Published in final edited form as:

Med Decis Making. 2017 July ; 37(5): 512–522. doi:10.1177/0272989X16686767.

Regulator loss functions and hierarchical modeling for safety decision making

Laura A. Hatfield, PhD,

Department of Health Care Policy, Harvard Medical School

Christine M. Baugh, MPH,

Interfaculty Initiative in Health Policy, Harvard University

Vanessa Azzone, PhD, and

Department of Health Care Policy, Harvard Medical School

Sharon-Lise T. Normand, PhD

Department of Health Care Policy, Harvard Medical School and Department of Biostatistics, Harvard T.H. Chan School of Public Health

Abstract

Background—Regulators must act to protect the public when evidence indicates safety problems with medical devices. This requires complex trade-offs among risks and benefits, which conventional safety surveillance methods do not incorporate.

Objective—To combine explicit regulator loss functions with statistical evidence on medical device safety signals to improve decision-making.

Methods—In the Hospital Cost and Utilization Project National Inpatient Sample, we select pediatric inpatient admissions and identify adverse medical device events (AMDEs). We fit hierarchical Bayesian models to the annual hospital-level AMDE rates, accounting for patient and hospital characteristics. These models produce expected AMDE rates (a safety target), against which we compare the observed rates in a test year to compute a safety signal. We specify a set of loss functions that quantify the costs and benefits of each action as a function of the safety signal. We integrate the loss functions over the posterior distribution of the safety signal to obtain the posterior (Bayes) risk; the preferred action has the smallest Bayes risk. Using simulation and an analysis of AMDE data, we compare our minimum-risk decisions to a conventional Z score approach for classifying safety signals.

Results—The two rules produced different actions for nearly half of hospitals (45%). In the simulation, decisions that minimize Bayes risk outperform Z score-based decisions, even when the loss functions or hierarchical models are misspecified.

Limitations—Our method is sensitive to the choice of loss functions; eliciting quantitative inputs to the loss functions from regulators is challenging.

Address correspondence to Laura A. Hatfield, PhD, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115 USA; telephone: 617-432-0006; fax: 617-432-0173; hatfield@hcp.med.harvard.edu.

This work was conducted in the Department of Health Care Policy, Harvard Medical School.

Conclusions—A decision theoretic approach to acting on safety signals is potentially promising, but requires careful specification of loss functions in consultation with subject matter experts.

Introduction

When hospitals fail to meet safety expectations, regulators act to protect the public's health. In Massachusetts in 2005, the state Department of Public Health compelled UMass Memorial Medical Center to stop performing coronary bypass surgeries after analysis showed that mortality and infections were too high, even after adjusting for patient characteristics.¹ As reported in the Boston Globe, "...hospital executives and doctors realized in 2003 that more than 4 percent of patients who had undergone coronary artery bypass surgery died within 30 days of their operations, and that the rate of chest infections among heart surgery patients was 6.4 percent. Hospitals on average report mortality rates between 2 percent and 2.5 percent and chest infection rates of between 1 percent and 4 percent, public health and hospital officials said."² A shutdown is a serious action in response to a serious safety issue; less intensive actions are appropriate for smaller safety signals. We develop a decision theoretic approach to monitor safety performance and respond with appropriate actions. Our general procedure is to 1) fit hierarchical Bayesian models to observed safety performance in previous years (training years), 2) produce safety performance targets in a current year (test year) from the fitted models, 3) compute a safety signal that compares the observed safety performance in the test year to the target, 4) specify a loss function to quantify the costs and benefits of each action in light of the safety signal, and 5) choose the action that best balances costs and benefits. We demonstrate our proposed method by analyzing adverse medical device events (AMDEs) in pediatric hospital admissions, taking the perspective of a hospital regulator.

The rest of the paper proceeds as follows. Section introduces the particulars of post-market medical device safety surveillance, summarizes previous approaches, and distinguishes our problem and methods from these. In Section, we describe our data, hierarchical models, decision framework, and simulation study. Section gives the results of our applied data analysis and simulation study. Finally, in Section, we discuss the strengths and weaknesses of our method and suggest areas for future research.

Background

The US Food and Drug Administration (FDA), in addition to establishing safety and efficacy before treatments are marketed, also directs post-market surveillance to monitor drugs and devices used in medical care. Post-market surveillance is particularly important in devices because pre-market evidence is often lacking. Many devices enter the market based on technological equivalence to previous devices rather than randomized trials; device trials, when conducted, are typically small. Post-market device safety surveillance in children is of particular concern³ because pivotal device studies rarely include children,⁴ even though approved devices are often used to treat children.⁵

Devices are frequently used in hospital-based care. A variety of statistical approaches are used to compare hospital performance in general and hospital-based device safety in

particular. For example, hierarchical models identify hospitals in the “extreme” quantiles of the empirical distribution of hospital-level effects.⁶ Sequential probability ratio testing (SPRT) computes a running test statistic for the cumulative evidence against a null hypothesis of “performance as expected” and signals a problem when the test statistic crosses a pre-specified boundary.^{7,8} Other approaches are used in provider profiling, which quantifies the performance of health care providers on performance measures that may include safety.^{9,10,11,12}

Our objectives and approach differ from those outlined above. First, we aim to discern which action yields the best balance of real-world harms and benefits. This contrasts with approaches that rely on statistical significance (e.g., classifying “outlying” hospitals) and do not formally trade off practical costs and benefits (though they may trade off statistical costs and benefits such as bias and precision). Second, we use past data on a hospital’s safety performance to set a safety target, then take action based on comparing current safety performance to the target. This contrasts with SPRT, which continuously monitors accumulating evidence of a safety problem, rather than making a decision at a single time point based on a fixed set of historical performance data. Third, we identify only those hospitals with worse-than-expected performance that represents a departure from historical trends for that hospital. This is because our hospital safety monitoring perspective uses historical hospital performance to set hospital-specific benchmarks and detect actionable departures in a current year. Of course, other decision makers might want to rank hospitals on their whole performance record. For example, provider profiling seeks to characterize the relative performance of *all* hospitals, such as in pay-for-performance payment systems.

We apply a Bayesian decision theory framework to formalize the process of trading off real world costs and benefits of actions that may be taken in response to safety problems. The theoretical underpinnings of such a risk-minimizing approach to make *statistical* decisions are well established.¹³ However, this literature is largely confined to mathematical loss functions, e.g., squared error loss or absolute error loss, for the purpose of estimation and hypothesis testing, rather than real-world decision-making. Rarely are these ideas applied to regulatory safety settings in which decision-makers are asked to quantify the costs and benefits they consider in a safety decision. However, formal decision theory is applied in other areas, such as cost-effectiveness¹⁴ and value-of-information¹⁵ studies. Our paper presents a novel combination of a theoretical framework for decision-making with hierarchical modeling of data on safety performance to inform real-world safety decisions.

Methods

Hospital Data

The Agency for Healthcare Research and Quality (AHRQ) in the United States sponsors an annual National Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP).¹⁶ The project collects a nationally representative 20% sample of hospital discharges. The discharge summaries include diagnoses and procedures, patient demographics, discharge status, length of stay, and severity and comorbidity measures. Participating states vary by year; in 2011, 46 states and 1049 hospitals contributed data on 8

million admissions. Hospital identifiers in the data permit linkage to hospital characteristics from the American Hospital Association (AHA) Annual Survey Database.

Our sample includes admissions of patients younger than 21 years old in 2000 through 2011 (excluding births). The raw data contain 11959 observations on 4123 hospitals. We limit to hospitals with 3 or more years of at least 100 admissions in the “training years” of 2000 to 2010. These data form our “training data”: 4661 observations from 1200 hospitals. We also form a “test data” sample of the 251 hospitals that are included in the training data and also contribute data in 2011 (the “test year”).

Using a list of International Classification of Diseases 9th Edition (ICD-9) diagnosis codes^{17,18} (see Appendix A), we classify whether each admission involved an adverse medical device event (AMDE). Our outcome of interest, the AMDE rate, is defined in each hospital-year as the number of pediatric admissions in which any AMDE is observed divided by the total number of pediatric admissions.

We measure several hospital-level characteristics: hospital size (large, medium, or small) and binary indicators of private ownership, urban location, and teaching status. The remaining variables aggregate admission-level characteristics to the hospital-year level. Most measures use only the pediatric admissions in our sample: median length of stay, age, and number of chronic conditions,¹⁹ and percentage of pediatric discharges covered by government insurance (mostly Medicaid), admitted from the emergency department (ED), and ending with death in the hospital. Because states implemented different privacy rules and coding schemes, for observations with missing percent of ED admissions ($n=599$, 13%) or median age ($n=565$, 12%), we use mean imputation. Finally, we compute the proportion of all admissions that were pediatric admissions in each hospital-year.

The funding agreement ensured the authors’ independence in designing the study, interpreting the data, writing, and publishing the report.

Hierarchical Models

We fit a hierarchical model to the training data. Let y_{jt} be the number of AMDE admissions at each hospital $j=1,\dots,J$ and year $t=1,\dots,T_j$ and n_{jt} be the total pediatric admissions. We begin by transforming the AMDE rate using a variance stabilizing transformation that corrects for zero counts (by adding 0.5 and 1 to the numerator and denominator respectively), followed by a normalizing transformation (taking the arcsin of the square root),²⁰

$$y_{jt}^* = \arcsin \left(\sqrt{\frac{y_{jt} + 0.5}{n_{jt} + 1}} \right). \quad (1)$$

The addition of 0.5 to the numerator and 1 to the denominator eliminates the large spike of observations with zero AMDE counts. We use these particular correction values because they improve coverage of Wald-type intervals around proportions and provide good bias-variance trade-off for small proportions.^{21,22} The transformed observations are very close to

normal (see Appendix E Figure 3), and we can approximate the error variance of the transformed observations using $\sigma_{jt}^2 = 1/(4n_{jt})$.²⁰ This allows us to write a Gaussian hierarchical linear model for the transformed observations,

$$\begin{aligned} y_{jt}^* &\sim \text{Normal}(\mu_{jt}, \sigma_{jt}^2) \\ \mu_{jt} &= \mathbf{x}_{jt}'\boldsymbol{\beta} + b_{0j} + b_{1j} \text{year}_t \\ \mathbf{b}_j &\sim \text{Normal}((\alpha_0, \alpha_1)', \Sigma). \end{aligned} \quad (2)$$

The global coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_B)'$ govern the effects of covariates $\mathbf{x}_{jt}$, including hospital characteristics and time-varying summaries of the patient population. The hospital-level coefficients $\mathbf{b}_j = (b_{0j}, b_{1j})'$ capture idiosyncratic rates and time trends. We specify a bivariate Gaussian shrinkage distribution on these parameters with 2×2 covariance matrix Σ and hierarchical centering around global intercept and year parameters $(\alpha_0, \alpha_1)'$. The model also produces a posterior predicted (transformed) AMDE rate for each hospital in the test year τ , described below.

We fit these models using the `<monospace>rjags</monospace>` package^{23,24} in `<monospace>R</monospace>` v 3.1.3, see model code in Appendix C. We assess convergence using autocorrelation function plots, Brooks-Gelman-Rubin statistics,²⁵ and trace plots. The results reported here reflect 2000 adaptation iterations, 5000 burn-in iterations, and 5000 posterior iterations from each of 3 parallel chains.

We considered a variety of alternative model specifications, including a binomial model for the proportion of AMDE admissions and a Poisson model for the count of AMDE admissions (with an offset for total pediatric admissions). We examined different combinations of hospital-level intercepts and slopes and explored latent mixture models for the random effects. We compared models on the basis of convergence and fit statistics such as DIC²⁶. Our preferred model specification (transformed data with a Gaussian likelihood and hospital-level random slopes and intercepts) had superior performance on all the measures we considered (results not shown).

Decision Framework

A complete decision theoretic framework requires 1) a set of possible actions, 2) a parameter for the safety signal, and 3) loss functions for each action given the safety signal. We describe each of these in turn, followed by details of Bayes risk and the two decision rules we compare.

Possible actions—We define a space of actions $a \in \mathcal{A} = \{0, \dots, 4\}$ that includes taking No Action $a = 0$, Investigating the hospital by collecting more data $a = 1$, Auditing the hospital's safety procedures $a = 2$, providing Remediation (e.g., staff training) $a = 3$, or closing the hospital (Shutdown) $a = 4$.

Safety signal—We compare the observed (transformed) AMDE rate $y_{j\tau}^*$ in hospital j in test year τ to the posterior predicted (transformed) AMDE rate from the model in Eq. (2),

$$\tilde{y}_{j\tau}^* \sim \text{Normal}(\mu_{j\tau}, \sigma_{j\tau}^2). \quad (3)$$

where $\mu_{j\tau} = \mathbf{x}_{j\tau}'\beta + b_{0j} + b_{1j} \text{year}_{\tau}$. That is, we assume that the likelihood governing safety performance is the same in the training and test years. A safety signal can be defined as any measure of the discrepancy between the observed and predicted rates; for simplicity, we use a simple difference between the observation and the target, $\delta_{j\tau} = y_{j\tau}^* - \tilde{y}_{j\tau}^*$. Note that because we have the posterior distribution of the predicted (transformed) AMDE rate, we also obtain the posterior distribution of the safety signal, rather than a single value.

Loss functions—For safety signal $\delta_{j\tau}$, a loss function $\ell(a|\delta_{j\tau})$ expresses the total cost of each action a for every possible value of the safety signal. Appendix B details a set of loss functions designed to represent the costs to regulators, hospitals, patients, and society of the actions described above. In general, the loss functions impose a cost of harming patients that increases with $\delta_{j\tau} > 0$. The actions that require regulator action have both fixed and variable costs to the regulator that represent the resources required to implement the actions. Some actions produce useful information (Investigate, Audit), which enters the loss function as a negative term (i.e., a benefit). Remediation reduces the AMDE rate and Shutdown substitutes the AMDE rate of an average hospital (i.e., one with average hospital-level random intercept and slope; we call this the “community rate”) for the rate of the hospital under consideration. The loss function for Shutdown also includes a penalty for diverting patients that scales with hospital size. Figure 1 illustrates the loss functions for a small and large hospital; the units of the loss function are arbitrary, but we have formulated these in terms of dollars. Appendix B provides details of these functions.

Bayes risk—Bayesian decision theory establishes that the best decision minimizes the Bayes risk,²⁷ which is the loss function integrated over the posterior distribution of the safety signal $f(\delta|\mathbf{y}_{-\tau}^*)$:

$$E_{\delta|\mathbf{y}_{-\tau}^*}[\ell(a|\delta)] = \int \ell(a|\delta) f(\delta|\mathbf{y}_{-\tau}^*) d\delta, \quad (4)$$

where a is the action and $\mathbf{y}_{-\tau}^*$ is the training data. We have a collection of MCMC samples from the distribution of the safety signal, so we simply compute the value of the loss function for every sample and average over them to obtain the posterior risk. Further details are given in Appendix C.

Alternative decision rule—An alternative decision rule uses Z scores,

$$Z_{j\tau} = \frac{y_{j\tau}^* - E(y_{j\tau}^*)}{\sigma_{j\tau}} = 2\sqrt{n_{j\tau}}(y_{j\tau}^* - \mu_{j\tau}), \quad (5)$$

where $\mu_{j\tau}$ is the posterior predicted mean safety target from Eq. (3). Conventional Z score rules use statistical significance thresholds to decide whether an observed increase is actionable. Z scores are commonly used in studies of safety surveillance, though most of these studies make a simple classification of “significant/not significant”.^{28,29,30,11} For example, only 5% of Z scores should lie above the 95th percentile of a standard normal distribution, so a rule for identifying poor-performing hospitals might identify those with $Z_{j\tau} > q_{.95} = 1.64$. To extend beyond the binary case, we map the range of Z scores onto the action space by defining ranges. For example, $Z < 1$: No Action, $1 \leq Z < 1.5$: Investigate, $1.5 \leq Z < 2$: Audit, $2 \leq Z < 3$: Remediate, $3 \leq Z$: Shutdown. This approach is sensible only because we can order the actions in terms of increasing “severity” corresponding to increasing safety signals. Moreover, such a mapping does incorporate hospital features without further refinement (i.e., setting different thresholds for different hospital types).

In our applied example and simulation study, we set these thresholds to produce the same marginal distribution of decisions as the loss function method. That is, if the minimum-risk decision is No Action for $x\%$ of hospitals, we set the Z score threshold for No Action at the x th percentile of the Z score distribution so that the Z score action will also be No Action for $x\%$ of hospitals. Note, however, that *different hospitals* may be given a No Action decision under the two rules. This way, we compare the difference in the allocation of actions among hospitals, rather than the overall distribution of the actions.

Simulation Study

We also perform a simulation study to compare the minimum risk and Z score approaches. Appendix D contains full details of the simulation, which we summarize here. Each iteration of the simulation begins by taking a random sample of hospitals from the HCUP data. The simulated training data set begins with the observed covariates of each sampled hospital in each year in which they contributed information to the HCUP database (prior to 2011) $\mathbf{x}_{jt}$, $t \in \mathcal{T}_j$. Then we assign each hospital a random slope and intercept b_{1j} and b_{0j} ; these are the estimated slopes and intercepts from the model fit to the real data. We also assign a vector of covariates $\mathbf{x}_{j\tau}$ in a test year τ , as described in Appendix D. To generate the safety performance, we write a data-generating model that combines a set of true parameters β^{true} with the hospital characteristics, slopes, and intercepts to set the “true safety target”:

$\mu_{jt}^{true} = \beta^{true} \mathbf{x}_{jt} + b_{0j} + b_{1j} year_t$. We take draws from a normal distribution centered at this truth as the simulated (transformed) safety performance $y_{jt}^* \sim N(\mu_{jt}^{true}, 1/4n_{jt})$. The difference between the simulated safety performance and the true safety target in the test year is the “true safety signal” $\delta_{j\tau}^{true} = y_{j\tau}^* - \mu_{j\tau}^{true}$.

Next we *estimate* safety targets from the simulated safety performance, as we would do in real data analysis where the data-generating model is not known. We fit two different models to the simulated safety performance in the training years: 1) a full model that includes all the predictors in the data-generating model and 2) a reduced model that omits some predictors. This permits examination of the sensitivity of our approach to misspecification of the model for safety targets. The full and reduced models each produce a posterior expected safety

target for each hospital in the test year and thus a posterior distribution for the safety signal, $\delta_{j\tau}^{full}$ and $\delta_{j\tau}^{red}$, respectively.

We assume that there exists a true loss function, $\ell^{true}(a/\delta)$, that correctly summarizes the values of a hypothetical decision maker. In real data analysis, it is unlikely that we will be able to elicit this exact function from a decision-maker because, for example, some of the decision features may not be measured or because the elicitation process is error-prone. Thus we also specify a misspecified loss function $\ell^{mis}(a/\delta)$.

Given the two *methods* (minimum risk and Z score), three safety signals from different *models* (true data-generating, estimated with the full model, and estimated with the reduced model), and two *loss* functions (true and misspecified), we obtain a set of 12 actions for each hospital in the test year, which we denote by $a_{j\tau}^{(method),(model),(loss)}$. The gold standard action is the one that minimizes risk using the true loss function and the true safety signal, $a_{j\tau}^{risk,true,true}$. For every action, we can compute the “true risk”, that is, the risk of the chosen action computed using the true loss function and the true safety signal, $\ell^{true}(a_{j\tau}^{(method),(model),(loss)}|\delta_{j\tau}^{true})$. Note that because the true safety signal is known exactly, loss and risk are identical (i.e., there is no uncertainty over which to average). We then compare among actions using *excess risk*– the difference between the true risk of each action and the true risk of the gold standard action, i.e.,

$$\Delta_{j\tau}^{(method),(model),(loss)} = \ell^{true}(a_{j\tau}^{(method),(model),(loss)}|\delta_{j\tau}^{true}) - \ell^{true}(a_{j\tau}^{risk,true,true}|\delta_{j\tau}^{true}).$$

By definition, the excess risk measure is zero whenever an action is the same as the gold standard action. In each iteration of the simulation, we average excess risk over all the sampled hospitals’ actions in the simulated test year, separately for all 11 non-gold-standard actions.

Results

HCUP data analysis

Our analytical sample includes mostly public (55%), urban (68%), large (50%), non-teaching hospitals (73%). Table 1 summarizes hospital-year characteristics in the training data. AMDE rates are generally low, but vary across hospitals and year: 41% of hospital-years had no AMDE admissions, while the highest 5% of hospital-years had 205 AMDEs per 10,000 pediatric admissions.

Appendix E Figure 1 shows the posterior medians and two-sided 95% credible intervals for the global coefficients β . The predictors of higher AMDE rates include pediatric patients with more chronic conditions and more in-hospital mortality as well as being a teaching hospital and having a higher proportion of pediatric admissions among all admissions. Higher proportions of pediatric admissions from the emergency department, a greater proportion of pediatric admissions covered by government insurance, and long lengths of stay are all associated with lower AMDE rates. In general, AMDE rates are decreasing over

time. Correlations among the coefficients are modest (Appendix E Figure 2), except for the indicators of large and medium hospitals ($\rho=0.74$) and those for private ownership and teaching status ($\rho=0.53$).

The hierarchical model produces posterior prediction intervals for the test year AMDE rate in each hospital, shown in Figure 2. These are on the transformed rate scale of the linear model and the legend gives the counts of observations that are inside, above, or below the posterior prediction intervals. We use one-sided 95% intervals to reflect the focus on facilities with worse-than-expected safety.

Hospital A has observed AMDE rate furthest *below* its prediction interval, 1.6 per 100 pediatric admissions compared to the lower bound of 2.6. This is a medium, public, urban, teaching hospital with 20279 admissions in 2011, of which 97% were pediatric. The low AMDE rate in the test year is a departure from the AMDE rates in the training years, 2.7, 2.8, 3.1 per 100 pediatric admissions in 2003, 2004, 2005.

Hospital B has AMDE rate furthest *above* its prediction interval, 0.9 per 100 pediatric admissions compared to the upper bound of 0.6. Hospital B is a small, private, urban, non-teaching hospital, at which children comprised 5% of the 4622 admissions in 2011. Its high AMDE rate in the test year is a departure from its low AMDE rates in the training years, 0, 0.2, 0 per 100 pediatric admissions in 2004, 2005, 2006.

Table 2 shows the percent of test year observations having each action chosen using both the Bayes minimum risk (rows) and the Z score (columns). The marginal proportions of actions are the same for both rules: 65% No Action, 22% Investigate, 8% Audit, 4% Remediate, and 1% Shutdown. Among the 251 test year observations, the two decision rules produced different actions for 114 (45%) hospitals. The most concordant decision (i.e., along the diagonal of the table) between the two rules was No Action, while the most common discrepancies were for hospitals where the Z score rule indicated No Action while the Bayes minimum-risk decision was Investigate (13.9%) or vice versa (15.5%).

Figure 3 shows the distribution of the risks for actions chosen using these two decision rules. The risk differences for the commonly discrepant decisions (Investigate versus No Action) are usually quite small. For example, the average risk difference among hospitals where the Bayes action was Investigate and the Z score indicated No Action was only \$489. (Recall that the risk may be interpreted as expected loss in dollars.) Compare this to the largest average risk difference, \$19166, when the Bayes decision is Remediate but the Z score indicates Shutdown (which occurred in only 0.4% of hospitals).

Hospital C has a Z score action with the largest posterior risk. This is a large, private, urban, non-teaching hospital with 16% pediatric admissions among its 15553 admissions in 2011. The observed rate of 0.4 AMDE per 100 pediatric admissions is above the prediction interval (0 to 0.3 per 100 pediatric admissions) and the Z score of 3.6 was above the threshold of 1.8 for a Shutdown decision. This hospital's test year performance is a departure from the training year AMDE rates: 0, 0.1, 0.1, 0 in 2001, 2003, 2005, 2006. The Bayes minimum-risk decision for this hospital is Remediate because of the loss function's

penalty for diverting patients and the reference to the community AMDE rate, which was similar to the observed rate at Hospital C.

Hospital D is the only hospital for which both rules chose Shutdown. This is a small, public, urban, teaching hospital with previous AMDE rates of 5.6, 3.8, 4 in 2002, 2004, 2006. The test year rate of 4.7 is a departure from this trend, and well outside the prediction interval of (0.8, 4.2). In 2011, the hospital had only 1080 total admissions and compared very unfavorably to the community AMDE rate of 1.6, making Shutdown a good decision from both the Z score and minimum risk perspectives.

Simulation study

Figure 4 displays the distribution (across 500 simulation iterations) of the average excess risk of each decision rule. The excess risk is 0 by definition for the gold standard case (minimum risk, true signal, true loss), so we do not display that in the figure. In general, the distribution of average excess risk for the Z score actions is larger and more spread out compared to the excess risk of the minimum-risk actions, with some Z actions having *much* larger excess risk. All of the largest Z score excess risks are cases in which the Z score indicates Shutdown and the minimum-risk action is less serious. We can also see the impact of model and loss function misspecification on both types of decision rules, which is particularly acute when both are misspecified (top right).

Discussion

We develop an approach to combine hierarchical modeling for safety targets with realistic loss functions that quantify the costs and benefits of taking action on safety signals. In our analysis of AMDE rates in pediatric admissions, we find substantial differences in the actions chosen for hospitals on the basis of a loss function that integrates costs and benefits and actions chosen solely on the strength of the statistical evidence for a safety problem. In our simulation study, we find much smaller excess risk of minimum-risk actions compared to actions chosen on the basis of Z scores, even when the loss function or the model are misspecified. This supports the use of minimum-risk decisions, even when misspecification of the models for safety targets or the loss functions is a concern.

In our comparison of these two decision rules, we have made two key implementation choices driven mainly by limitations of the Z score method. First, the Z score method can only produce more serious actions in response to stronger safety signals. This implies that the marginal proportion of actions should decrease with intensity, assuming that serious safety problems are rare. Thus we set the parameters of the loss functions to produce decisions with this same feature: i.e., there is a decreasing relationship between the seriousness of the action and the marginal probability of that action.

Second, the Z score method requires us to fix thresholds for each action, meaning the marginal distribution of actions is determined in advance. For example, in a binary decision setting, placing the threshold for action vs no action at the 95th percentile of the standard normal implies that we will act 5% of the time under a Z score rule. The minimum-risk method does not require any *ex ante* decisions about the probability of each action. Thus,

having obtained the marginal distribution of actions by applying the minimum-risk method, we set the Z score thresholds to produce this same distribution of actions. This focuses the comparison between the two rules on the allocation of the “right” hospitals to each action, rather than on the overall distribution of actions.

The performance of the two rules could be de-coupled by setting the Z score thresholds based on other criteria. However, modifying the basic Z score rule is beyond the scope of this paper. Our minimum risk method has a sound theoretical underpinning and allows decision-makers to assign value to factors of the decision on a common and meaningful scale, such as dollars. By contrast, adjusting Z score thresholds to allow different thresholds for hospitals with different characteristics would lack this theoretical grounding and, we believe, be even more difficult than eliciting loss functions.

This study has several limitations, related both to our applied data example and our proposed method. The first and most challenging obstacle is specifying the inputs to and functional form of the loss functions. In the early phases of this project, we met with patient safety advocates and hospital administrators to understand the criteria used in making decisions about safety data. They described informal processes, implemented by networks of decision-makers using diverse (and often insufficient) data with largely *ad hoc* criteria. Such “messy” processes are difficult to formalize and quantify in loss functions. Indeed, the decision-makers we consulted were more concerned with the enormous burden and perceived illegitimacy of external quality and safety measurement requirements (in the case of the hospital safety officer) or the poor quality of reported data from hospitals and lack of resources for taking action (in the case of the safety advocates). Clearly, the state hospital regulatory environment is a fraught scenario in which to formalize safety decisions. Where possible, we incorporated these decision-maker inputs, but further work is needed to incorporate regulator input into realistic loss functions.

However, we can imagine more tractable scenarios in which the data are clearer, stakeholder interests are aligned, and decision-makers are empowered. For example, a medical device manufacturer tracks orders for, deployments of, and problems with its products. Corporate decision makers may take a range of actions in response to safety problems including intensified postmarket data collection, forensic engineering studies of returned faulty products, “Dear Doctor” letters, or voluntary recall. A manufacturer could better quantify the costs and benefits of lost revenue, internal resources use, and patient harm. Moreover, with internal decision makers acting on internal evidence, the incentives for formalized processes are better aligned.

In this study, we use nationally representative administrative data on pediatric inpatient care. Although a valuable source of broadly representative information about care delivery in the real world, it lacks clinical detail and more detailed device information. This limits our ability to identify particular devices or patient characteristics associated with more intensive care or vulnerability to safety issues. Furthermore, only device problems that generate a billable event in the course of an admission are accurately coded in our data. In addition, our data do not allow us to identify children’s specialty hospitals, though we can accommodate these higher rates through the hospital-level random effects.

In addition, we limit to hospitals with sufficient pediatric admissions and enough years of observations to make longitudinal analysis tractable, i.e., at least 3 years of at least 100 observations each. This is because the AMDE proportions are so low in most hospitals that yearly admission counts lower than 100 would render any method unreliable. For instance, in our sample, the median number of AMDE admissions is 14 out of every 10,000, or 0.0014. With 100 admissions, this translates to less than one (0.14) ADME in each year. This limits the applicability of these methods for very rare outcomes, particularly if only a fraction of the hospital population is at risk (e.g., only pediatric patients, as in this study).

We also limit to covariates that are available in our data for all states, which excludes variables that were omitted from the database according to state-specific reporting limitations. The set of available covariates resembles those used in previous studies of medical device adverse events. One previous study of pediatric AMDE in emergency department visits³¹ included “medical specialty, device category, and device classification, as well as patient demographic characteristics, injury diagnosis, and patient disposition.” That study focused on incidence-level reports and national estimates of counts, rather than hospital-level rates. A second previous study of pediatric specialty hospitals included age, gender, race, length of stay, disposition, chronic conditions, presence of common diagnoses, and number of admissions by patient.¹⁷ In addition to the observable characteristics of hospitals, our model accounts for unobservable (stable) differences in hospitals through the hospital-level random effects.

In summary, our proposed combination of a formal decision theoretic framework with hierarchical modeling for safety targets is a promising technique to formalize and improve decision making in postmarket safety surveillance. This paper is an initial step in the development of robust decision rules that are statistically grounded and properly reflect the values of decision-makers. In ongoing work, we are exploring the specification and robustness of loss functions and extending these ideas to other decision settings.

Acknowledgments

The authors wish to thank Thomas Kelley and Jeannie Biniek for helpful comments on the draft and Alyna Chien, Jisun Jang, and Mark Schuster for sharing code to construct the chronic condition indicator variables. We are also grateful to two anonymous reviewers and the associate editor for their constructive comments.

LAH, VA, and S-LTN were funded by a grant from the US Food and Drug Administration (U01 FD004493-01) as part of work under the Medical Device Epidemiology Network Methodology Center. CMB is supported by the National Institute of Mental Health (T32MH019733). The funding agreements ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Appendix A: Diagnosis codes

Table 1 lists the billing codes used to identify AMDE admissions from hospital billing claims.

Table 1

ICD-9 diagnosis codes use to identify adverse medical device events (AMDE) in claims

Category	Codes
Cardiac	996, 996.01, 996.02, 996.04, 996.09, 996.61, 996.71, 996.72
Dialysis	996.56, 996.57, 996.68, 996.73
Genitourinary	996.3, 996.31, 996.32, 996.39, 996.64, 996.65, 996.76
Nervous system	996.2, 996.63, 996.75
Ocular	996.53
Orthopedic	996.4, 996.4, 996.41, 996.42, 996.43, 996.44, 996.45, 996.46, 996.47, 996.49, 996.66, 996.67, 996.77, 996.78
Vascular access	996.1, 996.62, 996.74
Other	996.59, 996.69, 996.7, 996.79, E878.1, E878.2, E879.4

Appendix B: Loss function details

A regulator faces a sequential decision problem; after an investigation, audit, or remediation, she must decide whether to continue escalating the response. For example, more intensive data collection may reveal reporting errors indicating the true safety signal is larger (smaller) than thought and thus more (less) action is needed. Or the remediation measures may successfully address the safety concerns and bring the counts in line with expectations. Thus after any action, we re-compute the safety signal in light of new evidence and additional accumulated AMDE cases, and make a new decision. Our loss functions therefore need not incorporate subsequent losses or actions; the Bayesian updating of evidence ensures that at each subsequent decision point, we act appropriately based on the full posterior distribution at that time.

Because we make decisions at the hospital level, we can incorporate hospital features in our loss functions. Shutting down a hospital is a drastic action that incurs a cost that depends on the size and importance of the hospital in the local health care market. For example, a small rural hospital that triages and treats the majority of patients in a sparsely populated area might incur larger costs in a Shutdown than a small specialty hospital that takes selected cases in an urban environment with many alternative providers. In addition, large facilities in any market entail more loss under Shutdown because more patients will be affected.

Suppose that we form a prediction for the year τ and then evaluate the data collected through the first half of the year to make our decision. Assuming cases accrue uniformly over the year, we will have observed half the year's AMDE cases and will compare this to half of our prediction $\tilde{y}_{j\tau}/2$. Next we will consider the losses of each of our actions for true safety signals. For ease of explication, we will refer to expected cases as $\tilde{y}$ and observed as y in the following (i.e., ignoring the halving and data transformation), the safety signal as $\delta = (y - \tilde{y})$, and the number of pediatric admissions in the first half of the year at each hospital as n .

Table 1

Parameters of the loss functions illustrated in Fig 1 of the main text and used in the real data analysis as well as the true and misspecified loss functions in the simulation.

Parameter	Description	Real data	True	Misspecified
c^{harm}	Cost of AMDE (per excess case)	\$1000	\$1000	\$1100
c^{div}	Cost of diversion (per patient)	\$20	\$20	\$15
c^{info}	Benefit of information (per case audited/investigated)	\$-250	\$-250	\$-200
c^{reg}	Cost of regulator effort (Investigate, Audit, or Remediate)	\$200	\$200	\$300
c^{audit}	Cost of auditing (per AMDE case)	\$350	\$350	\$400
c^{remed}	Cost of remediation (per admission)	\$50	\$50	\$100
c^{iso}	Regulator cost multiplier for isolated hospitals		0.5	0
c^{spec}	Cost of shutting down specialty hospital		\$10000	\$0

If we take No Action ($a = 0$), we incur no loss if a hospital is performing at or above expectations ($y \geq \bar{y}$). If a hospital is performing below expectations, each excess AMDE case ($y - \bar{y} > 0$) experiences loss c^{harm} , which summarizes the morbidity and mortality harms. The loss function is constant 0 for negative signals and increasing for positive signals:

$$\ell(a=0|\delta) = c^{harm} \delta^+, \quad (1)$$

where δ^+ indicates only the positive part of δ , i.e., $\delta I_{\delta > 0}$.

One might argue that *any* AMDE is unnecessary, but we are willing to accept reasonable and necessary risks of medical technology, which our predictions implicitly incorporate. Thus our loss function increases in the number of patients harmed *above* usual rates.

Assume that our protocol for increased data collection ($a = 1$) requires a review of AMDE cases at a hospital. There are three key components of the loss function: effort required to collect data, the value of the information it produces, and the ongoing harm to patients that accrues during the review. Case review entails a fixed cost to the regulator, c^{reg} , to coordinate the investigation and then act on the results. It also requires some fraction of the hospital's staff; we assume a hospital can afford to examine a number of AMDE cases up to 5% of their total pediatric population. This produces c^{info} worth of information on each case reviewed. In this simple formulation, we conceive of information benefit as the inverse of variance, i.e., precision, which is proportional to $\sqrt{n}$. Finally, we assume that excess AMDE cases continue to accrue at the same rate. Thus the loss function is

$$\ell(a=1|\delta) = c^{reg} - c^{info} \sqrt{\min(.05n, y)} + c^{harm} \delta^+. \quad (2)$$

Auditing a hospital's safety procedures ($a = 2$) for 2 months resembles data collection, with a few differences. The same fixed costs accrue to the regulator c^{reg} but in addition, we must

compensate an external audit team, rather than relying on hospital personnel. We assume these costs c^{audit} scale with the log of the number of pediatric admissions during the audit, because the audit must cover the range of clinical services in order to identify potential safety concerns (recall that n is the number of pediatric admissions in the first 6 months). As with investigation, the information produced also scales with the square root of the cases reviewed. In the audit, this includes every pediatric admission for 2 months, so information is proportional to $\sqrt{n/3}$, and the excess harms to patients continue to accrue. Thus, the loss function is

$$\ell(a=2|\delta) = c^{reg} + c^{audit} \log n / 3 - 0.5 c^{info} \sqrt{n/3} + c^{harm} \delta^+. \quad (3)$$

Implementing remediation ($a = 3$) incurs the same fixed supervisory cost to the regulator c^{reg} , plus additional cost to design and supervise a remediation strategy, c^{remed} , because this is more complex than the previous two actions. The costs scale with the log of pediatric admissions n because remediation must be implemented more widely. The benefit of remediation is not in terms of information produced, but cases reduced. We assume remediation halves the rate of excess AMDE, thus reducing harms. Thus,

$$\ell(a=3|\delta) = c^{reg} + c^{remed} \log n + c^{harm} 0.5 \delta^+. \quad (4)$$

If we shut down ($a = 4$) a hospital that is performing as well as or better than expected, the n patients who would have been treated at the affected hospital each experience loss c^{div} , which summarizes additional travel costs, disruption of care, inability to access usual providers, etc. In addition, diverted patients are exposed to the community average AMDE rate, $\mu = \mathbf{x}'_{j\tau} \beta + \beta_0 + \beta_1 year_\tau$. We compute the additional increase in harms to patients at the community level ($\delta^{comm} = (n\mu - \tilde{y})$) versus the hospital under consideration.

$$\ell(a=4|\delta) = c^{div} n + c^{harm} (\delta^{comm} - \delta). \quad (5)$$

The harm of shutting down a hospital increases with the size of the hospital and decreases with the number of patients who would be harmed by not shutting it down. If the “excess” rate at a very good hospital equals the community average, we do not harm any fewer patients by diverting them into the community. Conversely, even if a hospital performs better than expectations on AMDE, patients may be better served by going to other community hospitals if the hospital in question has very poor expected performance.

Appendix C: Modeling

We extend the notation for the Bayesian model in Eq. (2). Denote the training data as

$\mathbf{y}_{-\tau}^* = (y_{11}^*, \dots, y_{1\tau-1}^*, y_{21}^*, \dots, y_{j\tau-1}^*)'$. The global parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \Sigma)'$ and the

hospital-level parameters are $\mathbf{b} = (\mathbf{b}'_1, \dots, \mathbf{b}'_J)'$. By Bayes rule, the joint posterior distribution of the parameters given the training data is

$$f(\theta, \mathbf{b} | \mathbf{y}^*_{-\tau}) \propto f(\mathbf{y}^*_{-\tau} | \theta, \mathbf{b}) f(\mathbf{b} | \theta) \pi(\theta) \quad (1)$$

where $f(\mathbf{y}^*_{-\tau} | \theta, \mathbf{b})$ is the likelihood from (2), $f(\mathbf{b} | \theta)$ is the random effect distribution, and $\pi(\theta)$ is the prior.

We assume that the likelihood governing the safety signal is the same in the training and test years, $f(\mathbf{y}^*_{-\tau} | \theta, \mathbf{b})$, and we integrate over the posterior distribution of the parameters to obtain the posterior predictive distribution of the safety targets,

$$f(\tilde{\mathbf{y}}^*_{\tau} | \mathbf{y}^*_{-\tau}) = \int f(\tilde{\mathbf{y}}^*_{\tau} | \theta, \mathbf{b}) f(\theta, \mathbf{b} | \mathbf{y}^*_{-\tau}) d\theta d\mathbf{b}. \quad (2)$$

The safety signal $\boldsymbol{\delta} = (\delta_{1\tau}, \dots, \delta_{J\tau})'$ is the difference between the observed test year data $\mathbf{y}^*_{\tau} = (y^*_{1\tau}, \dots, y^*_{J\tau})'$ and the safety targets, $\tilde{\mathbf{y}}^*_{\tau} = (\tilde{y}^*_{1\tau}, \dots, \tilde{y}^*_{J\tau})'$. Thus the posterior distribution of the safety signal $\boldsymbol{\delta}$, $f(\boldsymbol{\delta} | \mathbf{y}^*_{-\tau})$, is a transformation of the posterior of the safety targets using the observed test year data. We fit the model using Markov chain Monte Carlo (MCMC) to obtain samples from (1) and (2). Therefore, it is straightforward to obtain samples from the posterior predictive distribution of the safety signal by subtracting posterior predictive samples from the corresponding observed $\mathbf{y}^*_{\tau}$.

Below is the `JAGS` model code used to fit our hierarchical Bayesian model with random slopes and intercepts to the transformed AMDE rates and produce test year predictions.

```
model{
  # Training model
  for (i in 1:N){
    R[i] ~ dnorm(mu[i], tau[i])
    mu[i] <- X[i,3:q1]%%beta[3:q1] + b[u[i],1] + b[u[i],2]*X[i,2]
  }
  # Prediction model for holdout observations
  for (j in 1:N.h){
    Rpred.h[j] ~ dnorm(mu.h[j], tau.h[j])
    mu.h[j] <- X.h[j,3:q1]%%beta[3:q1] +
      b[u.h[j],1] + b[u.h[j],2]*X.h[j,2]
  }
  # Random effects distributions
  for (k in 1:J){
    b[k,1] ~ dnorm(beta[1], bPrec[1])
```

```

    b[k,2] ~ dnorm(beta[2],bPrec[2])
  }
  # Priors
  beta[1:q1] ~ dnmnorm(betaMean[1:q1],betaPrec[1:q1,1:q1])
  bPrec[1] <- 1/(bSig[1]*bSig[1])
  bPrec[2] <- 1/(bSig[2]*bSig[2])
  bSig[1] ~ dnorm(0,1)T(0,)
  bSig[2] ~ dnorm(0,1)T(0,)
}

```

Appendix D: Simulation details

Sampling from HCUP data

For each iteration of the simulation, we draw a random sample of hospitals from the HCUP data, which we index by $j = 1, \dots, J$. For all the years in which each sampled hospital contributed data to HCUP, $\mathcal{T}_j$, we take the observed covariates, $\mathbf{x}_{jt}$, i.e., median length of stay, urban location, etc., from HCUP. Not all hospitals have observations in HCUP in 2011, but we wish to compute the safety target and simulate safety performance for *every* sampled hospital in this test year. Thus we simulate covariate values for each sampled hospital in the test year, $\mathbf{x}_{j\tau}$, by adding random noise to the average of the hospital's covariates in the previous years (leaving hospital characteristics fixed, i.e., size, private, urban, teaching). Thus each sampled hospital has a complete set of covariates that include both observed values (for the years in which the hospital contributed data) and simulated values (in the test year), $\{\mathbf{x}_{jt}, t \in \mathcal{T}_j, \tau\}$. We also assign each sampled hospital a random slope and intercept b_{1j} and b_{0j} ; we use estimates of these parameters from the model fit to the real data.

Simulated safety performance

We simulate true (transformed) safety performance for each sampled hospital using a data-generating model that uses fixed coefficients β and the simulated hospital-specific parameters b_{0j} and b_{1j} :

$$\mu_{jt}^{true} = \mathbf{x}_{jt}\beta + b_{0j} + b_{1j}year_t.$$

Note that this quantity is unobservable in real data analysis, but we can use it as a gold standard benchmark in the simulation study. The simulated observable (transformed) safety performance is a random draw from a normal distribution centered at the true safety target:

$$y_{jt}^* \sim N(\mu_{j\tau}^{true}, 1/4n_{jt}).$$

The value of the true safety signal in the test year is the difference between the simulated observable performance and the true expected performance under the data-generating model.

$$\delta_{j\tau}^{true} = y_{j\tau}^* - \mu_{j\tau}^{true}.$$

The complete set of simulated data for each sampled hospital is $\{y_{jt}^*, \mathbf{x}_{jt}, t \in \mathcal{T}_j, \tau\}$.

Modeled targets

Next we consider *estimating* the safety target for each hospital by fitting models to the simulated data. This reproduces the process that occurs in real data analysis, in which the true safety target is not known. We first fit a full model that includes all the same predictors as the data-generating model to the simulated training data. This model produces M samples from the joint posterior distribution of β , b_{0j} and b_{1j} :

$$\{\beta^{full(m)}, b_{0j}^{full(m)}, b_{1j}^{full(m)}, m=1, \dots, M, j=1, \dots, J\}.$$

Combining these posterior samples with the simulated covariates in each hospital in the test year yields samples from the posterior distribution of the safety target,

$$\mu_{j\tau}^{full(m)} = \mathbf{x}_{j\tau} \beta^{full(m)} + b_{0j}^{full(m)} + b_{1j}^{full(m)} year_{\tau}$$

and thus of the safety signal,

$$\delta_{j\tau}^{full(m)} = y_{j\tau}^* - \mu_{j\tau}^{full(m)}.$$

We also fit a reduced model that is missing some of the predictors in β . This allows us to study the sensitivity of our methods to misspecification of the safety target model, specifically, the impact of excluding important predictors of hospital safety performance. We fit this reduced model to the simulated training data using only a subset $\mathbf{x}^{red} \subset \mathbf{x}$ of the predictors. This model produces M posterior samples from the distribution of the subset of coefficients $\beta^{red} \subset \beta$ corresponding to $\mathbf{x}^{red}$ as well as b_{0j} and b_{1j} :

$$\{\beta^{red(m)}, b_{0j}^{red(m)}, b_{1j}^{red(m)}, m=1, \dots, M, j=1, \dots, J\}.$$

Combining these posterior samples with the simulated covariates in the test year yields the posterior distribution of the safety target,

$$\mu_{j\tau}^{red(m)} = \mathbf{x}_{j\tau}^{red} \beta^{red(m)} + b_{0j}^{red(m)} + b_{1j}^{red(m)} year_{\tau}$$

and thus of the safety signal,

$$\delta_{j\tau}^{red(m)} = y_{j\tau}^* - \mu_{j\tau}^{red(m)}.$$

Figure 1 displays the coefficients of the full and reduced models, which are the same as the random slopes and intercepts model fit to the ADME data except for four interactions added to the full model (and omitted from the reduced model). (Note that all of the predictors are standardized, so each coefficient represents the impact of a 1 standard deviation change in the predictor on the transformed AMDE rate.) Two of these interactions increase the expected AMDE rate with 1) the interaction of the median chronic condition count and percent of pediatric admissions among all admissions and 2) specialty designation, defined as teaching hospitals with a fraction of pediatric admissions at least two standard deviations above average. The other interaction coefficients decrease the rate of AMDE with 1) the interaction of length of stage and median age and 2) the combination of urban and teaching status.

Figure 1.
Coefficients of the full and reduced simulation models

Loss functions

We assume that there exists a true loss function that correctly parameterizes the true preferences of a hypothetical decision-maker, $\ell^{true}(a/\delta_{j\tau})$. That is, for any possible action a ,

$\ell^{true}(a|\delta_{j\tau}^{true})$ correctly captures the decision maker's costs and benefits. In real applications, this function is unobservable, but in our simulation study, it serves as the ground truth.

The true loss functions for the simulation largely replicate the applied data loss functions, with a few modifications. For No Action, the loss function remains the same, constant 0 for negative signals and increasing for positive signals:

$$\ell(a=0|\delta)=c^{harm}\delta^+, \quad (1)$$

Investigations are more costly to the regulator if the facility is isolated ($I_{iso}=1$ if the hospital is not urban and 0 otherwise) and the investigation produces information on all the cases to date,

$$\ell(a=1|\delta)=(1+c^{iso}I_{iso})c^{reg}-c^{info}\sqrt{\min(.05n,y)}+c^{harm}\delta^+. \quad (2)$$

Audits and remediation have the same added cost for isolated facilities, but are otherwise the same as above,

$$\ell(a=2|\delta)=(1+c^{iso}I_{iso})c^{reg}+c^{audit}\log n/3-c^{info}\sqrt{n/3}+c^{harm}\delta^+, \quad (3)$$

and

$$\ell(a=3|\delta)=(1+c^{iso}I_{iso})c^{reg}+c^{remed}\log n+c^{harm}0.5\delta^+. \quad (4)$$

Shutdowns include an extra penalty for children's specialty hospitals c^{spec} ($I_{spec}=1$ for a teaching hospital with a large fraction of pediatric admissions and 0 otherwise),

$$\ell(a=4|\delta)=I_{spec}c^{spec}+c^{div}n+c^{harm}(\delta^{comm}-\delta). \quad (5)$$

We also consider a misspecified loss function, $\ell^{miss}(a|\delta_{j\tau})$ that does not correctly represent the true preferences of our hypothetical decision maker. It has both the wrong set of inputs and the wrong valuations of these costs and benefits. This loss function allows us to study the sensitivity of our methods to misspecification of the loss function, which is very likely in real applications.

Table 1 of Appendix B gives details of the true and misspecified loss functions. The major differences are that the misspecified loss function omits the additional penalty for closing specialty hospitals and does not impose additional regulator costs for isolated hospitals. In addition, the fixed regulator cost, patient harm cost, and information parameters are misspecified.

Choosing actions

The action that minimizes the risk using the true loss function and the true safety signal is, by definition, the true best action $a_{j\tau}^{risk,true,true}$. We define a Z score using the true safety signal, $Z_{j\tau}^{true} = 2\sqrt{n_{j\tau}}\delta_{j\tau}^{true}$, and cutpoints that produce the same distribution of actions as the true best minimum-risk rule. This produces the “best” Z score action for each hospital, $a_{j\tau}^{Z,true,true}$.

Analogous to the pair of actions described above, we define minimum-risk and Z score actions that use the modeled safety signals under various combinations of full and reduced models with true or misspecified loss functions. We use a superscripting $a_{j\tau}^{(method),(model),(loss)}$ to indicate the 1) method of the decision rule: minimum (risk) vs (Z) score, 2) model for the safety target: true vs (full) model vs (reduced) model, and 3) loss function: true vs (mis)specified. For example, the minimum-risk action that uses the full model to set the safety targets and the true loss function is denoted $a_{j\tau}^{risk,full,true}$. The remaining actions are $a_{j\tau}^{Z,full,true}$, $a_{j\tau}^{risk,red,true}$, $a_{j\tau}^{Z,red,true}$, $a_{j\tau}^{risk,full,mis}$, $a_{j\tau}^{Z,full,mis}$, $a_{j\tau}^{risk,red,mis}$, $a_{j\tau}^{Z,red,mis}$. For example, $a_{j\tau}^{risk,red,mis}$ is the action that minimizes the Bayes risk with the misspecified loss function and the reduced model safety signal $1/M \sum_{m=1}^M \ell^{mis}(a|\delta_{j\tau}^{red(m)})$. Similarly, $a_{j\tau}^{Z,red,mis}$ is the action that categorizes $Z_{j\tau}^{red} = 2\sqrt{n_{j\tau}}(y_{j\tau}^* - E(\mu^{red(m)}))$ according to cutpoints that produce the same overall distribution of actions as $a_{j\tau}^{risk,red,mis}$ across all $j = 1, \dots, J$.

Appendix E: Extra Figures

Figure 1. Coefficients (medians and 95% credible intervals) from a Bayesian mixed model with random intercepts and slopes for each hospital. All predictors are standardized so that the coefficient represents the change in the expected (transformed) AMDE rate for a 1 standard deviation change in the value of the predictor.

Figure 2.
Correlation among parameter posteriors from a Bayesian mixed model with random intercepts and slopes for each hospital.

Figure 3.
Normal q-q plot of standardized residuals from linear mixed model with random intercepts and slopes.

References

1. Ettinger WH, Hylka SM, Phillips RA, Harrison LH, Cry JA, Sussman AJ. When things go wrong: the impact of being a statistical outlier in publicly reported coronary artery bypass graft surgery mortality data. *American Journal of Medical Quality*. 2008; 23(2):90–5. [PubMed: 18245577]

2. Kowalczyk L, Smith S. Hospital halts heart surgeries due to deaths: High rate cited at Worcester facility. *The Boston Globe*. 2005 Sep 22.
3. Bates KE, Vetter VL, Li JS, Cummins S, Aguel F, Almond C, et al. Pediatric cardiovascular safety: Challenges in drug and device development and clinical application. *American Heart Journal*. 2012; 164(4):481–492. [PubMed: 23067905]
4. Hwang TJ, Kesselheim AS, Bourgeois FT. Postmarketing trials and pediatric device approvals. *Pediatrics*. 2014; 133(5)
5. Sutherell JS, Hirsch R, Beekman r RH. Pediatric interventional cardiology in the United States is dependent on the off-label use of medical devices. *Congenital Heart Disease*. 2010; 5(1):2–7. [PubMed: 20136851]
6. Jones HE, Spiegelhalter DJ. The identification of “unusual” health-care providers from a hierarchical model. *The American Statistician*. 2011; 65(3):154–163.
7. Spiegelhalter D, Grigg O, Kinsman R, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *International Journal for Quality in Health Care*. 2003; 15(1):7–13. [PubMed: 12630796]
8. Matheny ME, Normand SLT, Gross TP, Marinac-Dabic D, Loyo-Berrios N, Vidi VD, et al. Evaluation of an automated safety surveillance system using risk adjusted sequential probability ratio testing. *BMC Medical Informatics and Decision Making*. 2011; 11:75. [PubMed: 22168892]
9. Bronskill SE, Normand SLT, Landrum MB, Rosenheck RA. Longitudinal profiles of health care providers. *Statistics in Medicine*. 2002; 21(8):1067–88. [PubMed: 11933034]
10. Paddock SM. Statistical benchmarks for health care provider performance assessment: A comparison of standard approaches to a hierarchical Bayesian histogram-based method. *Health Services Research*. 2014; 49(3):1056–1073. [PubMed: 24461071]
11. Kalbfleisch JD, Wolfe RA. On monitoring outcomes of medical providers. *Statistics in Biosciences*. 2013; 5(2):286–302.
12. Shahian DM, Normand SLT, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Annals of Thoracic Surgery*. 2001; 72(6):2155–68. [PubMed: 11789828]
13. Berger, JO. Statistical decision theory and Bayesian analysis. Springer Science & Business Media; 2013.
14. Soares MO, Dumville JC, Ashby RL, Iglesias CP, Bojke L, Adderley U, et al. Methods to assess cost-effectiveness and value of further research when data are sparse: negative-pressure wound therapy for severe pressure ulcers. *Medical Decision Making*. 2013; 33(3):415–36. [PubMed: 22927694]
15. Bennette CS, Veenstra DL, Basu A, Baker LH, Ramsey SD, Carlson JJ. Development and Evaluation of an Approach to Using Value of Information Analyses for Real-Time Prioritization Decisions Within SWOG, a Large Cancer Clinical Trials Cooperative Group. *Medical Decision Making*. 2016; 36(5):641–51. [PubMed: 27012232]
16. Healthcare Cost and Utilization Project (HCUP). HCUP Nationwide Inpatient Sample (NIS). Agency for Healthcare Research and Quality; 2011. www.hcup-us.ahrq.gov/nisoverview.jsp
17. Brady PW, Varadarajan K, Peterson LE, Lannon C, Gross T. Prevalence and nature of adverse medical device events in hospitalized children. *Journal of Hospital Medicine*. 2013; 8(7):390–3. [PubMed: 23744814]
18. Samore MH, Evans RS, Lassen A, Gould P, Lloyd J, Gardner RM, et al. Surveillance of medical device-related hazards and adverse events in hospitalized patients. *JAMA*. 2004; 291(3):325–34. [PubMed: 14734595]
19. Hwang W, Weller W, Ireys H, Anderson G. Out-of-pocket medical spending for care of chronic conditions. *Health Affairs*. 2001; 20(6):267–78.
20. Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, Blunt I, Wood C, Grigg O. Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society, Series A*. 2012; 175(1):1–47.
21. Kulinskaya E, Morgenthaler S, Staudte RG. Variance stabilizing the difference of two binomial proportions. *The American Statistician*. 2010; 64(4):350–356.

22. Böhning D, Viwatwongkasem C. Revisiting proportion estimators. *Statistical Methods in Medical Research*. 2005; 14(2):147–69. [PubMed: 15807149]
23. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 3rd International Workshop of Distributed Statistical Computing (DSC 2003); Vienna. 2003. vol. ISSN 1609-395X
24. Plummer, M. rjags: Bayesian graphical models using MCMC. R package version 4-5. 2016. Available from: <https://CRAN.R-project.org/package=rjags>
25. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*. 1998; 7(4):434–455.
26. Spiegelhalter DJ, Best NG, Carlin BP, Linde AV. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*. 2002; 64(4):583–639.
27. Carlin, BP., Louis, TA. *Bayesian Methods for Data Analysis*. 3rd. Boca Raton: Chapman & Hall, CRC; 2009.
28. Bardsley M, Spiegelhalter DJ, Blunt I, Chitnis X, Roberts A, Bharania S. Using routine intelligence to target inspection of healthcare providers in England. *Quality & Safety in Health Care*. 2009; 18(3):189–94. [PubMed: 19468000]
29. Gillis KD, Hixson JS. Efficacy of Statistical Outlier Analysis for Monitoring Quality of Care. *Journal of Business & Economic Statistics*. 1991; 9(3):241–252.
30. Berlowitz DR, Anderson JJ, Ash AS, Brandeis GH, Brand HK, Moskowitz MA. Reducing random variation in reported rates of pressure ulcer development. *Medical Care*. 1998; 36(6):818–25. [PubMed: 9630123]
31. Wang C, Hefflin B, Cope JU, Gross TP, Ritchie MB, Qi Y, et al. Emergency department visits for medical device-associated adverse events among children. *Pediatrics*. 2010; 126(2):247–59. [PubMed: 20660543]

Figure 1. Examples of loss functions for hypothetical hospitals of different sizes. The safety signal is measured as the number of AMDE admissions above the target, scaled as % of total pediatric admissions.

Figure 2. One-sided 95% posterior prediction intervals (line segments) and medians (tick marks) compared to observed test year (2011) observations (open symbols) from a Bayesian mixed model with random intercepts and slopes for each hospital. Total pediatric admissions in the test year are shown on the y-axis.

Figure 3. Distribution of posterior risks for actions in the test year (2011) chosen by a minimum risk rule (left) and a Z score rule (right). Each point represents the posterior risk of the action for a hospital chosen under the two different rules; the mean is given in red and a smoothed density is overlaid as a solid line. The risks for each action are computed under the loss function used in the minimum risk rule. The Z score thresholds are set to deliver the same distribution of actions as the minimum risk rule.

Figure 4. Excess risk relative to risk of minimum-risk action chosen using the true safety target and true loss function (i.e., gold standard).

Table 1

Selected percentiles of hospital-year variables in the training data. Denominators are pediatric admissions except the percent of pediatric admissions, which uses total admissions.

	25th %ile	Median	75th %ile
Median age (years)	11	14	17
Median length of stay (days)	3	3	3
Pediatric admissions (%)	5	8	12
Emergency admissions (%)	25	38	49
Government insurance (%)	35	51	64
Male (%)	32	39	46
Any chronic condition (%)	26	36	47
Multiple chronic conditions (%)	5	8	14
Died in hospital (per 10,000)	0	0	22
AMDE (per 10,000)	0	14	44

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Percent of test year hospitals with each action according to the minimum Bayes risk (rows) or Z score (columns) decision rules. Differences from sum-to-100 are due to rounding.

	No action	Investigate	Audit	Remediate	Shutdown	Bayes Total
No action	45.8	15.5	3.2	0.4	0.0	64.9
Investigate	13.9	5.2	2.8	0.4	0.0	22.3
Audit	5.2	1.2	0.4	0.8	0.0	7.6
Remediate	0.0	0.4	0.8	2.8	0.4	4.4
Shutdown	0.0	0.0	0.4	0.0	0.4	0.8
Z Score Total	64.9	22.3	7.6	4.4	0.8	100.0