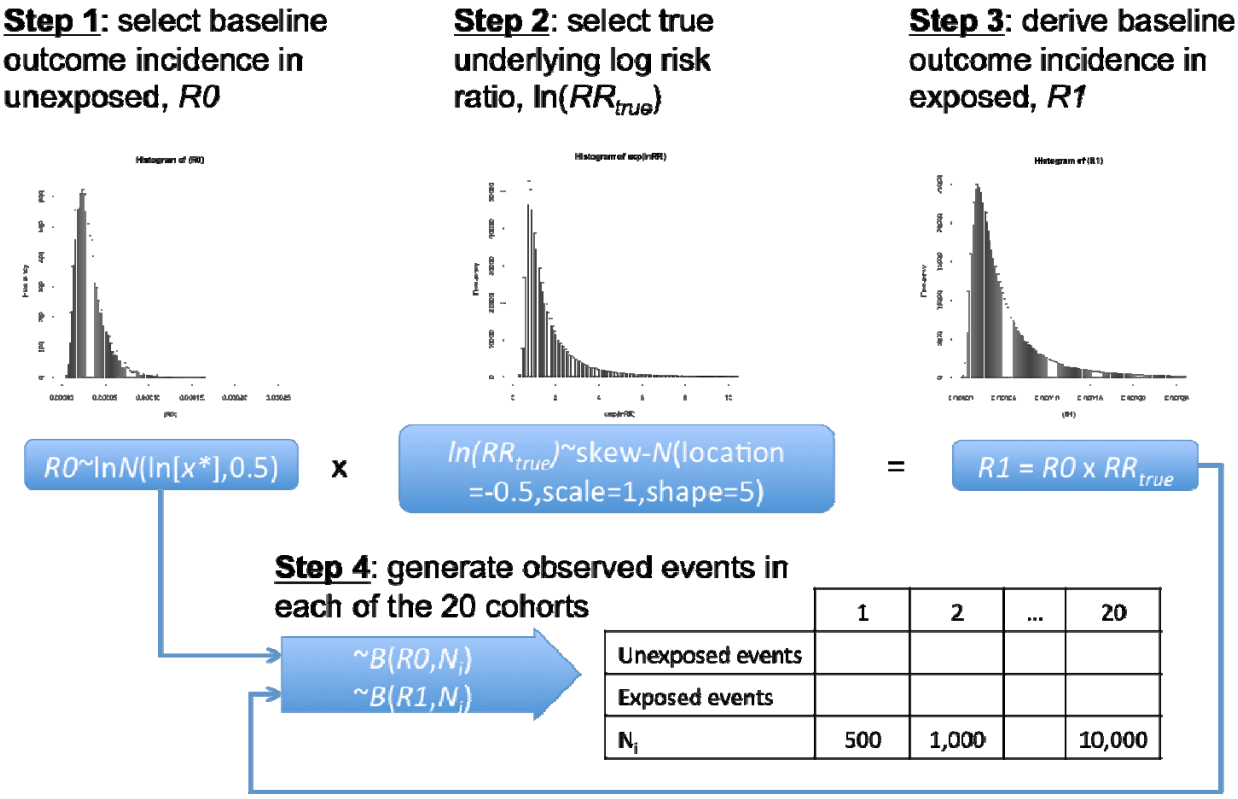
eAppendix A. Simulation study parameters

Parameter type	Notation	Values or distributions	Definition
Design	i	{1…20}	Indicator for each of 20 monitoring periods
	N_i	500*i	Number of exposed patients in each period, i
	$M0$	{3, 5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 150, 300, 500, 1000}	Expected total number of events among the unexposed across the i=20 periods
	θ	{1.00, 1.25, 1.50, 2.00}	Pre-defined signaling threshold
Known/assumed	$R0$	~log-normal(ln[x], 0.5)	Risk among the unexposed in each period
	$\ln(RR_{true})$	~skew-normal(location=-0.5, scale=1, shape=5)	True underlying log risk ratio
Derived	x	$M0/105000$[§]	Expected event risk among the unexposed
	$R1$	$R0*RR_{true}$	Risk among the exposed in each period
	RR_{true}	$exp(RR_{true})$	True underlying risk ratio
Stochastic	b_i	~B(N_i, $R0$)	Number of unexposed events in each period
	a_i	~B(N_i, $R1$)	Number of exposed events in each period

[§]105,000 is the sum of N_i over i=20 (i.e. 105,000 = 500 + 1,000 + … + 10,000)

eAppendix B. Flow chart of simulation study data generating process



*Distributions shown for *R0* and *R1* are empirical distributions based on scenarios for which the expected total event count among the unexposed equaled 3 (i.e. $x = 3/105{,}000$, where 105,000 is the total number of patients [*N*] across the 20 monitoring periods). We repeated this process 10,000 times for each of 60 combinations of 15 expected event counts in the unexposed and four alerting thresholds.

eAppendix C. Description of general classes of alerting algorithms.

Type I error-based approaches are generally used in retrospective pharmacoepidemiologic investigations that do not involve sequential testing. We implemented algorithms with a wide range of α values based on p-values from Fisher's exact test. We also applied group sequential monitoring methods that use cumulative non-linear α-spending functions, also with a wide range of α values, including the Pocock-like boundary and the O'Brien-Fleming-like boundary, as described by Proschan et al.[1] Such "stopping rules" are common for statistical monitoring of clinical trials and are intended to maintain a cumulative α-level while allowing for a pre-defined number of sequential tests of the data. This is achieved by partitioning the cumulative α-level over the monitoring periods according to some function. The Pocock-like boundary uses a convex function, which spends more α in the early periods, and the O'Brien-Fleming-like boundary uses a concave function, which saves α-spending for the later periods.

Sequential probability ratio tests, which were originally developed for monitoring the quality of manufactured goods, are now routinely used to monitor vaccine safety[2] and have recently been applied in the context of drug safety monitoring.[3] These procedures take the form of a running likelihood ratio that generates alerts when the ratio exceeds a pre-determined critical value. We implemented the maximum sequential probability ratio test as derived by Kulldorff et al. and used their published critical values.[4]

Statistical process control is another approach that was developed for quality control monitoring[5] and that has recently been applied in the healthcare setting.[6] Statistical process control methods compare observed variation in sampled units with expected variation from an underlying process to determine whether the process is out of control. We modified several statistical process control rules in a novel application to medical product safety monitoring. Finally, we modified measures of disproportionality, which have been used for drug safety monitoring based on spontaneous adverse event reports,[7] for application to sequential monitoring. Disproportionality approaches to analyzing spontaneous adverse event data typically rely on an estimated relative measure of association above some threshold combined with a lower confidence bound about that measure that surpasses a different threshold. We focused on the former of the two components and modified the approach by requiring consecutive estimates above a threshold to accommodate the sequential nature of prospective monitoring.

References

1. Proschan MA, Gordon Lan KK, Turk Wittes J. Statistical monitoring of clinical trials: a unified approach. New York: Springer Science+Business Media, LLC, 2006.
2. 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.
3. 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.
4. 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.
5. Oakland J. Statistical process control. 6th ed. Oxford, UK: Butterworth-Heinemann, 2008.

6. Carey RC. Improving healthcare with control charts: Basic and advanced SPC methods and case studies. Milwaukee: American Society for Quality, 2003.
7. 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.

eAppendix D. Description of event-based performance metric used to compare alerting algorithms in simulation study

Event-based performance (EBP) is a weighted average of event-based sensitivity and event-based specificity (Gagne JJ, 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. [In revision]).

To calculate EBP across the 10,000 iterations in each of the 60 scenarios, we used the following expression:

$$EBP = \frac{\sum_{j=1}^{k} a_j \cdot w_j}{\sum_{j=1}^{k} a_j + c_j} + \frac{\sum_{j=1}^{k} d_j \cdot (1 - w_j)}{\sum_{j=1}^{k} d_j + b_j}$$

where j is an individual scenario (i.e. iteration), k = 10,000, a_j is the number of exposed events that occurred in scenario j after alerting given that scenario j was one in which a safety issue of interest existed (i.e. $RR_{true} \geq \theta$), b_j is the number of exposed events that occurred in scenario j after alerting given that scenario j was one in which no true safety issue existed (i.e. $RR_{true} < \theta$), c_j is the number of exposed events that occurred in scenario j prior to or in the absence of alerting given that scenario j was one in which a safety issue of interest existed, d_j is the number of exposed events that occurred in scenario j prior to or in the absence of alerting given that scenario j was one in which no true safety issue existed, and w_j is a user-defined weight reflecting the tradeoffs in costs between false positive and false negative alerting. The table below depicts the relations among a_j, b_j, c_j, and d_j.

Table. Cross-classification of exposed events according to true safety status and alerting status of a particular method

		True medical product safety issue status	
		+	-
Alerting status	+	a_j	b_j
	-	c_j	d_j

Use of EPB requires stakeholders to pre-specify their preference (w), or range of preferences, for sensitivity versus specificity in a given scenario. The choice of weight is analogous to the α:β trade-off in a typical epidemiologic study, in which investigators often constrain Type I error (e.g. α = 0.05) and aim for sufficient power to limit Type II error (e.g. β = 0.20). In such a situation, investigators imply that Type I error (i.e. 1-specificity or likelihood of false positivity) is more important than Type II error (i.e. 1-sensitivity or likelihood of false negativity).

However, the relative consequences of false positives versus false negatives vary among scenario and depend on many factors including, the severity of the monitoring outcomes, the availability of treatment alternatives, and the relative benefit of the monitoring product. Specifying sensitivity versus specificity preference *a prior* prompts stakeholders to consider the trade-offs among these factors. Ideally, w would be determined by formal decision analysis.

eAppendix E. Data sources for and methods applied to monitoring of cerivastatin-induced rhabdomyolysis in two electronic healthcare databases

Data sources

We used data from New Jersey Medicare Parts A and B linked to the Pharmacy Assistance for the Aged and Disabled (PAAD) program and Pennsylvania Medicare data linked to the Pharmaceutical Assistance Contract for the Elderly (PACE) program. Both PACE and PAAD provide medications at minimal expense to elderly individuals with low income but who do not meet the Medicaid annual income threshold. The Medicare data include information on hospital and outpatient services and diagnoses. We reproduced monitoring from cerivastatin's marketing approval in June 1997 through its withdrawn from the US market in August 2001.

Monitoring framework

We mimicked prospective monitoring by dividing both databases into sequential data sets defined by claims occurring in each calendar quarter, beginning in 1998, when prescription dispensings for cerivastatin began appearing in the databases. We queried each sequential data set to identify new users of cerivastatin and atorvastatin, an active comparator with a low risk of rhabdomyolysis.[1] We defined new users as those initiating cerivastatin or atorvastatin with no use of any statin in the preceding 180 days. Because few cerivastatin prescriptions occurred in the first quarter of 1998, we combined data from quarters one and two to create the first monitoring period.

We matched cerivastatin and atorvastatin initiators on propensity for receiving cerivastatin. We constructed separate propensity score (PS) models in each sequential data set (i.e. we fit a separate PS model in each period within each database). In addition to age and sex, we included the following potential risk factors for statin-induced rhabdomyolysis as pre-defined covariates in the PS models: diagnosis of diabetes mellitus, liver disease, renal disease, hypothyroidism, and use of drugs that either cause or interact with statins to cause rhabdomyolysis.[2] We further enriched the PS models with empirically identified variables using the high-dimensional PS (hdPS) algorithm,[3] using the option that considers only covariate-exposure associations and prevalence of covariates.[4] In each model, we considered up to 100 baseline covariates from each of three domains – procedure codes, diagnosis codes, and drugs used. We matched new users of the cerivastatin to new users of atorvastatin within each period and database and then pooled the matched pairs within each period across the two databases. This approach is compatible with the privacy-maintaining PS-pooling approach described by Rassen et al.[5,6]

Patients contributed person-time until they experienced rhabdomyolysis, discontinued their index treatment (as defined by a gap in treatment of greater than 14 days), switched to a different statin, died, disenrolled, or at the end of the third quarter of 2001 when cerivastatin was withdrawn. We defined rhabdomyolysis using the algorithm for claims data validated by Andrade et al, which had a positive predictive value of 74% in a network of managed care organization databases.[7]

We analyzed data from each sequential dataset in turn as if they became available prospectively (**eFigure**). The first sequential dataset contained follow-up information through the end of June 1998 (i.e. the end of period 1) for those who initiated cerivastatin or atorvastatin in the first six

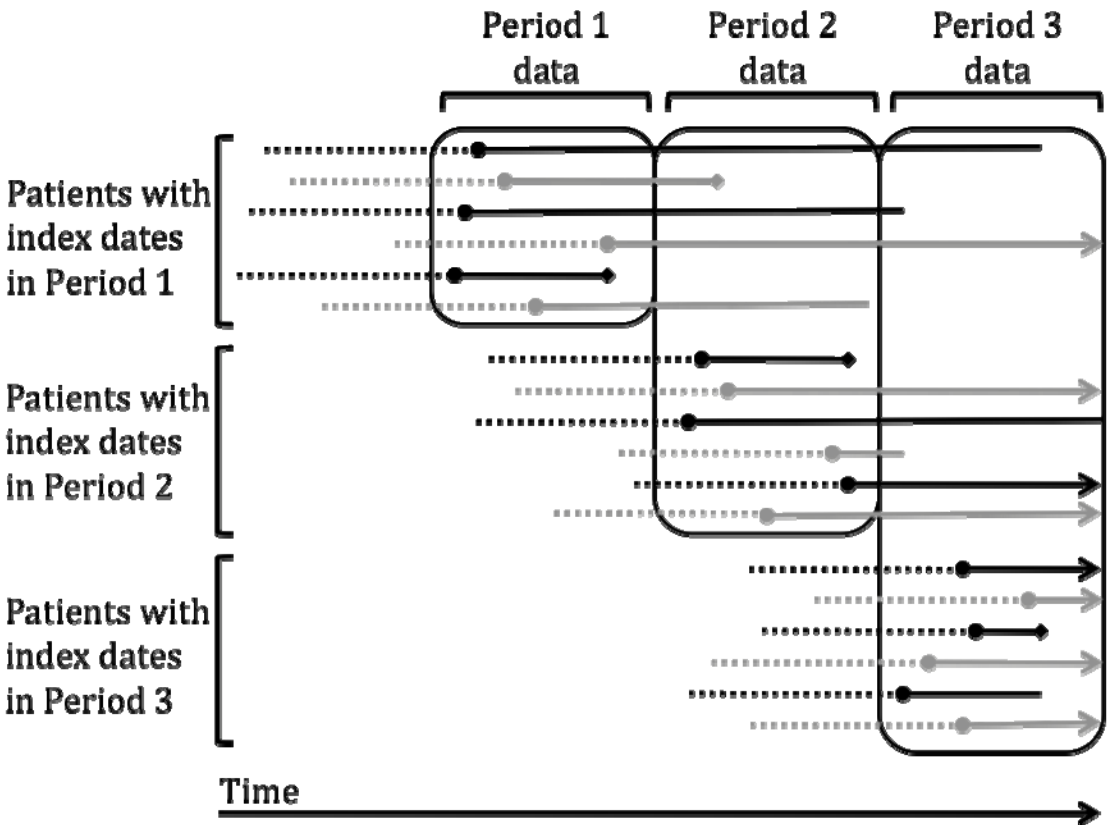
months of 1998 (i.e. the time covering period 1). The second dataset included follow-up information for patients who initiated cerivastatin or atorvastatin in the third quarter of 1998 plus continued follow-up information for those patients who initiated during the first period and whose follow-up continued into the second data set. We queried each dataset and extracted the number of matched cerivastatin and atorvastatin initiators, the eligible follow-up time, and the number of observed outcomes among matched patients. These data served as inputs into the alerting algorithms selected from the simulation study.

To select algorithms for application to this example, we restricted the simulation results to the 10,000 scenarios with an expected event frequency of 3 in the unexposed for the entire monitoring timeframe and a threshold of θ =1.0. We chose 3 events based on the observed number of matched patients in the first monitoring period (n=147), an expected event frequency of about 1 event per 20,000 statin-treated patients,[27] and by assuming that the number of matched patients would increase throughout the monitoring timeframe. We used θ = 1.0 because we were interested in detecting any elevation in risk. We selected the three algorithms with the highest EBP at three different weights (i.e. w = 0.05, w = 0.10, and w = 0.15) reflecting very high, high, and moderately high preferences for specificity over sensitivity.

References

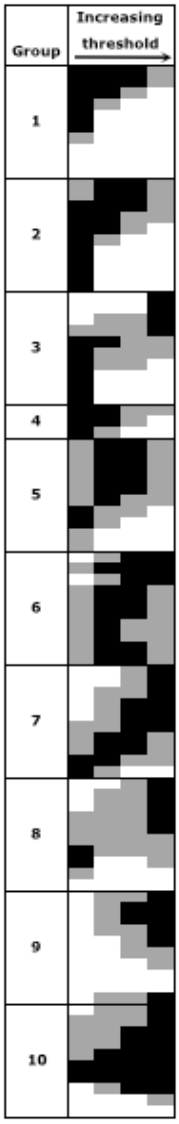
1. Graham DJ, Staffa JA, Shatin D, et al. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. *JAMA* 2004;292:2585-2590.
2. Schech S, Graham D, Staffa J, et al. Risk factors for statin-associated rhabdomyolysis. *Pharmacoepidemiol Drug Saf* 2007;16:352-8.
3. Schneeweiss 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.
4. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Observed performance of high-dimensional propensity score analyses of treatment effects in small samples. *Am J Epidemiol*
5. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf* 2010;19:848-857.
6. Rassen JA, Solomon DH, Curtis JR, Herrinton L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med Care* 2010;48(6 Suppl):S83-S89.
7. Andrade SE, Graham DJ, Staffa JA, et al. Health plan administrative databases can efficiently identify serious myopathy and rhabdomyolysis. *J Clin Epidemiol* 2005;58:171-4.

eFigure 1. Illustration of prospective monitoring with data updating at fixed calendar intervals and follow-up that may span multiple updating periods



Dotted lines represent the baseline covariate assessment period for hypothetical patients exposed to cerivastatin (black) and those exposed to atorvastatin (gray). Circles indicate the index drug initiation date and solid lines represent follow-up time. Diamonds represent events and arrowheads indicate that patients' follow-up would continue into period four data.

eFigure 2. Relative performance of alerting algorithms across different values of the alerting threshold (i.e. 1, 1.25, 1.5, 2).



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). The value of the alerting threshold increases from left to right and the preference weight (w) is held constant at $w = 0.10$ across all cells.

eFigure 3. Overall sensitivity of each algorithm across all 600,000 scenarios.