medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Hospital profiling using Bayesian decision theory

Johannes Hengelbrock[1], Johannes Rauh[1], Jona Cederbaum[1], Maximilian Kähler[1], Michael Höhle[1,2]

[1] Federal Institute for Quality Assurance and Transparency in Healthcare, Berlin.

[2] Department of Mathematics, Stockholm University, Sweden

Abstract

Background

For evaluating the quality of care provided by hospitals, special interest lies in the identification of performance outliers. We study a setting where the decision to classify hospitals as performance outliers or non-outliers is based on the observed result of a single binary quality indicator.

Methods

We propose to embed the classification of providers into a Bayesian decision theoretical framework which enables the derivation of optimal decision rules with respect to the expected decision consequences. We argue that these consequences depend upon for which pathway to quality improvement the profiling of hospitals takes place. We propose paradigmatic utility functions for the two pathways *external reporting* and *change in care delivery* and compare the resulting optimal decision rules with regard to their threshold values, sensitivity and specificity. We further apply them to the area of hip replacement surgeries, for which we re-evaluate hospital performances for five quality indicators in Germany during 2017.

Results

Based on the utilities we assigned to the classification decisions, the decision rule for change in care delivery classifies more high-volume providers as outliers compared to the decision rule for external reporting, with consequences for both sensitivity and specificity. The re-evaluation of the five quality indicators illustrates that classification decisions are highly dependent upon the underlying utilities.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Conclusion

Analyzing the classification of hospitals as a decision theoretic problem and considering pathway-specific consequences of decisions can help to derive an appropriate decision rule. Contrasting decision rules with regard to their underlying assumptions about the utilities of classification consequences can be helpful to make implicit assumptions transparent and justifiable.

Keywords: hospital profiling, quality assurance, quality of care, Bayesian decision theory

Word count: 4,978

Introduction

The evaluation of the quality of care that patients receive from hospitals or other healthcare providers has gained increasing public and political attention (1,2). Measurements of quality of care are used for *external reporting* as well as for initiating *change in care delivery* (in the follwing referred to as *change*) with the common aim of quality improvement (3)[1]. For these purposes, special interest lies in the identification of performance outliers, either compared to a national average or a pre-specified target (5,6). For estimating and classifying hospital performances based on observed results, various frequentist and Bayesian statistical methods have been proposed and compared (5,7–10). However, little attention has been paid to the question of how much statistical evidence is needed in order to classify a hospital as performance outlier. Often, simply a pre-specified level of statistical significance is used as classification threshold, usually without relation to the consequences of the classification decision (9,11). Extending the work of (8) who compares different Bayesian methods for hospital profiling, we propose to explicitly define a utility function that quantifies the expected utility associated with a classification as outlier in relation to that as non-outlier. By formulating the classification as a decision problem, optimal decision rules can be derived in a Bayesian framework that are sensitive to prior information as well as to the pathway-specific utility function.

In the following, we first embed the profiling of hospitals into a general Bayesian decision theory framework and propose two alternative ways of specifying a utility function: one for the pathway of external reporting and one for that of change. Using an exemplary quality indicator, we then illustrate how the optimal decision rule depends on the utility function. In order to illustrate the practical consequences, we apply the framework to data on hip replacement surgeries that was collected for hospital profiling in Germany during 2017 and re-evaluate five indicators of quality of the mandatory German hospital profiling program. Finally, we discuss the insights such an formal decision theoretic view can provide for the practical process of the evaluation of quality of care, as well as its limitations.

[1] External reporting is sometimes also referred to as *accountability* or *selection* and change as *improvement* (3,4)

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

As starting point, we assume that a binary *quality indicator* is used to measure a certain aspect of the quality of care provided by a hospital (or some other healthcare provider), for which data is collected and analyzed. Typical examples include the adherence to specified processes (12,13) or outcome indicators such as 30-day mortality after some procedure (14). Data is collected for a pre-specified time period and the population of interest are all patients fulfilling certain inclusion criteria defined for the indicator within the selected time period. In the following, O_j denotes the binary outcome of interest of patient $j, j = 1, \ldots, J$, at a specific hospital. In the case of an annual evaluation of the quality of care, for instance, J is the number of patients which are considered as part of the quality indicator in the year of interest. We assume that the outcome variables are independent and identically distributed Bernoulli variables with outcome probability $\mathrm{P}(O_j = 1) = \theta$. This implies that individual patient characteristics are assumed to have no relevant effect on the outcome of interest. Our framework can however be extended to risk adjusted quality indicators with patient-specific event probabilities. In the following, O denotes the sum of the outcome variables of interest, $O = \sum_{j=1}^{J} O_j$, that is the number of events of interest at the specific hospital. By assumption, $O|J, \theta \sim \mathrm{Bin}(J, \theta)$.

In the case of non-risk adjusted quality indicators, the evaluation of hospital performance is based on the comparison of an estimate of the outcome probability θ with one or multiple pre-specified target values, which are the same for all providers. HIn our work, we focus on the case of binary classification based on a single pre-specified target value R, but the framework can easily be extended to multiple categories. In the binary case, hospitals are often classified into *performance outliers* and *non-outliers*, respectively (9). Furthermore, we assume that the target value R is the upper bound of a reference range and that the goal is to identify hospitals with true underlying values of θ above the target as performance outliers. We assume in the following that the target value R is a specified value known before the start of the data collection (2). Alternatively, R can also be an estimated mean or some quantile of the observed indicator results of several providers (5). Interest lies in the underlying value θ (opposed to just the observed rate o/J with $O = o$), because we are interested in the inherent quality processes at the hospital. This view corresponds to a so called analytic analysis framework in the spirit of (15).

The classification decision dividing hospitals into performance outliers and non-outliers can be conceptually formulated as an influence diagram (16,17), in which nodes represent random variables, arrows indicate conditional dependencies of the joint probability distribution and thus the absence of arrows indicates conditional independence. Figure 1 illustrates the case of a non-risk adjusted quality indicator operationalized as a proportion:

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

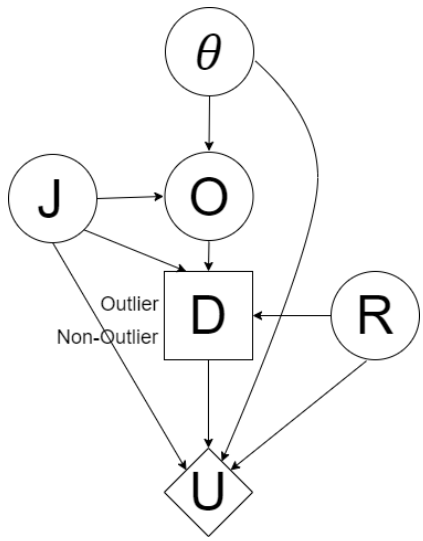

Figure 1: Influence diagram for the classification of a single hospital

$D(\cdot)$ denotes the decision to classify a hospital into one of the two categories $D \in \{\text{outlier, non-outlier}\}$, and U represents the utility function. Note that the number of patients J and the latent parameter of interest θ are assumed to be conditionally independent; this assumption can be relaxed, if necessary. From Figure 1, it follows that the classification decision depends upon the parameter of interest θ, the number of treated patients J, the observed number of outcomes of interest, the target value R as well as the utility function.

In a Bayesian context, for all random variables without parent nodes in the graph, a prior distribution has to be specified, reflecting all relevant prior knowledge. Since the decision D takes into account R and J, we may assume that R and J are not random. Thus, formally, we use point mass priors for R and J. Therefore, only a prior distribution for θ is really needed.

Given a treated number of patients J and an observed number of events of interest $O = o$, an optimal decision rule for the classification of a single hospital can be derived by maximizing the expected utility associated with the two classification decisions. This leads to a Bayes optimal decision rule (18):

$$\begin{aligned} d^*(o, J, R, \theta) &= \operatorname*{argmax}_{d \in D} \mathrm{E}[U(d, J, R, \theta)] \\ &= \operatorname*{argmax}_{d \in D} \int_0^1 U(d, J, R, \theta) f(\theta | o, J) \,\mathrm{d}\,\theta. \end{aligned}$$

Here, $f(\theta|o, J)$ denotes the density of the posterior distribution of θ given o, J and a prior distribution of θ.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

For the simultaneous classification of multiple providers $i = 1, \ldots, I$, the Bayes optimal decision rule is a vector of length I, denoted as $\boldsymbol{d}^*()$. For multiple providers, we assume an additive utility function

$$\sum_{i=1}^{I} U(d_i, o_i, J_i, R, \theta_i),$$

with the expected utility of each individual decision satisfying $\mathrm{E}[U(d_i, o_i, J_i, R, \theta_i)] \geq 0$. Additivity of utilities is a simplifying assumption that does not necessarily hold in all decision contexts. In the discussion, we refer to situations in which simple additivity may not suitable.

The Bayes optimal decision rule is then defined as

$$\boldsymbol{d}^*(o_i, J_i, R, \theta_i) = \underset{d_1,\ldots,d_I}{\operatorname{argmax}} \sum_{i=1}^{I} \mathrm{E}[U(d_i, o_i, J_i, R, \theta_i)].$$

Because each term in the summation is only a function of the expected utility for one single provider i, this can be written as

$$\boldsymbol{d}^*(o_i, J_i, R, \theta_i) = \left[\underset{d_1 \in D}{\operatorname{argmax}}\, \mathrm{E}[U(d_1, o_1, J_1, R, \theta_1)], \ldots, \underset{d_I \in D}{\operatorname{argmax}}\, \mathrm{E}[U(d_I, o_I, J_I, R, \theta_I)]\right].$$

In other words, the optimal decision rule is equal to the assemblance of all individually optimal decisions, which is why we concentrate on the expected utility of a single provider hereafter.

Consequences of classification decisions

Besides the specification of a prior distribution for θ and the choice of a target value R, the specification of a utility function is a central aspect for deriving an optimal decision rule. For the definition of utilities associated with different decisions, a time period for which decision consequences are assumed to have an effect needs to be defined. For our application to the two pathways, we assume that profiling of hospitals is carried out on a yearly basis and therefore consider the consequences of the decision to be effective only until the subsequent classification. Here and in the following, the term *utility* is used as a generic term for all positive effects of a decision compared to the effects of competing decisions, including financial consequences, transparency, good quality (or improvement) of health care, and so on.

According to the influence diagram (Fig. 1), the utility U may depend on the case load J, the parameter θ, the reference value R and the decision D. It is convenient to represent U as a 2 x 2 matrix with rows corresponding to the decision D and columns distinguishing whether θ lies below or above R.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

		$\theta \leq R$	$\theta > R$
Classification	Non-Outlier	$\tilde{U}_{0,0}(J,\theta,R)$	$\tilde{U}_{0,1}(J,\theta,R)$
	Outlier	$\tilde{U}_{1,0}(J,\theta,R)$	$\tilde{U}_{1,1}(J,\theta,R)$

Therefore, the optimal decision, which optimizes the expected utility, is given by:

$$d(o,J,R) = \mathrm{I}(\mathrm{E}[\tilde{U}(d=0,J,\theta,R)|o,J,R] < \mathrm{E}[\tilde{U}(d=1,J,\theta,R)|o,J,R]), \tag{1}$$

where $\mathrm{I}(\cdot)$ is a 0-1 indicator function, where the value 1 is interpreted as the decision *outlier*. Following from that, it is mainly the difference $\tilde{U}(d=1,J,\theta,R) - \tilde{U}(d=0,J,\theta,R)$ that is important for the classification decision. In other words, when arguing about the optimal decision rule, we may normalize the utility functions and assume that within each column of the above matrix, one of the two entries is equal to zero. Usually, if the reference value R is well aligned with the classification problem, the utility of a correct classification should be larger than the utility of a wrong classification.

$$\tilde{U}_{0,0}(J,\theta,R) \geq \tilde{U}_{1,0}(J,\theta,R) \quad \text{and} \quad \tilde{U}_{1,1}(J,\theta,R) \geq \tilde{U}_{0,1}(J,\theta,R).$$

Together with the normalization, we obtain a utility matrix of the following form:

		$\theta \leq R$	$\theta > R$
Classification	Non-Outlier	$U_{0,0}(J,\theta,R)$	0
	Outlier	0	$U_{1,1}(J,\theta,R)$

where $U_{0,0}$ and $U_{1,1}$ should be positive, at least when θ is far away from R.

If $U_{0,0}$ and $U_{1,1}$ do not depend on θ, then it is possible to express the optimal decision rule (1) in terms of the posterior probability $\mathrm{P}(\theta \leq R|o,J)$:

$$\begin{aligned} d(o,J,R) &= \mathrm{I}(U_{0,0}(J,R)\cdot \mathrm{P}(\theta \leq R|o,J) < U_{1,1}(J,R)\cdot \mathrm{P}(\theta > R|o,J)) \\ &= \mathrm{I}\left(\mathrm{P}(\theta \leq R|o,J) < \frac{U_{1,1}(J,R)}{U_{1,1}(J,R)+U_{0,0}(J,R)}\right). \end{aligned} \tag{2}$$

Assigning specific values to the utilities associated with classifications decisions is difficult, because the consequences are manifold and differ between the affected actors (such as patients, payers, hospitals

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

and others). However, every decision rule either explicitly or implicitly relies on a consideration of those utilities and making those considerations explicit can lead to more transparent assessments. Below, we give examples of how the functions $U_{0,0}$, $U_{1,1}$ can be chosen to model simple situations. For doing so, we consider two different pathways to improve quality of care based on measurements: external reporting and change. For both, we do not aim at deriving generally valid utility functions. Instead, we illustrate how different assumptions on the functional dependencies of the utilities from θ and J affect the choice of an optimal decision rule.

External reporting In the pathway of external reporting, measurements of quality of care are used to reveal and compare the performance of hospitals (4). In general, external reporting and the classification of hospital performances can help to guide patient decision regarding their hospital choice (19) or support decisions for financial rewards or punishment (3,13). For simplicity, we focus in the following on external reporting with binary consequences such as accreditation of hospitals or selection for a financial reward or punishment and thus assume that the utilities are independent from the underlying quality of care of the hospital, θ. We allow the utilities to depend upon the number of future patients J' because it is reasonable to assume that overall, decisions have a stronger effect when they affect more future patients. As described above, we are only interested in the decision consequences that have an effect until the subsequent hospital classification. Thus, J' refers to the number of treated patients within the next year. For this unknown quantity, usually the current number J can serve as a good estimate and will be used in the following.

Thus, one possible way to define the associated utility in the pathway of external reporting is displayed in Table 1:

		$\theta \leq R$	$\theta > R$
Classification	Non-Outlier	$\eta \cdot J$	0
	Outlier	0	$\mu \cdot J$

Table 1: Utility in the pathway of external reporting

with $\mu, \eta \geq 0$. As described above, $U_{0,0} = \eta \cdot J$ represents the utility of a correct non-outlier classification relative to that of an incorrect outlier classification. Thus, utilities arising independently of the classification decision (like the costs of collecting and analyzing data) are not included in $U_{0,0}$.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

According to (2), the optimal decision rule is given by

$$d_{\text{report}}(o, J, R, \theta) = \mathrm{I}\left(P(\theta \leq R|o, J) < \frac{\mu}{\mu + \eta}\right).$$

It is straightforward to show that the decision rule that we derive below is also optimal in the case in which $U_{0,0}$ and $U_{1,1}$ depend on J in the same functional way, in the sense that the ratio $U_{0,0}/U_{1,1}$ is independent of J. Therefore, we omit J as factor from Table 1 in the following.

Thus, the optimal decision rule weighs the probability of the healthcare provider being a performance outlier against the relation of expected utilities $\mu/(\mu + \eta)$, in the following referred to as α. If, for instance, hospitals are classified as performance outliers if the probability of them being an outlier is less than $\alpha = 5\%$, this corresponds to a utility ratio of correct outlier to correct non-outlier classifications of 1 to 19.

Change in care delivery In the pathway of change, classifications of hospitals usually intend to target resources for improvement and change processes for those providers where they have the most impact. Because the goal is to improve the quality of care of the hospitals under consideration, those who benefit most from change are future patients treated by those hospitals. In order to quantify the expected benefit, assumptions about the effects of change are necessary. These can either be made based on the evaluation of earlier actions to improve the quality of healthcare or, if no prior information is available, have to be based on a best guess. If no prior evidence is available, we suggest the following: We assume that measures of change can align the parameter of interest θ to the target value R, resulting in the improvement factor $(\theta - R)_+ := \max(0, \theta - R)$. Thus, a correct outlier classification (compared to the false classification as a non-outlier) is associated with a utility depending on the potential for improvement through change, the number of patients that benefit from these improvements and some multiplication factor δ: $\delta \cdot (\theta - R)_+ \cdot J'$, with $\delta \geq 0$. Again, we use J as an approximation for J'.

The utility of the classification as non-outlier (compared to as outlier), on the other hand, consists of the resources that are saved given that no measures of change are implemented (denoted as ρ). This applies regardless of whether the classification was correct or incorrect.

Following from that, one reasonable way to quantify the utilities associated with the classification decision for the pathway of change is displayed in Table 2, in which the utility of a correct outlier classification includes the resources that are necessary for the implementation of the measures of change (ρ):

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

		$\theta \leq R$	$\theta > R$
Classification	Non-Outlier	ρ	0
	Outlier	0	$\delta \cdot (\theta - R)_+ \cdot J - \rho$

Table 2: Utility in the pathway of change

with $\rho, \delta \geq 0$. Based on this, the optimal decision rule can be derived as:

$$d_{\text{change}}(o, J, R, \theta) = \mathrm{I}\left(\rho < \int_0^1 \delta \cdot (\theta - R)_+ \cdot J \cdot f(\theta, o, J)\,\mathrm{d}\,\theta\right).$$

Again, not the absolute utilities but only the relation of ρ and δ, in the following referred to as $\rho = \rho/\delta$, has to be defined in order to derive a utility-efficient decision rule. Here, ρ can be interpreted as the relation of utilities between those of saving resources by not implementing measures of change for one hospital and those of its future patients benefiting from measures of change. In contrast to the pathway of external reporting, the utilities of change are assumed to be directly dependent upon the outcome probability θ. In this case, it is not possible to give a simple expression for d_{change} as a function of posterior probabilities.

Comparison of different decision rules

In order to derive optimal decision rules d_{report} and d_{change}, the expected utilities associated with a classification of a hospital as performance outlier have to be set into relation to those associated with a classification as non-outlier. For the paradigmatic utility functions for the two pathways described above, this is controlled by the two parameters α and ρ, respectively. Given those and a fixed number of treated patients J, critical values $o_{\min}$ for the observed number of outcomes, o, can be calculated which lead to the classification as performance outlier:

$$o_{\min}(J, R, L) = \underset{0 \leq o \leq J}{\operatorname{argmin}}\{d(o, J, R, L) = 1\}.$$

Hospitals with observed $o \geq o_{\min}$ are classified as performance outliers, where the concrete value of $o_{\min}$ is pathway-specific.

To illustrate the effects of the hospital volume (J) and the choice of a decision rule (d_{report} and d_{change}), the critical values $o_{\min}$ are displayed in a funnel plot (5). As an example, we use a exemplary quality indicator with a target value (up to) $R = 15\%$. Because we assume no prior knowledge about the

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

parameter, we use the non-informative Jeffrey's prior as prior $\text{Beta}(1/2, 1/2)$ distribution for θ.

In addition to the decision rules d_{report} and d_{change}, we analyze the decisions resulting from a naive decision rule:

$$d_{\text{naive}}(O = o, J, R) = \text{I}\left(\frac{o}{J} > R\right).$$

This naive rule has been traditionally used for classification decisions in Germany (20), which is why we compare its results to that of the decision rules described above.

We compare the three decision rules with respect to their critical values $o_{\min}$ as well as their sensitivity and specificity in three scenarios: with underlying quality of care within (*in control*: $\theta = 10\%$), outside (*out of control*: $\theta = 20\%$), and at the border of the reference area (*borderline*: $\theta = R = 15\%$). Sensitivity is here defined as the probability of identifying outlying hospitals as performance outliers (and applies thus only to the out-of-control scenario), and specificity as probability of identifying non-outliers as non-outliers (and applies to the in-control and borderline scenario). For each scenario, sensitivity and $1 -$ specificity are calculated for a varying number of patients treated by the hospital. Again, Jeffrey's prior is used.

We apply the above proposed decision rules to five quality indicators covering aspects of stationary hip replacement surgeries in Germany in the year 2017. For all hospitals, providing relevant services for these indicators, we calculate for each indicator the proportion of hospitals which are classified as performance outliers given various values for ρ and α and using Jeffrey's prior for θ. Because in Germany, d_{naive} is currently used as decision rule and all classifications of performance outliers are subsequently validated in a qualitative step, we can compare the positive-predictive-value (PPV) of the proposed decision rules for various values of ρ and α with respect to this qualitative assessment. The validation consists of a structured dialogue between the hospital and the responsible quality assurance agency and aims at determining whether the classification as outlier is really due to deficits in the quality of care (1,2). PPV is calculated as the proportion of hospitals whose classification as performance outlier is confirmed by the qualitative validation.

Results

Figure 2 shows the critical values $o_{\min}$ depending on J for an indicator with $R = 15\%$ and for specific values of ρ and α:

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

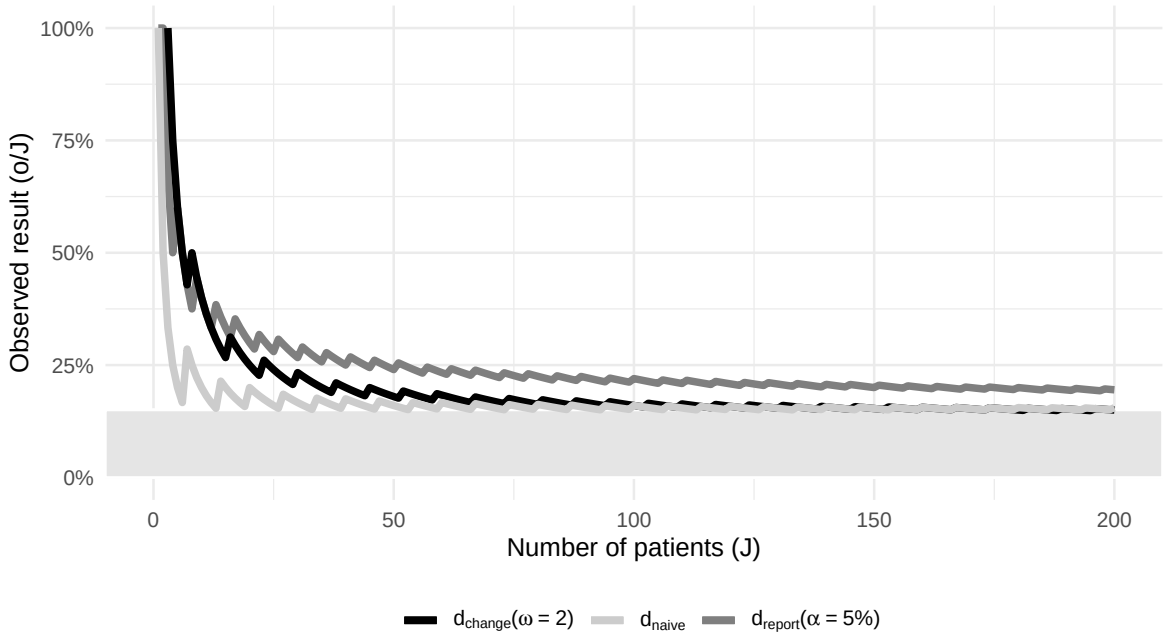

Figure 2: Funnelplot for a binary indicator with $R = 15\%$

With increasing number of patients J, the critical value of all three decision rules approaches the target area (grey area in Figure 2). For low-volume providers, d_{change} leads to more liberal decisions compared to d_{report}, with d_{naive} yielding the lowest critical values. On the other hand, d_{change} converges more quickly to R than d_{report} for higher values of J. This means that for the specific values $\alpha = 5\%$ and $\rho = 2$, low-volume hospitals are more often classified as performance outliers based on d_{report} compared to d_{change}, while the opposite is true for high-volume hospitals. The critical value of d_{naive}, in turn, is the most conservative over almost all J.

This is also reflected in the sensitivity and specificity of all three decision rules displayed in Figure 3:

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

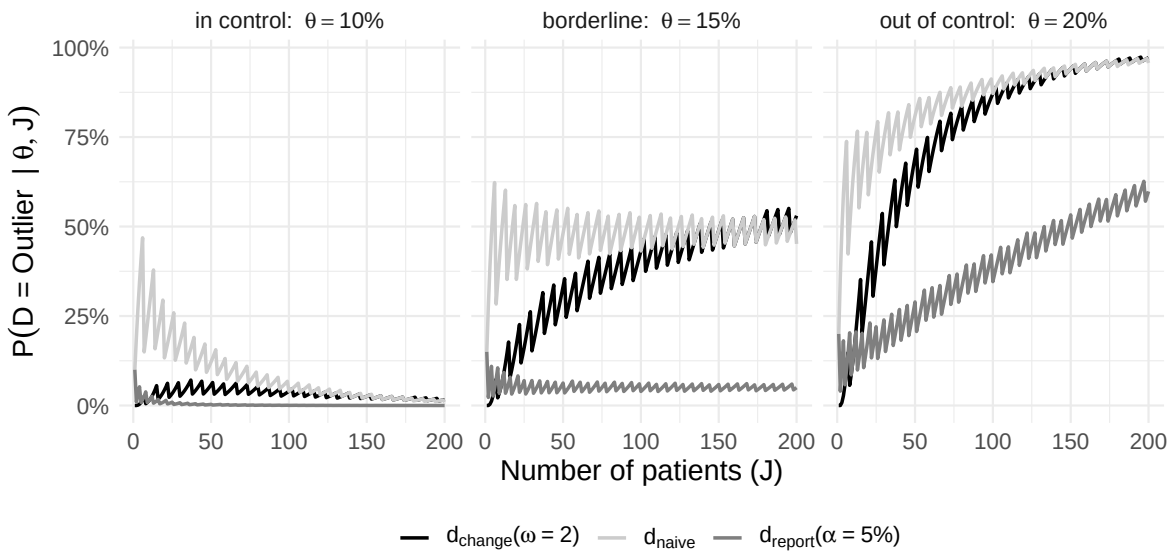

Figure 3: Sensitivity and specificity as a function of the number of cases J

For the in-control-scenario, the probability of a false-positive classification are low for both d_{report} and d_{change}, indicating a high specificity. For d_{naive}, however, the specificity is comparatively low for small values of J. For the borderline-scenario, the decision rule d_{report} for the pathway of external reporting has, on average, a probability of 5% for false-positive classifications, corresponding to a specificity of 95%. The decision rule for the pathway of change, d_{change}, has a higher specificity for low-volume providers. With increasing number of patients, however, the probability of false-positive classifications increases, due to the fact that the expected utility of false-negative classifications increase with J. This, on the other hand, means that d_{change} has a high sensitivity for high-volume providers, which can be seen in the out-of-control-scenario. There, the sensitivity of d_{report} also increases with higher numbers of patients, but not as quickly as that of d_{change}. For low-volume providers, however, the probability of false-negative classifications is again higher for d_{change}. For d_{naive}, the specificity is low if the true parameters lies within the reference area, but the sensitivity is high even for small-volume providers compared to both other decision rules if the true parameter lies outside of the reference area.

The above figures only show the consequences of single choices $\alpha = 5\%$ and $\rho = 2$ for the parameters, but the results are qualitatively similar for other values. Consequences of alternative choices for the parameters can be investigated via two Shiny apps (21).[2] Also for different choices of α and ρ, the general conclusion holds that for the pathway of change, outlier classifications are more frequent for high- compared to low-volume providers compared to the pathway of external reporting, with similar

[2]https://iqtig.shinyapps.io/funnel_plot/ and https://iqtig.shinyapps.io/sensitivity_specificity

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

consequences for both sensitivity and specificity as presented above.

Application to hospital profiling in Germany

In total, 1,277 hospitals were profiled based on data collected in 2017 for the five considered quality indicators for hip replacement surgeries. An overview over all indicators is given in Table 3:

QI	Quality indicator	Reference area
1	Patient has no indication for elective hip replacement surgery	$\leq 10\%$
2	Patient has no indication for hip revision surgery	$\leq 14\%$
3	Length of stay prior to surgery for hip fracture is > 48h	$\leq 15\%$
4	No measures for fall prevention taken	$\leq 20\%$
5	Patient is not mobile on discharge	$\leq 5\%$

Table 3: Quality indicators for hip replacement surgeries

Over all hospitals and indicators, 1,024 out of 6,035 hospital results were classified as performance outliers based on the naive decision rule d_{naive}. In the qualitative validiation of those outlier classifications, only 15% were subsequently confirmed as actual outliers.

For all hospitals and quality indicators, the number of hospitals that would have been classified as outliers using d_{report} and d_{change} is displayed for various values of ρ and α in Figure 4:

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

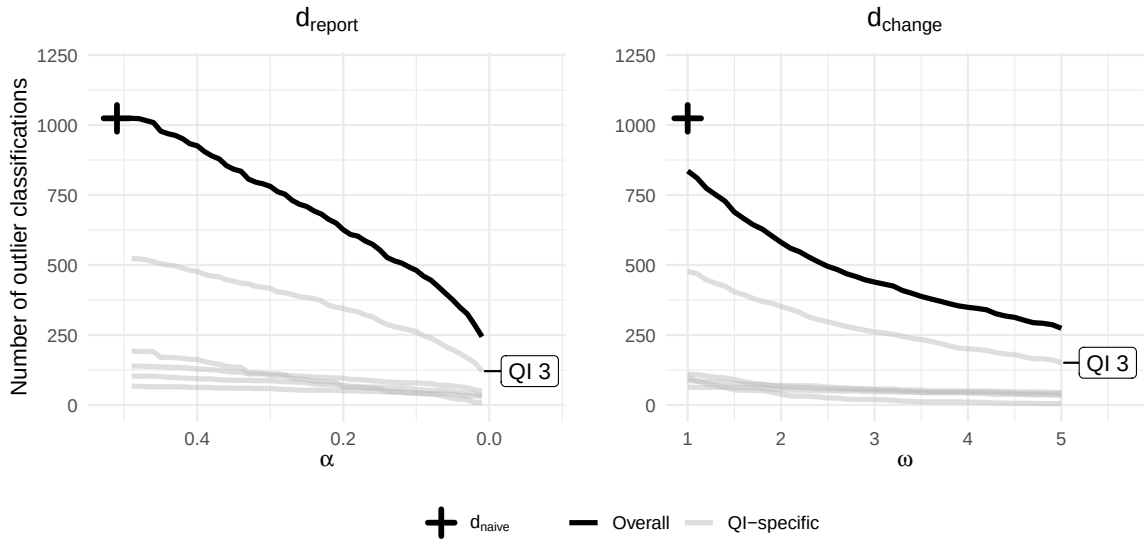

Figure 4: Number of outlier classifications as functions of the parameters α and ω

For all considered values, the naive decision rule is the most conservative, yielding the highest number of outlier classifications (1,024). For the two other decision rules, the number of outlier classifications decreases with decreasing α and increasing ρ, respectively. This is especially true for QI 3 (length of stay prior to surgery), for which the number of outlier classifications is generally the highest among all indicators. For all other indicators, the number of outlier classifications is lower but decreases similarly with decreasing α and increasing ρ, respectively.

Figure 5 displays the PPV, that is the proportion of outlier classifications that were confirmed as outliers in the qualitative validation step:

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

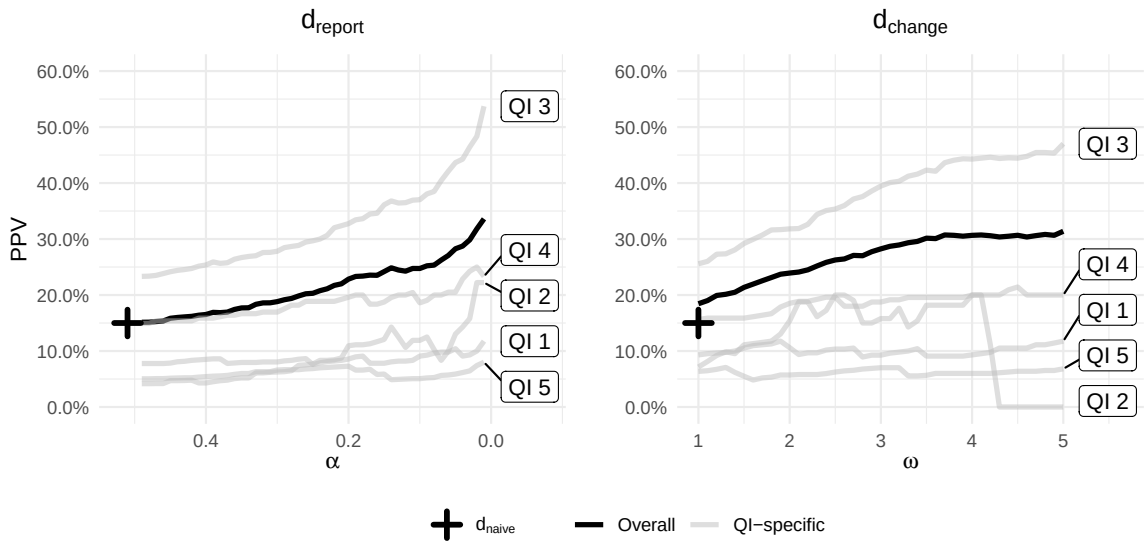

Figure 5: PPV as functions of the parameters α and ω

For both decision rules, the overall PPV is low for large values of α and low values of ρ but increases to up above 30% for both d_{report} and d_{change}. Indicator 3 (length of stay prior to surgery) is the indicator for which the PPV increases the most, from about 25% for $\rho = 1$ and $\alpha = .5$ and up to and above 50% for $\rho = 5$ and $\alpha = .01$. For all other indicators, there is at most a weak relationship between the PPV and the values of α and ρ.

Overall, even with conservative decision rules that assume much lower utility for false-positive compared to false-negative classifications, the PPV is relatively low across all hospitals and quality indicators. Only the PPV of a single indicators shows a strong relationship between the PPV and increasing or decreasing values of ρ and α, respectively.

Discussion

The classification of hospitals on the basis of measurements of process or outcome indicators is a decision under uncertainty, because the observed data only allows probabilistic statements about the true parameter of interest. Bayesian decision theory provides a framework to derive decision rules that are optimal with respect to the expected utility associated with each decision. Because the utilities are a crucial part of the decision rule, they should be reflected upon and incorporated into the decision rule wherever possible in order to make decisions transparent and justifiable. This also means that the decision rule should be pathway-specific rather than be based on standard rules for classification decisions that make (possibly not realistic) implicit assumptions about the underlying utilities. For the

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

profiling of hospitals, we suggest different utility functions for two pathways to quality improvement. For the pathway of external reporting, we assumed that only correct classifications are associated with positive consequences, which depend on the number of affected patients in a simple way but are independent of the underlying quality of care of the hospital. For the pathway of change, in contrast, we argued that the consequences of some classification decisions can be viewed as not only being dependent on the number of affected patients but also on the underlying quality of care. Based on the two differing utility functions, the proposed decision theoretic framework leads to different decision rules. In practice, arriving at a single utility function can be difficult, especially because consequences for different parties (such as patients, hospitals, sponsors, investigators, policy makers) have to be considered and weighted against each other. In these cases, it can be useful to define a set of fixed decision rules (based on values such as $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.4\}$ and some integer values of ρ, respectively) and, by doing so, to contrast different underlying utilities in order to arrive at a rule that suits best to the decision problem at hand among a set of options.

For the simultaneous classification of multiple hospitals, we assumed an additive utility function. In practice, the assumption of simple additivity may not always be a appropriate. For instance, the quality and effects of measures of change can be dependent upon the total number of outlier classifications due to limited resources. In this case, the utility of an outlier classifications directly depends upon the overall number of outlier classifications, which should be incorporated into the utility function. Furthermore, we ignored that the decision rule itself can have an effect on the quality of care provided by hospitals. If, for instance, small-volume hospitals below a certain number of treated cases irrespectively of their performance are never classified as outliers, the absence of quality assurance at these hospitals could lead to a decline in the quality of care. In this case, it can be reasonable to modify the decision rule in order to avoid undesired effects or use additional instruments.

Besides the specification of a utility function, the incorporation of prior knowledge is another crucial aspect in this Bayesian decision-theoretical framework. If a non-informative prior is used, the prior distribution has little effect on the posterior distribution of θ for middle- and high-volume providers. If multiple providers are profiled simultaneously, hierarchical Bayesian models are an alternative to let the prior distribution be informed by the observed data.

The application of the presented framework to the German setting of hospital profiling in the area of hip replacement surgeries illustrated that assumptions about the utility function have a high impact on the number of outlier classifications: the lower the utility of correct non-outlier in relation to correct outlier classifications are assumed to be, the more providers are classified as performance outliers. One consequence of a low utility for correct non-outlier in relation to correct outlier classifications (resulting in large values for α and small values for ρ, respectively) is a lower PPV of the classification decision.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

However, the relationship between PPV and values of α and ρ was pronounced only for one of the five quality indicator and weak or non-existent for all others. This is surprising, because one could assume that more liberal threshold values should lead to a higher PPV among those still classified as outliers. One reason for the weak relationship between PPV and values of α and ρ could be that the validation of results classified as outliers is a relatively informal process with plenty of heterogeneity and, hence, cannot serve as a gold standard for the classification of true outliers (20). Also, it is possible that the indicators' inclusion criteria are not accurate enough to exclude cases from the indicators which are explained by special circumstances. Either way, the weak relationship between the PPV and α and ρ could be reason to review the validity of the indicators or the qualitative validation of them.

The Bayesian framework discussed here can easily be extended to other types of quality indicators, including risk-adjusted indicators, the aggregation of several indicators or more complex decisions than a binary classification. Futhermore, more complex sequential decision frameworks can be employed to incorporate the effects of decisions on subsequential decisions or adaptive behaviours from hospitals.

References

1. Bramesfeld A, Wensing M, Bartels P, Bobzin H, Grenier C, Heugren M, et al. Mandatory national quality improvement systems using indicators: An initial assessment in Europe and Israel. Health Policy. 2016;120(11):1256–69.

2. Beaussier A-L, Demeritt D, Griffiths A, Rothstein H. Steering by their own lights: Why regulators across Europe use different indicators to measure healthcare quality. Health Policy. 2020;124(5):501–10.

3. Berwick D, James B, Coye M. Connections between quality measurement and improvement. Medical Care. 2003;41:I–30–I–38.

4. Solberg LI, Mosser G, McDonald S. The three faces of performance measurement: Improvement, accountability, and research. The Joint Commission Journal on Quality Improvement. 1997;23(3):135–47.

5. 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 (Statistics in Society). 2012;175(1):1–47.

6. Brakenhoff TB, Roes KCB, Moons KGM, Groenwold RHH. Outlier classification performance of risk adjustment methods when profiling multiple providers. BMC Medical Research Methodology. 2018;18:54.

7. Normand S-LT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

care: Issues and applications. Journal of the American Statistical Association. 1997;92(439):803–14.

8. Austin PC. A comparison of Bayesian methods for profiling hospital performance. Medical Decision Making. 2002;22(2):163–72.

9. Racz MJ, Sedransk J. Bayesian and frequentist methods for provider profiling using risk-adjusted assessments of medical outcomes. Journal of the American Statistical Association. 2010;105(489):48–58.

10. Austin PC. "The reliability and validity of Bayesian measures for hospital profiling: a Monte Carlo assessment". Journal of Statistical Planning and Inference. 2005;128(1):109–22.

11. Shahian DM, Normand S-LT. What is a performance outlier? BMJ Quality & Safety. 2015;24(2):95–9.

12. Ngantcha M, Le-Pogam M-A, Calmus S, Grenier C, Evrard I, Lamarche-Vadel A, et al. Hospital quality measures: Are process indicators associated with hospital standardized mortality ratios in French acute care hospitals? BMC Health Services Research. 2017;17(1):578.

13. Wang C, Su S, Li X, Li J, Bao X, Liu M. Identifying performance outliers for stroke care based on composite score of process indicators: An observational study in China. Journal of General Internal Medicine. 2020;35:2621–8.

14. Iezzoni LI. Risk adjustment for measuring healthcare outcome. Chicago: Health Administration Press; 2013.

15. Deming WE. On the distinction between enumerative and analytic surveys. Journal of the American Statistical Association. 1953;48(262):244–55.

16. Jensen FV, Nielsen TD. Bayesian networks and decision graphs. Second Edition. Berlin [u. a.]: Springer; 2007.

17. Pearl J. Causality: Models, reasoning, and inference. Second Edition. Cambridge, GB [u. a.]: Cambridge University Press; 2009.

18. Berger JO. Statistical Decision Theory and Bayesian Analysis. 2nd ed. Springer-Verlag; 1980.

19. Ash AS, Fienberg SE, Louis TA, Normand ST, Stukel TA, Utts J. Statistical issues in assessing hospital performance. Commissioned by the committee of presidents of statistical societies [Internet]. Centers for Medicare; Medicaid Services; Committee of Presidents of Statistical Societies; 2012. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Hospital QualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf

20. Hengelbrock J, Höhle M. Evaluating quality of hospital care using time-to-event endpoints based

medRxiv preprint doi: https://doi.org/10.1101/2021.06.23.21259367; this version posted June 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

on patient follow-up data. Health Services and Outcomes Research Methodology. 2019;19:197–214.

21. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. shiny: Web Application Framework for R [Internet]. 2021. Available from: https://shiny.rstudio.com/