including agencies such as the Food and Drug Administration (FDA) in the US; public datasets such as the FDA Adverse Event Reporting System (FAERS); and VigiBase, the World Health Organization's (WHO's) global database of drug adverse event reports. However, these sources

cannot be used for real-time monitoring of adverse drug events. A key motivation for studying drug adverse events using social media data is the recent success in the use of online search query logs in reliably predicting the outbreak of influenza,[1] sometimes days or weeks ahead of traditional approaches. White and his colleagues[2] used a special software device to track mentions of drugs and symptoms of interest as users entered their web search query, in order to analyze potential drug-to-drug interactions. Nikfarjam and Gonzalez[3] analyzed user comments on health-related social networks (namely, dailystrength) to extract ADR mentions. These efforts focused on one data source, and did not consider cases where information could be originating from multiple data sources.

Given the heterogeneous nature of social media data sources, one key question is how to relate and/or combine the information about drugs and their adverse reactions as derived from these diverse sources. This is essentially an exercise in data fusion. Personal views and sentiments about drugs and their adverse reactions are often expressed in diverse social media venues, such as blogs, Twitter, chat rooms, newsgroups, web forums, social networks, and so on. Further, the media itself may be multilinguistic. Thus, the "raw text" available in each source could differ significantly, ranging from structured language (for example, a news article) to unstructured text (Twitter). A key issue in social media analytics is the design of fusion mechanisms that combine the evidence from these different sources prior to rendering a final decision. This is a general problem in data mining, pattern recognition, and machine learning involving multiple sources of evidence.[4,5]

Here, we propose a peak-labeling fusion scheme for consolidating information about drugs and their adverse reactions using the correlation between local neighborhoods within signals from different social media sources. We use Twitter and search query data as specific example sources in our experiments.

Challenges in Signal Fusion for Social Media Analysis of Drug Events

The core challenges in signal fusion can be traced to the very nature of social media data, and the specific constraints enforced by our problem of analyzing and detecting drug reactions from such sources.

Diversity of Signal Sources

A standard problem in social media analytics is the diversity of social media data sources. The difficulty in signal fusion comes from the significant differences in the nature and type of signals generated from each source, and differences in their reliability; hence, we require methods to appropriately weigh the importance of each signal.

Temporal Resolution

Although some sources offer better local temporal information, others offer better global temporal information. Yet, other sources may require variable temporal resolutions for their analysis. Determining which resolution works best for a given source is a difficult task, which is further compounded when multiple sources are being considered. Another issue is the potential time lag between sources, given possible differences in their time stamps for the same general information.

Signal Normalization

The signal values across individual sources may vary significantly. For instance, a very popular data source (such as Twitter) could have significantly higher word occurrence frequencies compared to other sources, such as a search query. Clearly, combining information from such sources requires normalizing the individual signals prior to signal fusion.

Noise and Signal Extraction

Even after normalization, some sources may still exhibit relatively low signal-to-noise ratios, compared to others. The noise could depend on the nature of the source data, and the methods and processes required for extracting the signals. Search queries would typically have a relatively higher signal-to-noise ratio than Twitter, given the problems of redundancy, spamming, and credibility in the latter, and the various preprocessing steps that are needed before valid signals can be extracted from Twitter. A related problem is the fact that some of the social media signals could be in different languages, requiring techniques for analyzing multiple languages, and using multicultural considerations in sentiment analysis.

Redundancy and Correlation

Given that contributors to the different media sources could overlap, it is inevitable that the views and sentiments expressed on a subset of sources will be similar, with possible redundancy and correlation. However, there may be other reasons for overlap. For instance, if signals from Twitter users known to be located in one region of a country are found to correlate with signals from Facebook users in a different region or country, perhaps with some time delay, this might not be due to cross-channel redundancies or overlap in the user communities, and should thus be given a serious consideration. Detecting such a correlation, and distinguishing it from that due to redundancies or overlaps in user communities is difficult, but could result in significant improvement in the quality of the fused signals.

Computational Problem

Given the number, diversity, and size of datasets involved in computational analysis of drug adverse events, and the combinatorial possibilities in combing information from various data sources, scalability, and efficiency in computation pose further challenges. This undermines one of the key advantages of using social media for drug adverse event analysis: the timely (and possibly real-time) event detection. This calls for improved data structures for representing the available information, and efficient algorithms for analyzing adverse drug events based on such representations. The use of distributed processing, cloud infrastructure, and inexpensive processing nodes in various computational grid exchanges is another approach to this problem.

Fusion Rules

Given the various types and levels of fusion,[4,5] a related computational problem in social media analysis of drug events is in deciding on the appropriate fusion technique(s) to apply, and the level at which fusion needs to be performed.

Signal Generation

For signal generation, we use the simple drug-ADR reference model, based on a predefined list of keywords for

- human anatomy,
- drug reactions, and
- drug administration problems.

That is, for a given data source, we consider joint references to a given drug (or its various aliases) and a keyword from each of the three keyword sets. We record the number of such references, over a given time period—say days, weeks, or months, depending on the type of source. We then form a time series by normalizing these counts into empirical probabilities and z-scores.

For search query, signals are generated based on publicly available information on general Web search queries. We generated query signals for each of the 46 drugs and adverse events in our reduced FDA adverse event dataset (see the description of the dataset below). For Twitter signals, we first collected all tweets pertaining to the 46 events using Topsy—a third-party tool with Twitter fire hose access and a longitudinal archive of tweets. We collected all tweets dating back to 2008 for all drug names associated with the 46 events. Some of the collected tweets were spam, including advertisements for fly-by-night websites selling shady medications. These were removed using rule-based filters. Furthermore, duplicate tweets were removed. After filtering duplicates and spam, the resulting test bed encompassed approximately 2 million tweets.

To identify potential ADR mentions, lexicons were developed for anatomy-related terms, reactions, and drug administration keywords. The lexicons, which were developed by research assistants with backgrounds in biology and medicine, were used to tag the tweets. For example, the tweet "Pradaxa caused me to experience severe internal bleeding!" would be tagged as "`<DRUG> caused me to experience severe <ANATOMY><REACTION>`". For word-sense disambiguation, we used the CMU part-of-speech tagger designed specifically for tweets,[6] to help improve the likelihood that anatomy and side-effect tags were applied appropriately.

For an adverse event E, given a time window $t_i \in T$, let $D(d)$ represent the number of drug names associated with event E that appear in tweet d. Let $W = \{d_1 \ldots d_n\}$ signify the set of tweets occurring during t_i, where each $D(d_j) \geq 1$. Further, let $A(d_j)$ and $R(d_j)$ represent the number of anatomy and reaction terms present in d_j, respectively. The total raw score for time t_i is then computed as $s(t_i) = \sum_{j=1}^{n}\left(D(d_j) + A(d_j) + R(d_j)\right)$. We convert each $s(t_i)$ to a z-score $z(t_i) = (s(t_i) - \mu)/\sigma$, where μ and σ are the mean and standard deviation, respectively, across all t_i in T for which $s(t_i) > 0$. For a given event time series, a simple approach will be to consider an alert at time t_i if $z(t_i) > \tau_T$, where τ_T is a threshold for Twitter signals. The threshold can vary and depend on the resolution of the signals—such as daily, weekly, and monthly time windows. Figure 11 shows an annotated Twitter signal for the drug Yasmin from 2009 to 2012, along with the corresponding search query signal.

Signal Fusion via Peak Labeling

Each data source is viewed as a channel of information. For multiple channels, the results of the aforementioned independently extracted signals can be combined using a simple fusion rule—for instance, majority rule at the decision level. This, however, neither considers potential correlations between the time series, nor differences between channels in terms of salience and frequency. Our approach to this problem is to fuse the signals by considering the potential local correlation that may exist between time series from different channels.

In fusing two signals, we would like to determine when the signals have similar spiking trends. Thus, for signals (say S_1 and S_2) along the same time axis, we scan each signal left-to-right and "label" the strength of the signal at each time instant t by considering peaks in the backward-window Δ_B preceding t and the forward-window Δ_F following t. We define a peak as a segment of a signal with $\geq\tau_A$ increasing ascents followed by $\geq\tau_D$ decreasing descents. We analyze windows in the vicinity of the detected peaks to determine the best numerical label for each

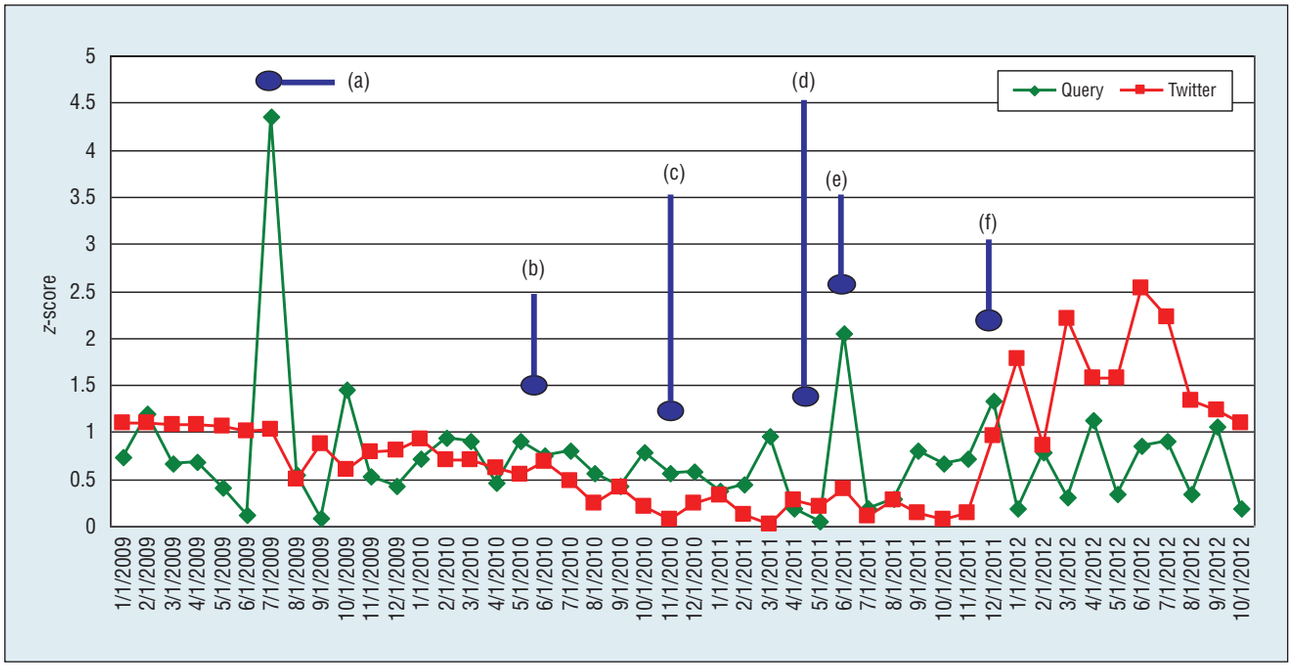

Figure 11. Annotated Twitter and query monthly signals for the drug Yasmin, from 2009 to 2012. (a) Discussion of Yasmin lawsuits. (b) A Canadian woman sues Bayer. (c) An Oklahoma City investigative news local report. (d) A *British Medical Journal* article is published. (e) A Food and Drug Administration review begins. (f) The Health Canada news release.

time instance, with smaller labels signifying that S_1 and S_2 have more similar spiking trends (that is, more local correlation). The fused signal from S_1 and S_2 is then represented by the labels corresponding to the consecutive windows of the signals. We define the labels (signal strengths) as follows:

1. Simultaneous peaks in Δ_F AND both S_1 AND S_2 have nearby peaks in Δ_B.
2. Simultaneous peaks in Δ_F AND either S_1 OR S_2 has nearby peaks in Δ_B.
3. Simultaneous peaks in Δ_F AND both S_1 AND S_2 have at least one nearby peak in Δ_B.
4. At least one simultaneous peak in Δ_F AND both S_1 AND S_2 have at least one nearby peak in Δ_B.
5. At least one simultaneous peak in Δ_F AND either S_1 OR S_2 has at least one nearby peak in Δ_B.
6. Simultaneous peaks in Δ_F.
7. At least one simultaneous peak in Δ_F.
8. S_1 has at least one peak in Δ_F and S_2 has at least one nearby peak in Δ_B; OR vice versa.
9. At least one nearby peak in Δ_B of both S_1 AND S_2.
10. Otherwise.

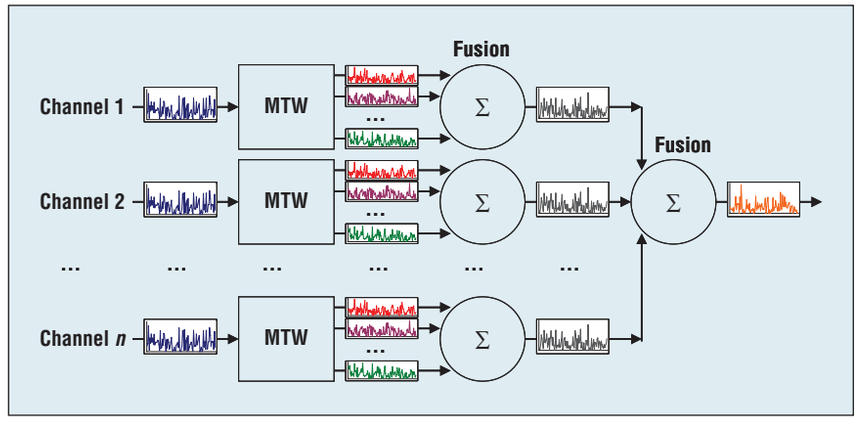

Figure 12. Multilevel cross-channel signal fusion using within-channel multiple temporal windows (MTWs).

Based on the fusion rule, our overall strategy is to apply our algorithm at different levels. First, we consider multiple temporal windows (MTW) within the same time series, and generate one fused signal using the fusion algorithm. This process is repeated for each source channel. Second, we apply the same fusion algorithm on the resulting signals from the different channels to obtain one overall fused signal (see Figure 12).

For each time period, we considered both the anatomy and reaction

Table 1. Performance statistics for valid anatomy keywords, reaction keywords, and anatomy/reaction pairs detected using various signals.*

	Anatomy		Reaction		
Signal	Recall	Precision	Recall	Precision	Success Rate
fuse([*Q*, *T*], [52,104])	**0.6869**	0.1804	**0.6830**	**0.1094**	**0.6522**
fuse([*Q*, *T*], [*n*,52,104])	0.5432	0.1740	0.5718	0.1000	0.5435
fuse[*Q*, *T*], [*n*])	0.3035	0.1815	0.3303	0.0988	0.3478
fuse([*Q*], [*n*])	0.3697	**0.2108**	0.2874	0.1012	0.4783
fuse([*T*], [*n*])	0.2444	0.1811	0.2559	0.0860	0.3261

* In this scenario, fuse(*s*,*w*) fuses the signals $s \subseteq \{Q, T\}$ (Q = Query, T = Twitter) using window sizes $w \subseteq \{n$ (global), 52 (one year), 104 (two years)$\}$. The highest numbers in each performance category are in bold.

keywords detected. First, we rank the time periods in terms of the strength of the fused signals, and then select the top q temporal positions based on this ranking for further processing. From these we further ranked the temporal positions based on the number of reaction keywords and anatomy keywords observed by the system. From this ranking, we report the final set of keywords from the top k temporal windows. In the current work, we set $q = 30$, and $k = 1$.

Performance Analysis

We now describe our experiment, its performance, and the significance of the results.

Datasets

We collected data on all drugs that had an FDA drug alert for adverse drug events, between January and September 2012. There were 67 such events, including some events that involved multiple drug groups. In our experiments, we were concerned with drug and ADR related issues, so we eliminated the 18 events that were due to drug administration problems (such as issues with broken or counterfeit tablets). Three events were removed due to insignificant data from our sources. The resulting dataset contained 46 individual drugs in 22 drug groups. For each drug group, there exists a single FDA webpage describing the problem that triggered the alert, some background to the problem, and any available history on adverse events about the drug. We manually analyzed the FDA website corresponding to each event, identifying the important points in terms of anatomy and reaction keywords. These keywords from the FDA webpages provided the ground truths, based on which we analyzed the performance of our approach.

Performance Metrics

We collected social media data about the FDA drugs from 2008 to 2012. We considered the anatomy and reaction keywords detected by the system and compared them with the keywords in the reference FDA webpages. We measure performance using the information retrieval metrics of precision and recall. Similarly, we compute the precision and recall when the anatomy and reaction keywords are used jointly. For this case, we simply use a pairwise combination of all the reaction keywords and anatomy keywords from each temporal period in the top k result. Any pair that appeared in the FDA reference webpage is taken to be a correctly detected pair.

Overall Results

Table 1 shows the overall results of our fusion scheme, with regard to precision, recall, and the success rate. The results show the significance of fusing information from multiple channels. Best results are often produced using both Twitter and query channels. On their own, the results from the individual channels were not as strong. Perhaps, more importantly, performance is also affected by the use of within-channel fusion using multiple temporal windows, before performing the cross-channel fusion. Yet, not all multitemporal windows lead to improved results. For instance, the simultaneous use of the one-year, two-year, and global windows did not necessarily lead to the overall best results. The last column shows the overall success rate: that is, the percentage of drugs for which the system identified at least one problem as contained in the FDA reference webpage, at least three months before the FDA alert.

The table shows that the system has a 65 percent success rate using $q = 30$, $k = 1$. We observed that the success rate increased with increasing k, reaching about 93 percent at $k = 5$ (for brevity, these results are omitted).

Figure 13 shows the detailed precision and recall results. Although the method could not detect the adverse event for all drugs (at the current parameter values for q and k), it did quite well on some drugs, such as Advicor, Crestor, and Zocor.

Detection Time

The success rate indicates whether the system was able to detect the problem. Another key measure is the timeliness of the detected events. Clearly, the earlier such problems can be detected, the more time it gives the FDA or drug manufactures to respond. From Figure 14, we see that, in some cases the proposed system could

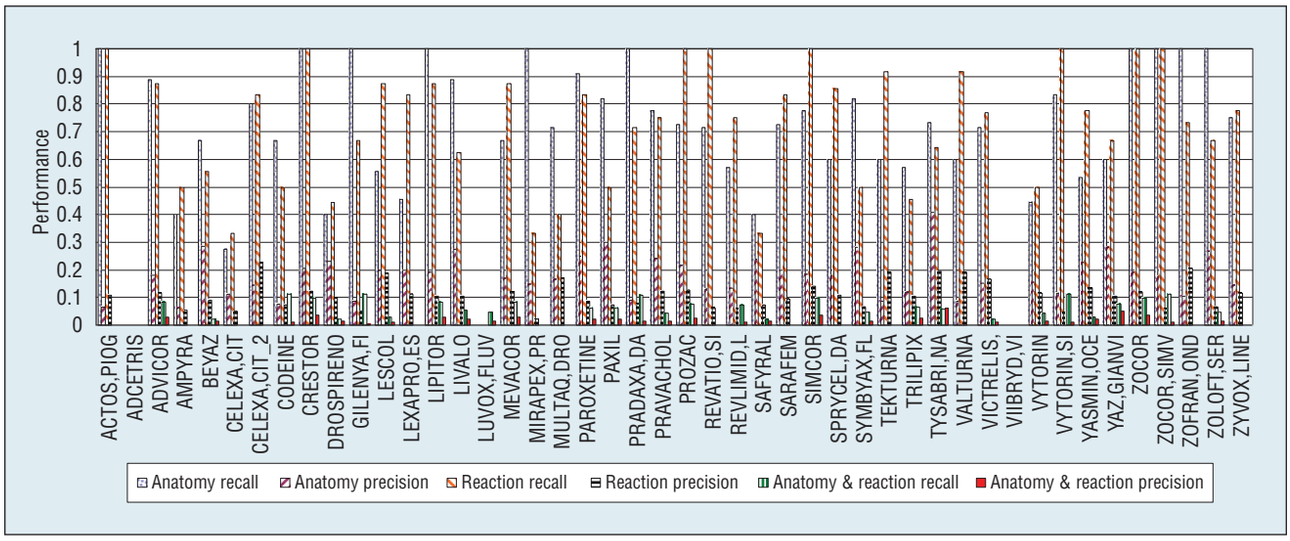

Figure 13. Performance (in terms of precision and recall) on individual drugs.

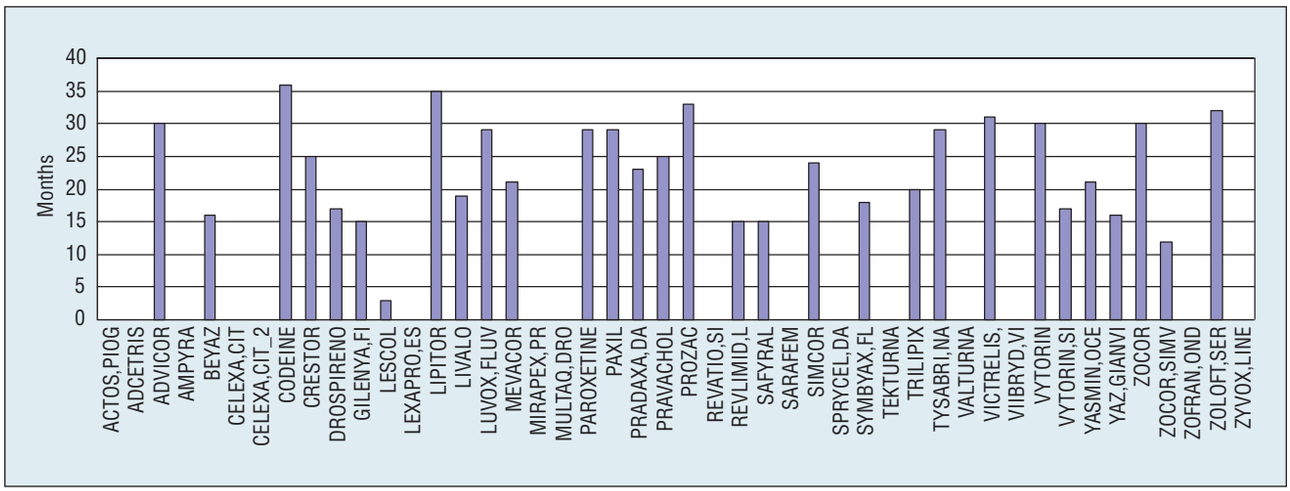

Figure 14. Performance (in terms of detection time) before the Food and Drug Administration alert (for individual drugs).

detect potential adverse drug events more than 30 months (for example, with Codeine and Lipitor) before the FDA alert. The time reported here is restricted to periods starting from early 2008. Increasing the time window could result in even earlier detection than the reported results.

We have discussed various issues involved in fusing social media signals, when the objective is to detect and monitor possible adverse drug events. We proposed a method to fuse multiple signals from different social media channels. Initial results are quite promising, with some adverse drug events detected years before FDA alerts. The results show that social media data informally provided by millions of users could provide another important source of data for drug safety studies. We are just at the beginning stages of this unconventional approach to drug surveillance. However, various challenges—such as the diversity of social media sources, noise, data redundancy, and correlation between sources—must be further considered for improved performance. ❐

Acknowledgments

We would like to thank the US National Science Foundation for their support through grants IIS-1236970 and IIS-1236983 entitled "Computational Public Drug Surveillance."

References

1. J. Ginsberg et al., "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature*, vol. 457, 2009, pp. 1012–1014.
2. R.W. White et al., "Web-Scale Pharmacovigilance: Listening to Signals

from the Crowd," *J. Am. Medical Informatics Assoc.*, vol. 20, no. 3, 2013, pp. 404–408.
3. A. Nikfarjam and G.H. Gonzalez, "Pattern Mining for Extraction of Mentions of Adverse Drug Reactions from User Comments," *Proc. Am. Medical Informatics Assoc. Ann. Symp.*, 2011, pp. 1019–1026.
4. A. Ross, K. Nandakumar, and A.K. Jain, *Handbook of Multibiometrics*, Springer, 2006.
5. J. Kittler et al., "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, 1998, pp. 226–239.
6. K. Gimpel et al., "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments," *Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies: Short Papers*, vol. 2, 2011, pp. 42–47.

Donald Adjeroh is a professor and graduate coordinator of computer science in the Lane Department of Computer Science and Electrical Engineering at West Virginia University. Contact him at don@csee.wvu.edu.

Richard Beal is a doctoral student in the Lane Department of Computer Science and Electrical Engineering at West Virginia University. Contact him at r.beal@computer.org.

Ahmed Abbasi is an associate professor and Director of the Center for Business Analytics in the McIntire School of Commerce at the University of Virginia. Contact him at abbasi@comm.virginia.edu.

Wanhong Zheng is an assistant professor and adult psychiatrist in the Department of Behavioral and Psychiatric Medicine at West Virginia University. Contact him at wzheng@hsc.wvu.edu.

Marie Abate is a professor in the Department of Clinical Pharmacy at West Virginia University. Contact her at mabate@hsc.wvu.edu.

Arun Ross is an associate professor in the Department of Computer Science and Engineering at Michigan State University. Contact him at rossarun@cse.msu.edu.