Published in final edited form as:
Epidemiology. 2012 March ; 23(2): 238–246. doi:10.1097/EDE.0b013e3182459d7d.

Active safety monitoring of new medical products using electronic healthcare data: Selecting alerting rules

Joshua J. Gagne[1,3], **Jeremy A. Rassen**[1], **Alexander M. Walker**[2,3], **Robert J. Glynn**[1,3], and **Sebastian Schneeweiss**[1,3]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

[2]World Health Information Science Consultants, LLC, Newton, MA

[3]Harvard School of Public Health, Boston, MA

Abstract

BACKGROUND—Active medical-product-safety surveillance systems are being developed to monitor many products and outcomes simultaneously in routinely collected longitudinal electronic healthcare data. These systems will rely on algorithms to generate alerts about potential safety concerns.

METHODS—We compared the performance of five classes of algorithms in simulated data using a sequential matched-cohort framework, and applied the results to two electronic healthcare databases to replicate monitoring of cerivastatin-induced rhabdomyolysis. We generated 600,000 simulated scenarios with varying expected event frequency in the unexposed, alerting threshold, and outcome risk in the exposed, and compared the alerting algorithms in each scenario type using an event-based performance metric.

RESULTS—We observed substantial variation in algorithm performance across the groups of scenarios. Relative performance varied by the event frequency and by user-defined preferences for sensitivity versus specificity. Type I error-based statistical testing procedures achieved higher event-based performance than other approaches in scenarios with few events, whereas statistical process control and disproportionality measures performed relatively better with frequent events. In the empirical data, we observed 6 cases of rhabdomyolysis among 4,294 person-years of follow-up, with all events occurring among cerivastatin-treated patients. All selected algorithms generated alerts before the drug was withdrawn from the market.

CONCLUSION—For active medical-product-safety monitoring in a sequential matched cohort framework, no single algorithm performed best in all scenarios. Alerting algorithm selection should be tailored to particular features of a product-outcome pair, including the expected event frequencies and trade-offs between false-positive and false-negative alerting.

The Sentinel Initiative ("Sentinel") is intended to improve the way in which the U.S. Food and Drug Administration (FDA) assesses medical product safety, by enabling near real-time surveillance of medical products and their outcome in routine care.[1,2] "Sentinel" will complement passive safety-monitoring systems, such as FDA's Adverse Event Reporting

Corresponding author: Joshua J. Gagne, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, 1620 Tremont Street, Suite 3030, Boston, MA 02120, (T) 617-278-0930, (F) 617-232-8620, jgagne1@partners.org.

SDC: Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the author.

System, by leveraging electronic healthcare data that are routinely and prospectively collected[3] and that are commonly used for drug-safety research.[4,5]

The main objective of active medical-product-safety monitoring is to alert stakeholders about which product-outcome associations warrant further attention. Two broad types of monitoring activities within electronic healthcare data can generate such alerts. Signal detection (or "all-by-all") activities involve mining electronic healthcare data for non-pre-specified associations among all possible product and outcome combinations. Targeted safety monitoring involves examining pre-specified product and outcome pairs, typically shortly after marketing authorization and informed by pre-approval data or by knowledge of similar compounds. FDA's Mini-Sentinel pilot has initially focused on the latter approach,[3] which FDA refers to as "signal refinement."

Whether evaluating pre-specified or non-pre-specified outcomes, confounding and other biases inherent in observational data threaten the validity of monitoring activities.[2] However, when evaluating pre-specified outcomes, prospective medical-product monitoring can resemble ordinary epidemiologic studies,[6] enabling the use of various design and analytic techniques to minimize false-positive and false-negative alerts due to bias.[7,8] Automated implementation of these processes will enable rapid and simultaneous monitoring of many pre-specified pairs.

In addition, automated alerting algorithms will be necessary to guide stakeholders to particular associations on which to focus subsequent evaluation. Rules to generate automated alerts must be both sensitive and specific, and should generate true-positive alerts quickly to facilitate timely public-health decision-making.[9] Many such rules have been proposed for active medical-product safety monitoring.[10,11] However, it is not known which algorithms perform best when coupled with semi-automated design and analytic processes for prospective drug-safety surveillance in electronic healthcare data.

In this study, we simulated serially accruing matched cohort data and compared the performance of alerting algorithms from five general classes to determine which are likely to be most useful for active medical-product monitoring in various scenarios. We then used the simulation results to select algorithms that we applied to empirical data for monitoring of rhabdomyolysis among patients newly treated with cerivastatin. Cerivastatin is a cholesterol-lowering drug that was withdrawn from the US market in 2001 because of its association with rhabdomyolysis.[12] We chose this example because it is a universally recognized drug-safety issue.

METHODS

Simulation study

Framework—We simulated data as they would accrue prospectively for newly marketed medical products in an electronic healthcare database. We used a sequential 1:1 matched cohort design, and simulated data-updating at fixed calendar intervals. In each of 20 sequential calendar intervals, we generated data for 1:1 matched users and nonusers of the monitored product of interest. The sequential matched cohort approach has previously been used in retrospective electronic healthcare data for pharmacoepidemiologic research during the early marketing time-period.[13]

The matched cohort design can be deployed rapidly in electronic healthcare data[7] and was found in a simulation study to be more efficient than other approaches for safety surveillance.[14] This design can also easily accommodate semi-automated confounding adjustment strategies that exploit the high-dimensional nature of electronic healthcare data.[15,16] Such automated variable-selection procedures can inform the construction of confounder summary scores, such as propensity scores[17] or disease risk scores,[18] on which exposed and unexposed patients are matched to balance observed baseline confounders.

The parameters in the simulation study are summarized in a table in eAppendix A (http://links.lww.com) and the data generating process is summarized in a flow chart in eAppendix B (http://links.lww.com). We created 60 monitoring scenarios characterized by 15 different expected unexposed event counts {3, 5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 150, 300, 500, 1000} and by four acceptable relative-risk thresholds (i.e. $\theta = 1, 1.25, 1.5, 2$). The event counts reflected the expected total number of events among unexposed patients across the 20 monitoring periods and ranged from extremely rare (i.e. 3) up to relatively common adverse drug events (i.e. 1,000), with respect to the frequency of the outcomes expected to be monitored by Sentinel-like systems. Stakeholders in active safety-monitoring systems may use acceptable risk thresholds if they plan to act only on associations of a certain magnitude after a thorough benefit-risk assessment. For example, stakeholders may be interested if an oral hypoglycemic agent increases the relative risk of myocardial infarction by 50%, but may determine *a priori* that a 10% increase in risk would not warrant subsequent action.[19] Thus, an alert generated for an association for which the true relative risk was 1.10 would be a false-positive alert when $\theta = 1.5$. For each of the 60 combinations of event counts and thresholds, we generated 10,000 sets of 20 sequential cohorts, for a total of 600,000 simulations.

Data generation—To simulate data for the 600,000 iterations, we first randomly selected the baseline event risk among the unexposed, *R0*, from a log-normal distribution, $\ln N(\ln[x], 0.5)$, where x was the median expected baseline risk, which equaled the expected event count (e.g. 3, …, 1,000) divided by the total number of unexposed patients in the 20 cohorts. We used a log-normal distribution so that the median expected total number of events among the unexposed was equal to the expected event count across all scenarios, and so that the expected total event count in any given scenario was never non-positive. We used a variance of 0.5 to regulate the right tail of the distribution.

We then selected a true underlying log risk ratio from a skewed-normal distribution,[20] with location of -0.5, scale of 1, and shape of 5 (the resulting percentiles of the true risk ratio [RR_{true}] distribution were 10%, 0.66; 25%, 0.83; 50%, 1.19; 75%, 1.92; 90%, 3.13). We used this distribution to reflect our prior belief that the majority of true underlying risk ratios among scenarios to be monitored in a Sentinel-like system would be just right of unity and asymmetrically distributed, with skewing to the right.

We multiplied *R0* by RR_{true} to obtain the underlying event incidence among the exposed, *R1*, for a given scenario. In each of the 20 sequential cohorts, we then generated the observed numbers of unexposed events from a binomial distribution with probability *R0* and number of trials, N, which was the number of exposed patients in the given cohort. We used a separate binomial distribution, with probability *R1* and number of trials N, to generate the numbers of observed exposed events in each cohort.

We set the numbers of exposed and unexposed patients (N) to 500 in the first monitoring period and increased N linearly across the 20 periods (to 10,000 in period 20), in order to model increasing use of the medical product in the early marketing period. We chose a large number of exposed and unexposed patients so that algorithm performance was driven almost exclusively by event counts, recognizing that there are multiple ways to accrue a set number of events (e.g. low event rate in a large population, or high event rate in a smaller population).

Alerting rules—We applied 10 groups of alerting rules, comprising 93 algorithm configurations, from five general classes, to each of the 600,000 individual monitoring scenarios. The five classes of rules comprised naïve Type I error-based approaches, group-sequential monitoring methods,[21] the maximum-sequential-probability ratio test,[10,22,23] statistical-process control rules,[24,25] and disproportionality measures.[26] The algorithms are described in eAppendix C (http://links.lww.com) and the specific construction and parameter values of each of the 93 algorithms are detailed in Table 1.

Performance evaluation—In each scenario, each of the 93 algorithm configurations had up to 20 opportunities to generate an alert. As described in Table 1, some algorithms used period-specific data as inputs (i.e. data from individual data cuts), whereas most used cumulative data. For a particular scenario, the cumulative cohort in any given period comprised all patients in prior cohorts up to and including the cohort for that period. In each scenario, we recorded whether each algorithm generated an alert, in which period alerting first occurred, and whether a true underlying causal relation of interest existed (i.e. $RR_{true} > \theta$). From this, we calculated general operating characteristics, including each algorithm's overall sensitivity (i.e. the proportion of scenarios in which $RR_{true} > \theta$ that the algorithm signaled) and specificity (i.e. the proportion of scenarios in which $RR_{true} \leq \theta$ that the algorithm did not signal). We plotted overall sensitivity and 1-(overall specificity) on a receiver-operating-characteristic curve (ROC) for each group of algorithms.

As we have previously argued,[14] comparisons of alerting-rule performance in prospective monitoring should consider time to alerting, because conventional measures of sensitivity and specificity can misrepresent the public health importance of earlier versus later alerting.[2] Therefore, we also computed event-based performance, which is a weighted average of event-based sensitivity and event-based specificity that accounts for time to alerting and accommodates a weight (w) reflecting user-specified tradeoffs between the relative costs of false positives (e.g. patients discontinue needed medication, or effective drugs are withdrawn from the market) versus false negatives (e.g. patients continue to be exposed to products that cause serious adverse events). Choosing w is analogous to choosing a cut point on an ROC curve. We provide a description of the metric in eAppendix D (http://links.lww.com). Finally, we examined the extent to which the algorithms' relative performance varied by selecting different w values, and we examined variation in relative performance across the range of expected event frequencies.

Application to empirical data

We used the simulation results to select algorithms that we then applied to empirical monitoring of cerivastatin-induced rhabdomyolysis. Cerivastatin was a cholesterol-lowering drug that was withdrawn from the US market in 2001 because of its association with rhabdomyolysis, a severe condition characterized by muscle breakdown.[12] We determined whether and when the selected algorithms would have generated alerts for this association had it been processed in an active monitoring system based on the sequential matched cohort design. We summarize the data sources and the methods in eAppendix E (http://links.lww.com).

RESULTS

Simulations

Among 600,000 scenarios, 370,405 (62%) were scenarios in which a true causal relation existed (i.e. $RR_{true} > 1.0$); of these, 251,879 (42%) were scenarios in which $RR_{true} > \theta$ (i.e. these were scenarios in which the true underlying risk ratio was greater than the signaling threshold). Figure 1 displays the overall sensitivity and 1-(overall specificity) for each group of algorithms across their parameter values. Overall sensitivities ranged from 0.1859 (group 7, X = 11; 11 consecutive period-specific estimates with *P*-values < 0.1587) to 0.9534 (group 9, X = 2; 2 consecutive cumulative effect estimates above the threshold, θ) and specificities ranged from 0.5759 (group 9, X = 2) to 0.9999 (group 10; X = 11; 11 consecutive period-specific estimates above the threshold, θ) across all 600,000 scenarios.

After accounting for time to alerting using an event-based performance metric,[9] the relative performance of the algorithms varied by the chosen weight (w), reflecting preference for sensitivity versus specificity (Figure 2A), by the expected event frequency (Figure 2B), and by the alerting threshold (eFigure 2, http://links.lww.com). In general, the most sensitive algorithms in a group (i.e. those at the bottom of each box) did not perform relatively well. In particular, naïve nominal Type 1 error approaches (group 1) generally performed poorly when large *P*-values were used, regardless of the weight (w) assigned. At lower event frequencies, α-based approaches with relatively large α or *P*-values tended to perform better than other approaches (Figure 2B). However, for frequent events, the statistical process control and disproportionality approaches tended to perform better than α-based algorithms. To assess empirical power of the algorithms, we compared their overall sensitivities, which varied widely both between and within groups of algorithms, with a range of 0.13 to 0.77 (eFigure 3, http://links.lww.com).

Simulation results from all 60 scenarios are retrievable through an on-line program at www.drugepi.org. In Table 2, we present the operating characteristics for the algorithms that attained highest event-based performance at three weights (w) among scenarios resembling monitoring of cerivastatin and rhabdomyolysis, where the expected event count was three and the alerting threshold was $\theta = 1.0$. The top two algorithms were the same at $w= 0.05$ and $w=0.10$. Alpha-based approaches, particularly the nominal Type I error-based algorithms and group-sequential methods, generally performed best in these scenarios. Overall, these seven alerting algorithms had very high specificity (range= 0.9696 – 0.9958) — reflecting, in part, the small weights that were chosen to minimize the likelihood of false-positive alerting. We applied all seven algorithms to the cerivastatin empirical data.

Application to cerivastatin and rhabdomyolysis

We monitored initiators of cerivastatin from 1998, when prescriptions for the drug began appearing in the databases, to mid-2001, when the drug was withdrawn from the US market. During this time, we observed 6 cases of rhabdomyolysis and severe myopathies among 3,530 pairs of propensity-score-matched cerivastatin and atorvastatin initiators, who contributed a total of 4,294 person-years of follow-up. All 6 events occurred among patients exposed to cerivastatin (Figure 3). All seven selected algorithms generated alerts by the end of the 13th monitoring period, corresponding to June 2001 — two months before cerivastatin was withdrawn from the market. One algorithm generated an alert as early as the close of the 10th monitoring period (September 2000). In Figure 3, we included several milestones in the history of this example, including the completion of an observational study by cerivastatin's manufacturer that failed to detect the association.[12] At the end of the entire monitoring timeframe, the estimated rate difference was 3.0 events (95% confidence interval [CI]= 0.6 – 5.4) per 1,000 person-years.

DISCUSSION

In a simulated, sequential, matched cohort framework we observed substantial variation in performance of algorithms that could be used to generate safety alerts in prospective medical product monitoring systems, such as the FDA's Sentinel System. Relative performance depended on the frequency of events in each set of simulations, whether an alerting threshold was used and its magnitude, and the tradeoffs in costs between false-positive and false-negative alerting, as reflected by values of a weight in the performance metric. No single algorithm performed best in all scenarios. Alerting-algorithm selection should be tailored to certain features of a monitoring scenario, particularly the expected event frequencies and trade-offs between false-positive and false-negative alerting.

When event frequencies were low, α-based approaches with large α-values tended to perform relatively better than other algorithms. When the expected event frequencies were high, statistical-process-control and disproportionality approaches tended to perform better. This is due, in part, to the built-in delays in alerting, which prevent these algorithms from generating chance alerts early in the early monitoring period. For example, an algorithm that requires information from five consecutive monitoring periods to generate an alert, by definition cannot generate an alert until at least the fifth monitoring period.

The algorithms' relative performance also depended on whether an alerting threshold was used and, if so, the value of this threshold. We compared the algorithms across four possible relative-risk threshold values: 1, 1.25, 1.5, and 2. A threshold of 1 implies that any indication of harm is of interest, regardless of how small. In some scenarios, stakeholders may decide that associations below a certain magnitude are not actionable. For example, stakeholders may decide that up to a 20% increased incidence of cardiovascular events among users of a particular oral hypoglycemic agent is acceptable.[19] By using an alerting threshold, the monitoring system can be programmed to generate alerts only if the observed association exceeds the specified acceptable risk threshold. The selection of a threshold should be based on a thorough benefit-risk assessment that considers, among many inputs, the severity of the monitoring event, the availability of alternative treatments, the relative benefit of the monitoring product compared with the alternatives, other safety considerations, and the background incidence of the monitoring outcome. We cannot recommend which threshold should be used in a particular monitoring scenario; we have made the simulated results for all 60 scenario types, defined by 15 expected event frequencies and by four alerting thresholds, available in an online look-up table (www.drugepi.org).

We used the simulation results to guide selection of the most appropriate algorithms for empirical monitoring of cerivastatin-induced rhabdomyolysis in two electronic healthcare databases. Over the entire monitoring timeframe, we observed a rate difference of 3.0 events (95% CI= 0.6 – 5.4) per 1,000 person-years, which is consistent with estimates from formal pharmacoepidemiogic studies.[27] Each of the chosen algorithms generated alerts before cerivastatin was withdrawn from the market. It is important to note that not all algorithms would have generated alerts in the empirical data. For example, the nominal *P*-value at the end of the monitoring timeframe was 0.0114, which was not small enough to generate alerts for many of the α-based algorithms, including four of the nominal Type 1 error approaches, seven of the Pocock-like rules, and five of the O'Brien-Fleming-like algorithms. In addition, several of the algorithms that were not selected (i.e. the three most sensitive disproportionality measures) would have generated alerts earlier than each of the selected algorithms. However, the simulation results are intended to guide algorithm selection after balancing user-specified preferences for sensitivity and specificity. Thus, the more sensitive algorithms would have generated more timely alerts in this case, but would be more likely to generate false positives in similar scenarios that lack true safety issues. Of the selected algorithms, one generated an alert after the addition of data from the third quarter of 2000, after only three observed events and nearly a year before cerivastatin was withdrawn from the market. The other algorithms generated alerts based on either four or five observed events and still before market withdrawal.

The algorithms we evaluated are not intended to prompt regulatory decisions, but rather to alert stakeholders to patterns that may warrant closer scrutiny. In a full-scale active-monitoring system, stakeholders may simultaneously monitor hundreds or thousands of potential associations, which will be too burdensome for continuous human processing. However, the features of a monitoring scenario that would prompt a reviewer to take a closer look at a particular association can be encoded into automated alerting rules. In this study, we compared 10 groups of rules from five general classes of existing sequential-monitoring approaches, coupled with four alerting thresholds. Future research should focus on incorporating other features of a monitoring scenario (e.g. requiring a certain number of events before alerting) and employing combinations of existing rules. For example, a rule could be devised such that alerting would occur only if at least four out of five alerting algorithms fired, the observed risk ratio was greater than a pre-specified threshold, and more than a pre-specified number of events were observed.

We made many assumptions that may limit the generalizability of our simulation study. First, we generated data assuming an absence of residual and unmeasured confounding and an absence of misclassification of exposure and outcomes, which is likely untenable in electronic healthcare data. However, because all of the alerting algorithms used the same basic exposure and outcome information as inputs, this assumption is not likely to affect their relative performance. In addition, an alerting threshold can allow users to incorporate confounding and misclassification that may bias effect estimates upward. For example, one may not be interested in relative effects of less than 1.25 because of concerns that they are confounded. Using an alerting threshold of 1.25 effectively redefines the reference standard by considering only estimates above this threshold. We allowed algorithms to generate alerts when their statistical criteria were fulfilled only if the observed relative frequency of exposed and unexposed events was above a pre-defined threshold. Most of the algorithms are configured to test the null hypothesis that the observed outcome frequency among the exposed is equal to that among the unexposed. It is possible that, when a threshold is used, algorithms may perform better if they are constructed to test the threshold as an alternative hypothesis. However, Kulldorff and colleagues[23] have demonstrated the practical limitations of alternative hypotheses in sequential monitoring, and the sensitivity of monitoring results to the choice of alternative hypotheses. While coupling algorithms that test the null hypothesis with the threshold criterion is one way around this issue, the optimal combination of alerting algorithms and thresholds for prospective safety monitoring requires further research.

We also assumed in the simulation study that hypothetical events occurred instantaneously, such that loss-to-follow-up could not occur and that all outcome information was available at the time of cohort formation. Many alternative modeling approaches are possible, including ones that incorporate long induction times, time-varying hazards, and informative censoring. We chose a much simpler approach as a starting point because active surveillance systems (such as Sentinel) that rely on electronic healthcare data may not be useful for identifying certain types of effects, such as long-term carcinogenic effects. However, the approach we simulated is also consistent with a setting in which the distribution of matched patients is uniform across the monitoring timeframe but the event risk increases over time. We simulated the data such that the number of matched exposed and unexposed patients and the number of events increased linearly across the 20 monitoring periods, to model data as they would stream through an active monitoring system of a newly marketed product and with data updates that occur at regular intervals defined by calendar time. It may be possible that other approaches, such as adding new data at the accrual of a set number of monitoring events (similar to monitoring approaches in clinical trials) may offer advantages for statistical efficiency. The relative performance of algorithms in this setting may differ from the setting that we considered; this should be evaluated in future work. However, across a large distributed-data-network in which many exposure and outcome pairs will be simultaneously monitored, synchronized querying at fixed calendar dates is likely the most logistically feasible approach.

Finally, we assumed a uniform event rate across the 20 cohorts in each scenario, which likely favors approaches that rely on cumulative data. Modification by time may occur if types of patients who use the treatments evolve with time and these characteristics are effect-measure modifiers. This may result in scenarios in which no true safety issue exists early in monitoring but arises in subsequent cohorts. Algorithms that use cumulative data may be insensitive to the late-appearing safety issue because of the prior accumulation of non-suggestive data. Algorithms that rely on period-specific estimates may be more capable of identifying the safety issue in a more timely manner, but this issue requires additional research.

Despite their limitations, the simulation results were useful in guiding the selection of algorithms that were then able to detect cerivastatin-induced rhabdomyolysis, a well-known drug-safety issue. A major concern of large-scale active monitoring systems is that they will generate an intractably large number of false-positive alerts. We plan to expand our approach to several empirical examples for which a generated alert would be a false-positive. The biggest threat to false positivity in a Sentinel-like system using electronic healthcare data is confounding by indication and other biases that arise from observational data. Our approach relies on a sequential-matched-cohort design, focuses on new users of the drug of interest, and matches them to users of an active comparator using propensity-score methods. Propensity-score-matching restricts patients exposed to the monitored drug to those with values that overlap with the distribution of scores for those exposed to the comparator drug. Thus, patients at the extremes of the propensity-score distributions may be excluded from analyses, yet they may be particularly susceptible to the drug-related event of interest. We suggest exploring event rates among these patients in sensitivity analyses. However, implicit restriction by matching can increase validity if observations in the tails of the distribution reflect the presence of unmeasured confounding.[28] The propensity-score-matched cohort approach offers several additional advantages for safety monitoring: (1) it enables application of many alerting algorithms without further adjustments; (2) it allows for monitoring of multiple outcomes within the same matched cohorts; and (3) it focuses on patients who are exchangeable with patients exposed to the active comparator, thereby reducing confounding and mitigating the burden of false positivity.[7] Moreover, for newly marketed drugs there are usually many more patients exposed to the active comparator, allowing for near-complete matching of patients exposed to the new drug.

In conclusion, the performance of existing alerting algorithms for prospective medical-product-safety monitoring vary when applied to a sequential, matched cohort framework for active surveillance. Algorithm selection for application to empirical monitoring should consider the expected frequency of the monitoring event, tradeoffs between sensitivity and specificity, and whether an alerting threshold should be used. Simulation results can help inform algorithm selection. By applying the results of the simulation study, as one component of an approach to prospective medical-product monitoring, we detected rhabdomyolysis associated with cerivastatin before the drug was withdrawn from the market.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support: This work was funded by grants from NIH to Dr. Schneeweiss (RC1-LM010351, RC1-RR028231, R01-LM010213, RC4-HL106376) and the Brigham and Women's Hospital-HealthCore Methods Development Collaboration. Dr. Rassen was funded by a K-award from AHRQ (1 K01 HS018088). Drs. Gagne, Glynn, Rassen, Walker, and Schneeweiss are co-investigators of the FDA-funded Mini-Sentinel project (PI: Dr. Richard Platt); however, no FDA funding supported this research and the opinions express here are those of the authors and not necessarily of Mini-Sentinel or FDA.

References

1. Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The new sentinel network--improving the evidence of medical-product safety. N Engl J Med. 2009; 361:645–7. [PubMed: 19635947]
2. Avorn J, Schneeweiss S. Managing drug-risk information--what to do with all those new numbers. N Engl J Med. 2009; 361:647–9. [PubMed: 19635948]
3. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System – a national resource for evidence development. N Engl J Med. 2011; 364:498–499. [PubMed: 21226658]
4. Bright RA, Avorn J, Everitt DE. Medicaid data as a resource for epidemiologic studies: strengths and limitations. J Clin Epidemiol. 1989; 42:937–45. [PubMed: 2681546]
5. Schneeweiss S, Avorn J. A review of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005; 58:323–37. [PubMed: 15862718]
6. Walker AM. Signal detection for vaccine side effects that have not been specified in advance. Pharmacoepidemiol Drug Saf. 2010; 19:311–7. [PubMed: 20014170]
7. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010; 19:858–868. [PubMed: 20681003]
8. Gagne JJ, Fireman B, Ryan P, et al. Study design considerations in an active medical product safety monitoring system. Pharmacoepidemiol Drug Saf. 2011 in press.
9. Gagne JJ, Walker AM, Glynn RJ, Rassen JA, Schneeweiss S. An event-based metric for comparing the performance of methods for medical product safety monitoring. Submitted: Pharmacoepidemiol Drug Saf.
10. Brown JS, Kulldorff M, Chan KA, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. Pharmacoepidemiol Drug Saf. 2007; 16:1275–84. [PubMed: 17955500]
11. Nelson J, Cook A, Yu O. Evaluation of signal detection methods for use in prospective post-licensure medical product safety surveillance. FDA Sentinel Initiative Safety Signal Identification Contract. 2009 March 31.
12. Psaty BM, Furberg CD, Ray WA, Weiss NS. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis. JAMA. 2004; 29:2622–2631. [PubMed: 15572720]
13. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. Pharmacoepidemiol Drug Saf. 2005; 14:465–476. [PubMed: 15651087]
14. McClure DL, Glanz JM, Xu S, Hambidge SJ, Mullooly JP, Baggs J. Comparison of epidemiologic methods for active surveillance of vaccine safety. Vaccine. 2008; 26:3341–5. [PubMed: 18462849]
15. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009; 20:512–22. [PubMed: 19487948]
16. Rassen JA, Schneeweiss S. Automated covariate adjustment in a distributed medical product safety surveillance systems. Pharmacoepidemiol Drug Saf. 2011 in press.
17. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006; 98:253–259. [PubMed: 16611199]

18. Cadarette SM, Gagne JJ, Solomon DH, Katz JN, Stürmer T. Confounder summary scores when comparing the effects of multiple drug exposures. Pharmacoepidemiol Drug Saf. 2010; 19:2–9. [PubMed: 19757416]

19. Home PD, Pocock SJ, Beck-Nielsen, et al. RECORD Study Team. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomized, open-label trial. Lancet. 2009; 373:2125–2135. [PubMed: 19501900]

20. Azzalini A, Valle D. The multivariate skew-normal distribution. Biometrika. 1998; 83:715–726.

21. Proschan, MA.; Gordon Lan, KK.; Turk Wittes, J. Statistical monitoring of clinical trials: a unified approach. New York: Springer Science+Business Media, LLC; 2006.

22. Lieu TA, Kulldorff M, Davis RL, et al. Real-time vaccine safety surveillance for the early detection of adverse events. Med Care. 2007; 45(10 Supl 2):S89–95. [PubMed: 17909389]

23. Kulldorff M, Davis RL, Kolcazk M, Lewis E, Lieu T, Platt R. A maximized sequential probability ratio test for drug and vaccine safety surveillance. Sequential Analysis. 2011; 30:58–78.

24. Oakland, J. Statistical process control. 6. Oxford, UK: Butterworth-Heinemann; 2008.

25. Carey, RC. Improving healthcare with control charts: Basic and advanced SPC methods and case studies. Milwaukee: American Society for Quality; 2003.

26. van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002; 11:3–10. [PubMed: 11998548]

27. Graham DJ, Staffa JA, Shatin D, et al. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. JAMA. 2004; 292:2585–2590. [PubMed: 15572716]

28. Stürmer, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a sumulation study. Am J Epidemiol. 2010; 172:843–854. [PubMed: 20716704]

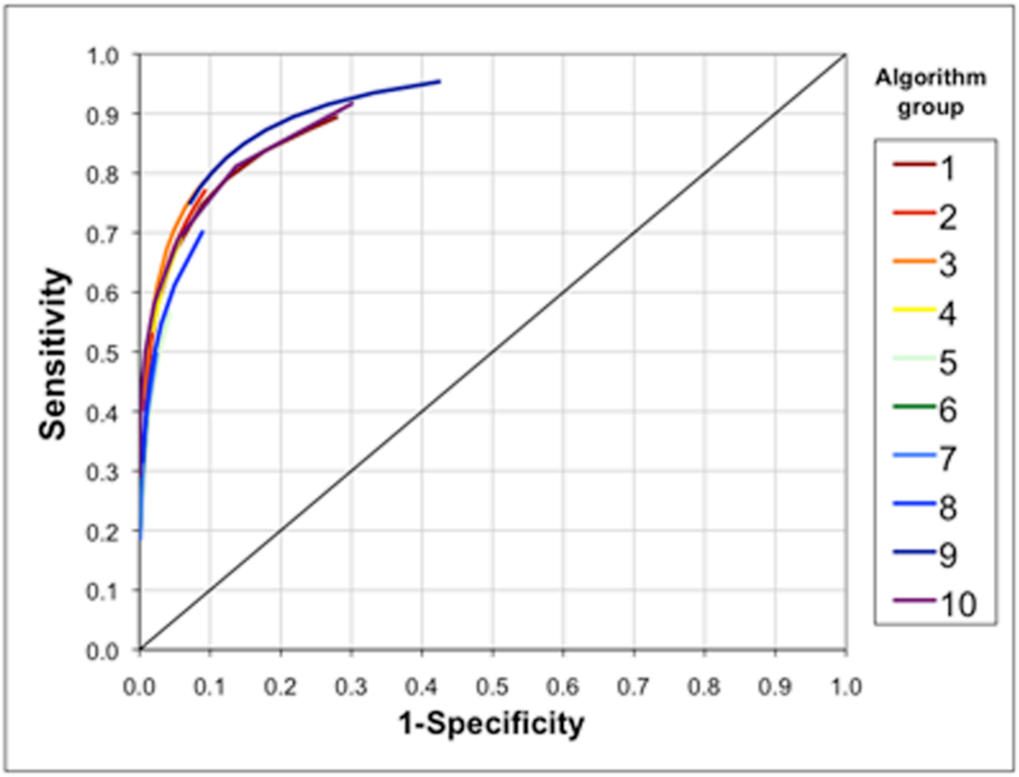


FIGURE 1.

Receiver-operating-characteristic curves based on overall sensitivity and overall specificity for 10 groups of alerting algorithms across all 600,000 simulated scenarios. Each arc comprises multiple points representing the various parameter values (e.g. different values of p, α, etc.) for each alerting algorithm group, as described in Table 1. Overall sensitivity and overall specificity do not incorporate time to alerting.

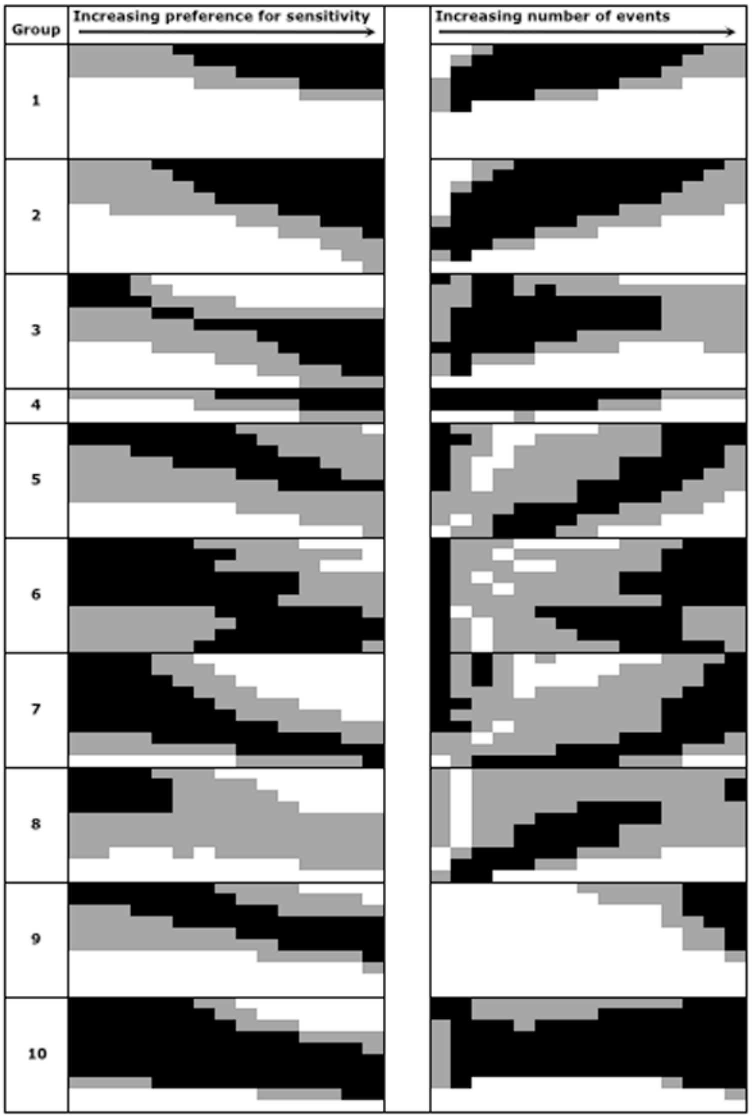
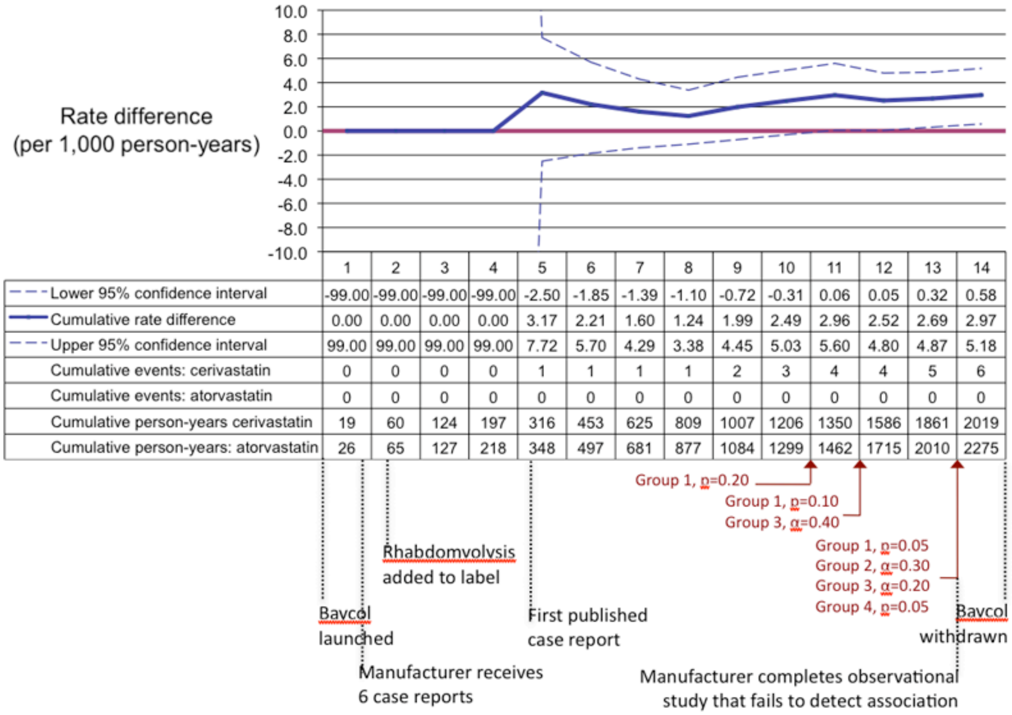


FIGURE 2.

Relative performance of alerting algorithms (A) across various preferences for sensitivity versus specificity (i.e. different weights [*w*]) and (B) across each of 15 sets of scenarios defined by expected event frequencies. Black cells represent relative performance in the top tertile, gray in the middle tertile, and white in the bottom tertile, using an event-based evaluation metric. Within each group (i.e. each box), algorithm sensitivity increases moving down the box (e.g. p increases, α increases, etc). A, The value for the weight defining the preference between sensitivity versus specificity in the evaluation metric increases from left to right from 0.02 (indicating very strong preference for specificity) to 0.30 (indicating very slight preference for specificity). B, The expected event frequency increases from left to right from 3 to 1000 and the preference weight (*w*) is held constant at *w* = 0.10 across all cells.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
- - - Lower 95% confidence interval	-99.00	-99.00	-99.00	-99.00	-2.50	-1.85	-1.39	-1.10	-0.72	-0.31	0.06	0.05	0.32	0.58
━ Cumulative rate difference	0.00	0.00	0.00	0.00	3.17	2.21	1.60	1.24	1.99	2.49	2.96	2.52	2.69	2.97
- - - Upper 95% confidence interval	99.00	99.00	99.00	99.00	7.72	5.70	4.29	3.38	4.45	5.03	5.60	4.80	4.87	5.18
Cumulative events: cerivastatin	0	0	0	0	1	1	1	1	2	3	4	4	5	6
Cumulative events: atorvastatin	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cumulative person-years cerivastatin	19	60	124	197	316	453	625	809	1007	1206	1350	1586	1861	2019
Cumulative person-years: atorvastatin	26	65	127	218	348	497	681	877	1084	1299	1462	1715	2010	2275

Group 1, p=0.20
Group 1, p=0.10
Group 3, α=0.40
Group 1, p=0.05
Group 2, α=0.30
Group 3, α=0.20
Group 4, p=0.05
Rhabdomyolysis added to label
Baycol launched
First published case report
Baycol withdrawn
Manufacturer receives 6 case reports
Manufacturer completes observational study that fails to detect association

FIGURE 3.

A reproduction of prospective monitoring of cerivastatin and rhabdomyolysis using retrospective data from two electronic healthcare databases from 1998 to 2001. In each monitoring period the numbers are updated in a cumulative fashion based on the data that became available during the corresponding calendar quarter. The black text below the table shows when each milestone in the history of this example occurred in relation to our monitoring periods.12

Table 1

Alerting algorithms evaluated in a sequential matched cohort simulation study

Type	Group no.	Specific algorithm	Parameter values
Naïve, fixed nominal Type I error levels	1	Alert when the exact *P*-value for the cumulative risk ratio < p and the observed cumulative risk ratio[a] (RR) > θ[b]	p = {0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40}
Group-sequential-Monitoring methods based on cumulative α-spending	2	Alert when the exact *P*-value, based on cumulative data, < alpha for that monitoring period as defined by the Pocock-like spending function based on cumulative alpha of α and assuming 20 equally-spaced monitoring periods, and the observed cumulative RR > θ	α = {0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40}
	3	Alert when the exact *P*-value, based on cumulative data, < alpha for that monitoring period as defined by the O'Brien-Fleming-like spending function based on cumulative alpha of α and assuming 20 equally-spaced monitoring periods, and the observed cumulative RR > θ	α = {0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40}
Sequential-probability-ratio test	4	Alert when the test statistic for the maximum sequential-probability-ratio test for binomial data exceeds the critical value based on α and using the appropriate matching ratio and the, observed cumulative RR > θ	α = {0.001, 0.01, 0.05}
Statistical-process-control rules[c]	5	Alert when the exact *P*-value for the period-specific RR < p and the observed cumulative RR > θ	p = {0.000032, 0.000088, 0.00233, 0.00577, 0.001349, 0.00298, 0.00621, 0.012224, 0.02275, 0.040059}
	6	Alert when the exact *P*-value for X out of Y consecutive period-specific RRs < 0.02275 and the observed cumulative RR > θ	X,Y = {(5,5), (4,5), (4,4), (3,5), (3,4), (3,3), (2,5), (2,4), (2,3), (2,2)}
	7	Alert when the exact *P*-value for X consecutive period-specific RRs < 0.1587 and the observed cumulative RR > θ	X = {11, 10, 9, 8, 7, 6, 5, 4, 3, 2}
	8	Alert when the exact *P*-value for X consecutive period-specific RRs < 1.0 and the observed cumulative RR > θ	X = {12, 11, 10, 9, 8, 7, 6, 5, 4, 3}
Disproportionality measures	9	Alert when X consecutive observed cumulative RRs exceed θ	X = {11, 10, 9, 8, 7, 6, 5, 4, 3, 2}
	10	Alert when X consecutive observed period-specific RRs exceed θ	X = {11, 10, 9, 8, 7, 6, 5, 4, 3, 2}

[a] The risk ratio is the observed risk ratio in the simulated data

[b] θ is a pre-defined signaling threshold equal to either 1, 1.25, 1.5, or 2, depending on the scenario

[c] The observed risk ratio corresponding to the p-values must have been indicative of harm

Table 2

Simulation results from 10,000 scenarios resembling monitoring of cerivastatin and rhabdomyolysis as defined by an acceptable risk threshold of 1 and a total expected number of events in the unexposed of 3[a]

w[b]	Algorithm	Overall sensitivity (95% CI)	Overall specificity (95% CI)	Event-based sensitivity (95% CI)[c]	Event-based specificity (95% CI)[d]	Event-based performance[e]
0.05	Group 1, p=0.05	0.2007 (0.1907–0.2107)	0.9958 (0.9937–0.9979)	0.2327 (0.2289–0.2365)	0.9975 (0.9965–0.9985)	0.9593
	Group 2, α=0.30	0.2013 (0.1913–0.2113)	0.9958 (0.9937–0.9979)	0.2286 (0.2248–0.2324)	0.9975 (0.9980–0.9988)	0.9591
	Group 3, α=0.20	0.2117 (0.2015–0.2218)	0.9955 (0.9934–0.9977)	0.2071 (0.2035–0.2108)	0.9825 (0.9975–0.9991)	0.9588
0.10	Group 1, p=0.05	0.2007 (0.1907–0.2107)	0.9958 (0.9937–0.9979)	0.2327 (0.2289–0.2365)	0.9975 (0.9965–0.9985)	0.9210
	Group 2, α=0.30	0.2013 (0.1913–0.2113)	0.9958 (0.9937–0.9979)	0.2286 (0.2248–0.2324)	0.9975 (0.9980–0.9988)	0.9206
	Group 4, p=0.05	0.2160 (0.2058–0.2263)	0.9921 (0.9893–0.9949)	0.2523 (0.2484–0.2562)	0.9947 (0.9932–0.9961)	0.9201
0.15	Group 1, p=0.20	0.3579 (0.3460–0.3698)	0.9696 (0.9641–0.9750)	0.3557 (0.3514–0.3600)	0.9805 (0.9777–0.9833)	0.8868
	Group 1, p=0.10	0.2635 (0.2526–0.2745)	0.9887 (0.9854–0.9921)	0.2822 (0.2782–0.2863)	0.9911 (0.9893–0.9930)	0.8848
	Group 3, p=0.40	0.2734 (0.2624–0.2847)	0.9885 (0.9851–0.9918)	0.2653 (0.2613–0.2692)	0.9928 (0.9911–0.9945)	0.8837

[a] Listed are the three algorithms with the highest event-based performance for each of three values of *w* (defined below) among the 10,000 simulation scenarios that were chosen to resemble monitoring for cerivastatin-induced rhabdomyolysis in which we expected to observe approximately three events among the unexposed and in which any excess relative risk among the exposed was of interest for alerting

[b] *w* is a user-defined weight that reflects trade-offs between the costs of false negatives and false positives; smaller weights reflect higher relative costs of false positives

[c] Event-based sensitivity is the proportion of observed exposed events in alert-worthy scenarios (i.e. scenarios in which a safety issue of interest exists; where the true underlying risk ratio ≥ the alerting threshold) that occurred after the given algorithm generated an alert

[d] Event-based specificity is the proportion of observed exposed events in non-alert-worthy scenarios (i.e. scenarios in which no safety issue of interest exists; where the true underlying risk ratio < the alerting threshold) that occurred before or in the absence of an alert by the given algorithm

[e] Event-based performance (EBP), which is an average of event-based sensitivity and event-based specificity, weighted by *w*, such that *EBP = (event-based sensitivity)*w + (event-based specificity)*(1-w)*