Health Services and Outcomes Research Methodology (2019) 19:197–214
https://doi.org/10.1007/s10742-019-00202-7

Evaluating quality of hospital care using time-to-event endpoints based on patient follow-up data

Johannes Hengelbrock[1] · **Michael Höhle**[1,2]

Received: 23 October 2018 / Revised: 26 April 2019 / Accepted: 18 June 2019 / Published online: 8 July 2019

Abstract

Revisions of hip and knee arthroplasty implants and cardiac pacemakers pose a large medical and economic burden for society. Consequently, the identification of health care providers with potential for quality improvements regarding the reduction of revision rates is a central aim of quality assurance in any healthcare system. Even though the time span between initial and possible subsequent operations is a classical time-to-event endpoint, hospital-specific quality indicators are in practice often measured as revisions within a fixed follow-up period and subsequently analyzed by traditional methods like proportions or logistic regression. Methods from survival analysis, in contrast, allow the inclusion of all observations, i.e. also those with early censoring or events, and make thus more efficient and more timely use of the available data than traditional methods. This may be obvious to a statistician but in an applied context with historic traditions, the introduction of more complicated methods needs a clear presentation of their added value. We demonstrate how standard survival methods like the Kaplan–Meier estimator and a multiplicative hazards model outperform traditional methods with regard to the identification of performance outliers. Following that, we use the proposed methods to analyze 640,000 hip and knee replacement operations with about 13,000 revisions between 2015 and 2016 in more than 1200 German hospitals in the annual evaluation of quality of care. Based on the results, performance outliers are identified which are to be further investigated qualitatively with regard to their provided quality of care and possible necessary measures for improvement. Survival analysis is a sound statistical framework for analyzing data in the context of quality assurance and survival methods outperform the statistical methods that are traditionally used in this area.

Keywords Quality assurance · Quality of care · Survival analysis · Stochastic process control

✉ Johannes Hengelbrock
johannes.hengelbrock@iqtig.org

Michael Höhle
michael.hoehle@iqtig.org

[1] Federal Institute for Quality Assurance and Transparency in Healthcare (IQTIG), Berlin, Germany

[2] Department of Mathematics, Stockholm University, Stockholm, Sweden

1 Introduction

The occurrence of subsequent operations to replace or repair implanted prostheses or cardiac pacemakers and their temporal distance from the initial operation is an important indicator for the quality of both the initial implantation operation as well as the implantation product. Thus, revision rates are frequently used as a measure of quality, in registry studies usually with interest in product quality (McGrory et al. 2016; Delaunay 2015; Tarasevicius et al. 2014; Junnila et al. 2016). Our focus is on the context of quality assurance for the initial implantation operation (Bernatz et al. 2015; Benbassat and Taragin 2000). Revisions within a fixed follow-up period are most commonly analyzed using traditional methods like proportions or logistic regression (Mehrotra et al. 2014; Bosco et al. 2014; Kristoffersen et al. 2015). It has, however, also been suggested to make use of survival analysis methods for analyzing hospital-specific time-to-event endpoints such as Kaplan–Meier curves (Kristoffersen et al. 2015), proportional hazard models (He and Schaubel 2014) or CUSUM charts based on parametric or non-parametric time-to-event methods (Gandy et al. 2010; Oliveira et al. 2016). Compared to traditional methods, methods from survival analysis allow the inclusion of observations with incomplete follow-up times and, thus, make more efficient use of the available data. This may be obvious to a statistician but in an applied context with historic traditions, the introduction of more complicated methods needs a clear presentation of their added value. Therefore, we show in the following how standard methods like the Kaplan–Meier estimator and a multiplicative hazard model can be utilized for analyzing an unadjusted and a risk-adjusted quality indicator, with time from initial to potential subsequent operations as outcome of interest. While these methods are standard tools in the context of registry studies (Ranstam and Robertsson 2017; Gwinnutt et al. 2017), they have rarely been used in the context of sequential hospital-specific quality assurance.

1.1 The context of quality assurance in healthcare in Germany

Since 2016 and on behalf of the Federal Joint Committee, the Federal Institute for Quality Assurance and Transparency in Healthcare (IQTIG) is responsible for evaluating quality in healthcare in Germany. Among its main responsibilities is the evaluation of stationary quality of care both on a national and hospital-specific level by means of quality indicators. These indicators are developed in collaboration with medical expert panels and evaluated on an annual basis. In the areas of hip and knee replacement operations as well as cardiac pacemaker implantations, stationary operations are collected with pseudonymized identifiers for the purpose of quality assurance since 2015. This allows the linkage of initial and subsequent operation and enables the analysis of hospital-specific revision rates, which is a common measure in the context of quality assurance (Benbassat and Taragin 2000; Kristoffersen et al. 2015) and is included in the German context of quality assurance since 2016.

Besides quantifying the nationwide quality of care, quality indicators are also used to identify hospital performance outliers, either based on fixed thresholds or on thresholds derived from the distribution of the hospital-specific results. In an initial quantitative screening step, performance outliers among hospitals are identified. These are then subsequently further investigated qualitatively in a so called *structured dialogue* to see if the provided quality of care complies with the required standard and, if not, which quality improving measures can be taken (Federal Joint Committee 2016).

As part of the process, data are collected on state level and are then transferred to the IQTIG, such that data are available on an annual basis. Since beginning 2015 until the end of 2016, about 368,000 hip and 273,000 knee replacement operations have been collected with valid, pseudonymized identifiers and containing about 8300 and 4900 matched revisions, respectively. For cardiac pacemakers, the collected data contains about 116,000 implantation and 5700 matched revisions.

Of these initial operations, those carried out at the beginning of 2015 had a longer exposure period for revisions compared to operations at the end of 2016, as illustrated in Fig. 1 in form of a Lexis diagram based on three hypothetical implantation operations. In the annual evaluation of the year 2016, the described indicators based on time from initial to (potential) subsequent operation have been evaluated for the first time. Thus, data from both years 2015 and 2016 were included, with December 31st 2016 as right-censoring date. Additional information on competing events, such as the death of a patient, are unfortunately not available by the current regulatory setup, which is why all analyses are limited to the occurrence of subsequent operations or the censoring of observations due to the end of data collection.

Traditional analyses like the simple proportion of revisions within a defined follow-up period, e.g. 90 days after initial operation, exclude observations with follow-up time shorter than these 90 days. Thus, in the hypothetical example of Fig. 1, the observation with initial operation at the end of 2016 would not be included. Methods from survival analysis provide a more efficient way of dealing with such early censoring in that they allow the inclusion of all observations and account for the difference in exposure time. In practice, this also means that the use of survival methodology allows a quicker reaction to increasing revision rates than the use of traditional methods because all revision can be included in the analysis, i.e. also those with incomplete follow-up time.

The remainder of this paper is organized as follows: in Sect. 2, we develop an unadjusted as well as a risk-adjusted quality indicator based on survival analysis methods and describe how hospitals are classified as performance outliers on their basis. Although the current regulatory framework in Germany requires the classification to be based on point estimates (Federal Joint Committee 2016), we also discuss the construction of appropriate

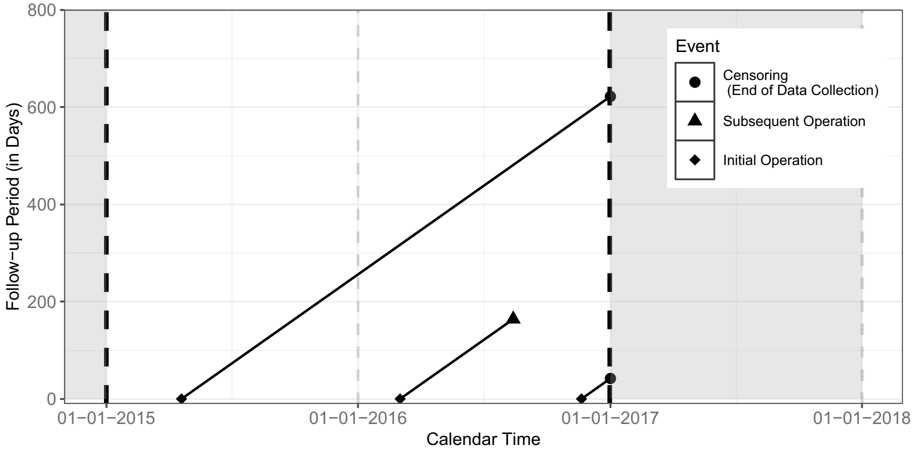

Fig. 1 Lexis diagram of three hypothetical observations

confidence intervals for assessing the estimation uncertainty associated with each hospital-specific estimate. Section 3 presents results from a simulation study in which the performance of the proposed methods is compared to that of a classification based on a simple proportion measure which takes into account only observations with complete follow-up time. Afterwards, we present results from the 2016 annual evaluation of the quality of care for an unadjusted (knee replacement operations) and a risk-adjusted quality indicator (hip replacement operations) in Germany and discuss the Cox model as alternative way to adjust for patient characteristics, which yields similar results in our case (Sect. 4). Section 5 concludes with summarizing the main findings and giving an outlook on possible future developments of these indicators.

2 Methods

Let T_{ij} denote the time from initial to subsequent operation of initial operation j in hospital $i, i = 1, \dots, m$ and $j = 1, \dots, n_i$, with n_i as number of initial implantation operation in hospital i. Furthermore, let C_{ij} be the time to (right-)censoring. One observes $T^*_{ij} = \min(T_{ij}, C_{ij})$, the event indicator $\delta_{ij} = I(T_{ij} \leq C_{ij})$ and a vector of covariates Z_{ij} that are relevant for the risk adjustment between hospitals. We assume that for a specific hospital i, T_{ij} are independent and identically distributed random variables with distribution function $F(t)$ and that censoring is independent of T_{ij}. Initial operations carried out on the same patient (e.g. knee or hip replacements on the right and left side of the same patient) are treated as independent observations and we limit the analyses to the first subsequent operation of each implantation. This is necessary because the responsible operation for all following ones cannot always be explicitly determined.

The statistical framework is developed for the context and requirements of the process of routine quality assurance in Germany. These requirements include the exact reproducibility of results, which, for instance, hampers the use of resampling or simulation based methods for obtaining estimates or confidence intervals. Also, results need to be easily interpretable for a broader, non-statistical audience, which advocates the orientation on previously established measures of quality indicators like rates or (risk-adjusted) standardized mortality ratios (IQTIG 2017). Therefore, we propose two types of indicators that closely resemble the interpretation of these two measures and still allow to account for varying observation times: the first based on the hospital-specific survival function and the second on a hazard based model with a multiplicative hospital-specific effect. Furthermore, small volume hospitals demand the use of confidence intervals that provide good coverage probabilities also in settings with only few cases.

Given this contextual setting, we first discuss the estimation of the survivor function and the multiplicative effect, including the construction of appropriate confidence intervals in the following. Based on the derived estimates, we then describe how performance outliers are identified within the regulatory context of quality assurance in Germany.

2.1 Estimating hospital-specific survival functions

In this context, the survival function $S_i(t) = P_i(T > t) = 1 - F_i(t)$ is a function of the probability that no subsequent operation will be carried out beyond t after an initial implantation operation in hospital i. A popular, non-parametric estimator for $S_i(t)$ was proposed by Kaplan and Meier (1958):

$$\hat{S}_i(t) = \begin{cases} 1 & \text{if } t < t_{i1}, \\ \prod_{t_{ki} \leq t} \left(1 - \frac{d_{ki}}{y_{ki}}\right) & \text{if } t_{i1} \leq t \end{cases},$$

with $t_{i1} < t_{i2} < \cdots < t_{iD}$ as distinct event times and d_{ki} and y_{ki} as number of events and observations at risk for a subsequent operation at t_{ki}, respectively. The Kaplan–Meier estimator is well defined for all $t \leq t_{i,\max}$, with $t_{i,\max}$ as largest observed study time of hospital i. If the largest time point in the data belongs to a censored observation, $\hat{S}_i(t > t_{i,\max})$ is not defined since the time of the last subsequent operation is not observed. In this case, we set $\hat{S}_i(t > t_{i,\max}) = \hat{S}_i(t_{i,\max})$ (Klein and Moeschberger 2003). Confidence intervals for the survival function are typically constructed based on the Greenwood formula (Klein and Moeschberger 2003). This method, however, only provides confidence intervals that have good coverage probabilities if the number of cases is large. To ensure good coverage of the confidence intervals also in small sample settings, we make use of the exact mid-p confidence intervals suggested by Fay and Brittain (2016). Without censoring, these confidence intervals reduce to exact mid-p confidence intervals for a binomial parameter, which are already used in a similar setting in the context of quality assurance in Germany (IQTIG 2016). In "Appendix", we briefly describe the construction of the exact mid-p confidence intervals. Using simulated data, Fay and Brittain (2016) show that these confidence intervals provide better coverage probabilities than competing methods, such as intervals based on the Greenwood variance or on the hybrid variance estimator described by Borkowf (2005). For details, we refer to Fay and Brittain (2016).

2.2 Comparing hospital-specific survival functions

Based on the Kaplan–Meier estimator for $S_i(t)$, the hospital-specific survival functions can be compared at specific time points τ, $\hat{S}_i(t = \tau)$. In principle, this comparison can account for uncertainty in the estimation of $\hat{S}_i(\tau)$, see e.g. (Goldstein and Spiegelhalter 1996). As mentioned, the regulatory guideline (Federal Joint Committee 2016) requires the classification of hospital results to not take possible estimation uncertainty and stochasticity into account in order for the procedure to have a high sensitivity (at the cost of a possibly low specificity). Therefore, the comparison of hospital results is solely based upon the point estimate $\hat{S}_i(t = \tau)$ in this context. Specifically, the threshold is calculated as $\alpha \cdot 100\%$ percentile of $\hat{S}_i(\tau)$ over all hospitals H for which $n_i \geq 20$:

$$\rho_\tau^s = Q\left(\alpha, \left\{\hat{S}_i(\tau)\right\}_{i \in H}\right),$$

and applied to all hospitals (i.e. also those with $n_i < 20$), with Q being the empirical quantile function. Hospital results are thus classified as performance outliers if $\hat{S}_i(\tau) < \rho_\tau^s$.

Besides estimating $S_i(t)$, comparing hospital results also involves choosing a specific time point τ at which estimates $\hat{S}_i(t = \tau)$ are to be compared. For some indicators this time point is already determined based on medical reasoning, in other cases it is left open. In these cases the specific choice should be clinically motivated but also needs to take into account the limitations of the available data: Due to the nature of right-censored survival data, the number of initial operations still at risk for revisions decreases with increasing time since initial operation and, thus, estimation uncertainty increases the larger τ is. On the other hand, differences in $\left\{\hat{S}_i(\tau)\right\}_{i \in H}$ accumulate over time (up to some point) and are less pronounced at very early time points. We quantified this trade-off between estimation precision and variability of hospital results by calculating the weighted mean squared difference between $\hat{S}_i(\tau)$ and $\hat{S}(\tau)$ as

$$D(\tau) = \sum_{i=1}^{m} n_i(\tau)\left[\hat{S}_i(\tau) - \hat{S}(\tau)\right]^2$$

(see Fig. 2 in Sect. 4.1), and with $\hat{S}(\tau)$ being the Kaplan–Meier estimate of the survival curve for patients of all hospitals combined. In addition to this measure, we analyzed the sensitivity and specificity of the classification based on $\hat{S}(\tau)$ at different time points τ using simulated data as explained in Sect. 3. Based on this as well as considering clinical reasons, τ was chosen individually for each quality indicator.

As an alternative, one could compare the expected number of years lost for an implant due to revision in [0, K] based on the quantities $K - \int_0^K \hat{S}_i(t)dt$ (see e.g. Andersen 2013). In the context of quality assurance, however, $\hat{S}_i(\tau)$ has the advantage of being interpretable in similar fashion to a simple proportion (adjusted for censoring), which is the preferred measure for many of the other quality indicators employed by the IQTIG.

2.3 Estimating hospital-specific effects on the hazard function

The comparison of hospital results based on $\hat{S}_i(t)$ presented in the previous section does not account for systematic differences between groups of patients across hospitals, which may be necessary for some of the considered types of implantations. For hip replacements, for instance, readmission rates are known to be lower for elective operations compared to implantations for the treatment of femoral fractures (Lim et al. 2016). Thus, hospital results should be compared under consideration of their respective patient population (Iezzoni 2013). While it is, in principle, possible to directly adjust survival curves for covariates (Xie and Liu 2005), risk adjustment can also be embedded within a multiplicative hazards model:

$$\lambda_{ij}(t) = \theta_i \lambda_j^*(t),$$

with $\theta_i > 0$ as (time-constant) hospital-specific effect, $\lambda_{ij}(t)$ as hazard rate of initial operation j in hospital i and $\lambda_j^*(t)$ as reference hazard rate for observations with covariate values

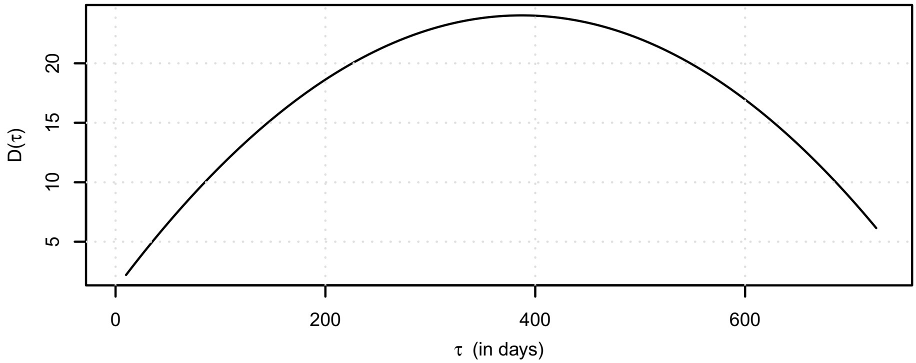

Fig. 2 (Smoothed) Trade-off between estimation precision and variability of hospital results. Values of τ represent the time point at which hospital results are compared and values of D(τ) quantify the trade-off between estimation precision and variability as explained in Sect. 2.2

equal to that of ij (but independent of i). For this model, originally proposed by Breslow (1975), the log-likelihood for right-censored data is

$$l(\theta_i) = o_i \log(\theta_i) + \theta_i \sum_{j=1}^{n_i} \log(S_{*,ij}(t_{ij})),$$

with $o_i = \sum_{j=1}^{n_i} \delta_{ij}$ and $S_{*,ij}(t_{ij})$ as reference survival function for observations with the same covariate values as observation ij (Breslow 1975). The maximum likelihood estimator is then

$$\hat{\theta}_i = \frac{o_i}{-\sum_{j=1}^{n_i} \log(S_{*,ij}(t_{ij}))} = \frac{o_i}{e_i},$$

and can be interpreted as a standardized mortality ratio (SMR). The model assumes the SMR to be constant over time, which may not always hold in practice. In case the SMR varies over time, $\hat{\theta}_i$ still provides a meaningful estimate that can be interpreted as a time-averaged effect over the analyzed time period. Right-censoring, however, can lead to a biased estimate in this case, which is why inverse probability of censoring weighting has been suggested as one way to recover the unbiased time-averaged estimate, see e.g. Dunkler et al. (2010). As shown in Sect. 4.2.1, the hospital-specific effects θ_i seem to be relatively constant over time in our data, which is why we proceeded with the unweighted estimate presented above.

We choose to estimate $S_{*,ij}(t_{ij})$ via stratified Kaplan–Meier curves, since the covariables consist of only categorical variables. Alternatively, the expected number of events can also be estimated via the cumulative hazards or Cox models with estimators for the cumulative baseline hazard, as for instance described in Agency for Healthcare Research and Quality Agency for Healthcare Research and Quality (2017), in which a Cox model is used to estimate the expected mortality of dialysis patients for comparing facility-specific SMRs (see also Berry (1983)). The SMR constitutes a form of indirect standardization and the calculation of the ratio of observed to expected numbers of events is currently the standard form to analyze risk-adjusted quality indicators in the German quality assurance (IQTIG 2017). Thus, the multiplicative hazards model fits well to existing methods already in use at the IQTIG. As an alternative, it would also be possible to calculate directly standardized estimates, for instance via the standardized rate ratio (He and Schaubel 2014), but these approaches are difficult in settings with empty strata.

Confidence intervals for θ_i can be constructed in multiple ways, for instance as Wald intervals based on the variance estimate of θ_i. Due to their better performance in a small-sample setting, we instead use likelihood-ratio based confidence intervals derived from the test statistic $T_i(\theta_0) = 2\{l_i(\hat{\theta}_i) - l_i(\theta_0)\}$, with null hypothesis $\theta_0 = 1$ and asymptotic distribution $T_i \sim \chi^2(1)$. Because the log-likelihood is not defined if $o_i = 0$, we set the lower interval limit to 0 in this case and define the upper limit as the upper interval of an exact binomial mid-p confidence interval in this case. Alternatively, the upper limit could be determined via an exact Poisson test. The binomial distribution, however, offers the advantage of taking into account the number of observations without subsequent operation.

2.4 Comparing hospital-specific effects on the hazard function

Based on $\hat{\theta}_i$, indirectly standardized hospital results can similarly to Sect. 2.2 be classified as performance outliers, based on whether $\hat{\theta}_i > \rho^\theta$, with $\rho^\theta = Q(1-\alpha, \{\hat{\theta}_i\}_{i \in H})$. By construction of the model, θ_i is assumed to be time-constant, hence, choosing a specific time point for comparing hospital results is not required.

2.5 Implementation

All computations were run using R version 3.4.1. (R Core Team 2017). Kaplan–Meier estimates with corresponding exact mid-p confidence intervals were calculated using the `bpcp`-package (Fay and Brittain 2016) and exact mid-p confidence intervals for the proportion estimate using the package `exactci` (Fay 2010). The SMR was calculated using the `glm`-function and all remaining survival models were fitted using the `survival`-package (Therneau 2015).

3 Simulation

In the following, we use the available data on knee replacement operations between January 1st, 2015, and December 31st, 2016, as reference point for simulating data and evaluating the performance of the proposed methods compared to a simpler classification based on the raw proportion of observed revisions. Since the Kaplan–Meier estimator is only proposed for quality indicators not adjusted for additional risk factors, we compare the proposed methods in a setting without risk adjustment. Using parameter estimates obtained from the original data, we simulated data with varying numbers of cases per hospital, dates of initial operations as well as times to revision or censoring, and evaluated the performance of the methods based on these simulated data, as described in detail below.

3.1 Simulation design

The number of hospitals who reportedly carried out knee implantation operations within the observation period 2015–2016 in Germany was $m = 1214$. For each hospital, we generated a random number of implantation operations n_i, with $n_i | n_i \geq 1 \sim \mathrm{NB}(s, p)$. The date of each implantation operation was sampled from the empirical distribution of dates. For all initial operations, the right-censoring date was set to December 31st, 2016. Time from initial to subsequent operation was assumed to follow a generalized Gamma distribution (Cox et al. 2007), i.e. $T_{ij} \sim \mathrm{GG}(\mu, \sigma, q)$, with probability density function

$$f(t|\mu,\sigma,q) = \frac{|q|(q^{-2})^{q^{-2}}}{\sigma t \Gamma(q^{-2})} \exp[q^{-2}(qw - \exp(qw))],$$

with $\gamma = \Gamma(q^{-2}, 1)$, $w = \log(q^2\gamma)/q$, $\Gamma(\cdot)$ denoting the gamma function and $\mu, q \in \mathbb{R}, \sigma > 0$. The GG-distribution contains the log-normal ($q = 0$), gamma ($q = 1$) and Weibull distribution ($q = \sigma$) as special cases and allows a more flexible modelling of the underlying hazard function compared to, for instance, the frequently used Weibull distribution which allows only monotonically increasing or decreasing hazards (Cox et al. 2007). In contrast to that, the GG-distribution allows flexible, bathtub-shaped hazard functions that reflect a high but reducing risk after operation with subsequently increasing risks over time (Briggs et al. 2004). Such shapes are often found in classical reliability theory (Shehla and Khan 2016).

For each implantation operation ij, we set the scale parameter to $\sigma_{ij} = \sigma, \sigma > 0$, and shape parameter to $q_{ij} = q$. The location parameter μ_{ij}, in turn, is allowed to vary by hospital, $\mu_{ij} = \exp(\alpha + \beta_i)$, with the hospital-specific effect β_i as normally distributed random variable $\beta_i \sim N(0, \phi)$, with $\phi > 0$. Thus, the variability of hospital performance (in terms of time to subsequent operation) increases with increasing ϕ, with larger β_i corresponding to longer times to subsequent operation (and, thus, better hospital performance). Because the hospital-specific effect only affects the location parameter, the underlying data generating model is equivalent to an accelerated failure time model in which $\exp(\alpha + \beta_i)$ has a multiplicative effect on the time to subsequent operation

$$S(t|\mu_i = \exp(\alpha + \beta_i)) = S(t \cdot \exp(-\exp(\alpha + \beta_i) \mid \mu_i = 0),$$

but not on the underlying hazard function. As long as the distribution does not simplify to a Weibull (or Exponential) distribution, this is different from the multiplicative hazard model in which $S(t|\mu_i = \exp(\alpha + \beta_i)) = S(t|\mu_i = 0)^{\exp(-\exp(\alpha+\beta_i))}$. Thus, the simulation design resembles the likely case that the modelling assumptions are not completely fulfilled in practice.

Estimating the parameters from the data on knee replacement operations by maximum (penalized) likelihood, we obtained $\hat{s} = 0.951$, $\hat{p} = 0.004$, $\hat{\alpha} = 2.510$, $\hat{\sigma} = 3.094$, $\hat{q} = 0.001$ and $\hat{\phi} = 0.304$, with which we simulated $K = 500$ datasets. For each simulation k, $k = 1, \ldots, K$, we computed $\hat{S}_{ik}(\tau)$, $\rho^s_{\tau,k}$, $\hat{\theta}_{ik}$ and ρ^θ_k as described above. For all methods and each simulation, hospitals results were classified into performance outliers and non-outliers. In addition to the classification method described above, we applied a second classification scheme in which the classification also involves the upper (UL_i) or lower limit (LL_i) of the estimated 90% confidence intervals (corresponding to a one-sided significance level of 5%). In this case, $\rho^s_{\tau,k}$ and ρ^θ_k are derived as explained above, but the classification is based on the evaluation of $\widehat{\mathrm{UL}}(S_i(t)) < \rho^s_{\tau,k}$ and $\widehat{\mathrm{LL}}(\theta_i) > \rho^\theta_k$, respectively, instead of on the point estimates.

We compared the performance of the proposed methods to a classification based on the simple proportion

$$r_i(\tau) = \sum_{j \in R_i(\tau)}^{n_i} \frac{\delta_{i,j}}{\mathrm{I}_{i,j}},$$

with $R_i(\tau)$ denoting the set of all observations with observed time greater or equal to τ, i.e. $R_i(\tau) = \{j \in \{1, \ldots, n_i\} \mid T^*_{ij} \geq \tau\}$ and indicator function $\mathrm{I}_{i,j}(j \in R_i(\tau))$. Again, the threshold value was calculated as the empirical $(1 - \alpha) \cdot 100\%$ quantile of the r_i values. For r_i, we calculate 90% exact mid-p binomial confidence intervals.

3.1.1 Measures of performance

As measures of performance, we calculate sensitivity and specificity, with for example sensitivity defined as:

$$\frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{i=1}^{m} I(\beta_{ik} < Q(\alpha, \{\beta_{ik}\}_{i=1}^{m})) \cdot I(\hat{\beta}_{ik} < Q(\alpha, \{\hat{\beta}_{ik}\}_{i=1}^{m}))}{\sum_{i=1}^{m} I(\beta_{ik} < Q(\alpha, \{\beta_{ik}\}_{i=1}^{m}))},$$

with $\hat{\beta}$ as estimated and β as true hospital-specific effect, and specificity accordingly. Additionally, we also calculated the positive (PPV) and negative predictive value (NPV) for the classification of performance outliers, with *positive* referring to a classification as outlier and *negative* as non-outlier, as well as the area under the receiver operating characteristics curve (AUC).

3.2 Simulation results

Table 1 displays the results for scenarios in which the calculation of the threshold ρ as well as the classification of hospital results are based on the point estimates. Overall, specificity as well as NPV of all methods is high, due to the fact that by design of the simulation, only the worst 5% of all hospital results are defined as outliers and all methods classify only somewhat more than 5% of all results as such. This also leads to a high AUC for all methods, leaving this measure with only little discriminatory power between the methods.

Sensitivity as well as PPV, in contrast, vary more substantially between the classification methods. For each investigated τ (except for $\tau = 90$), sensitivity and PPV of the Kaplan–Meier estimator are higher compared to the classification based on the raw proportion, due to the fact that the Kaplan–Meier estimator makes use of all available information, whereas the proportion is only calculated on the basis of implantations with censoring time greater than (or equal to) τ. This is also the reason why the sensitivity of the classification based on raw proportions is low for large τ.

For the Kaplan–Meier method, sensitivity and PPV is lowest for early and late time points within the observation period, indicating that the optimal balance of estimation certainty and variability of hospital-specific results lies somewhere in between. The SMR yields an even slightly higher sensitivity and PPV, even though its underlying assumption of time-constant multiplicative hospital effects on the hazard function does not hold in this generalized gamma simulation setting.

Table 2 displays results for scenarios in which the threshold ρ is calculated on the basis of point estimates but in which hospital-specific results are classified as outlier if the upper 90% confidence interval falls below ρ (Kaplan–Meier) or the lower 90% confidence interval exceeds τ (SMR and Proportion). As expected, this classification algorithm generally yields a higher specificity and lower sensitivity compared to the classification algorithm of Table 1 due to the fact that less hospitals are classified as performance outliers. This trade-off is a function of the level of significance α, with smaller values of α leading to

Table 1 Simulation results for the classification based on point estimates

Method	τ (in days)	Sensitivity	Specificity	PPV	NPV	AUC
SMR		0.856	0.990	0.824	0.992	0.990
Kaplan–Meier	180	0.842	0.990	0.814	0.992	0.986
Kaplan–Meier	365	0.837	0.989	0.801	0.992	0.987
Proportion	180	0.833	0.987	0.796	0.991	0.983
Proportion	90	0.830	0.988	0.799	0.991	0.982
Kaplan–Meier	90	0.830	0.990	0.808	0.991	0.983
Proportion	365	0.810	0.981	0.753	0.990	0.981
Kaplan–Meier	700	0.789	0.987	0.757	0.989	0.984
Proportion	700	0.576	0.859	0.444	0.981	0.904

Table 2 Simulation results for the classification based on confidence limits

Method	τ (in days)	Sensitivity	Specificity	PPV	NPV	AUC
SMR		0.566	0.999	0.977	0.978	0.986
Kaplan–Meier	180	0.531	1.000	0.984	0.976	0.981
Kaplan–Meier	90	0.515	1.000	0.986	0.975	0.979
Kaplan–Meier	365	0.509	1.000	0.985	0.975	0.982
Proportion	90	0.488	0.999	0.990	0.974	0.946
Proportion	180	0.476	0.998	0.989	0.974	0.952
Kaplan–Meier	700	0.446	1.000	0.981	0.972	0.981
Proportion	365	0.416	0.995	0.988	0.971	0.947
Proportion	700	0.085	0.897	0.930	0.955	0.830

higher specificity and lower sensitivity, and vice versa. This is also the case for the trade-off between PPV and NPV. When α converges to 1, upper and lower confidence intervals converge towards the point estimates and, thus, the results would converge towards that of Table 1.

3.3 Insights from the simulation study

The performance of the multiplicative hazards model in terms of the presented measures is slightly better than that of the analyzed alternative methods, even though its underlying assumption of time-constant multiplicative hospital-effects does not hold given the data generating process of the simulation design explored above. Thus, as long as the specification of the model does not deviate too much from the true underlying process, SMR and Kaplan–Meier yield similar results regarding the classification of performance outliers. For the latter method, however, the timing of when hospital results are compared does matter, with too early and too late points of time not being suited so well. This may change if additional effects of external influences on revision rates accumulate over time, which we have neglected in the simulation. In all cases, though, the survival method should be preferred over the calculation of simple proportions, as argued above. This insight might seem obvious for statisticians, but was necessary to illustrate empirically in our interdisciplinary context, where simple methods are often preferred because they can be easier to understand. With respect to the classification of performance outliers by point estimate or confidence interval, the statistician might also have a clear preference, but this proves to be a highly critical debate in practise, because the fear is that quality deficits in small volume providers might then be overlooked. Future analyses and possible extensions are needed to address these concerns.

4 Results: hip and knee replacement operations in Germany

As described earlier, patient follow-up data for hip and knee replacement operations as well as cardiac pacemaker implantations is being collected in Germany since 2015 and was for the first time analyzed within the annual nationwide quality evaluation in 2016. Thus, the observation period ranged from January 1st, 2015, to December 31st, 2016. In the following, we present an example for an unadjusted (knee replacement operations) as well as a risk-adjusted quality indicator (hip replacement operations) based on the methodology described above. Revisions with implantation operation before January 1st, 2015, could not be linked to their initial operation and its responsible hospital and are thus excluded from the analyses.

4.1 Knee replacement operations

Overall, we analyzed 273,148 knee replacement operations with 4866 revisions in 1214 different hospitals within the observation period. The number of initial operations per hospital varied between 1 and 2680 operations (median 151.5) over these 2 years. As explained above, Fig. 2 displays the (smoothed) trade-off between the number of observations still at risk and the variability of hospital results. This quantification indicates that τ should lie in the middle rather than directly at the beginning or the end of the observation period of 2 years after initial operation, which reflects the results of the simulation study above.

Figure 3 displays the Kaplan–Meier curve and a point-wise two-sided 95% exact mid-p confidence intervals (CI) over the whole period, with the vertical line indicating the time point for comparing the hospital-specific estimates, $\tau = 365$ days.

At $\tau = 365$ days, the nationwide estimate is $\hat{S}(\tau) = 0.982$ (95% CI 0.981–0.982). The hospital-specific estimates at $\tau = 365$ range from 0.5 to 1 and the critical value ρ^s_τ at $\tau = 365$, is 0.949. Thus, all hospital results below that, i.e. $\hat{S}_i(\tau = 365) < 0.949$, are classified as performance outliers. In the case of knee replacement operations, the result of 66 hospitals falls below this threshold value. The quality of care of these outliers was further scrutinized qualitatively within the framework of a *structured dialogue* as explained previously.

4.2 Hip replacement operations

Data on hip replacement operations consists of 368,267 implantation and 8289 revisions in 1353 different hospitals in the observation period. Of these, 76% were documented as elective operations and the remaining 24% as endoprosthetic treatment of a femoral fracture

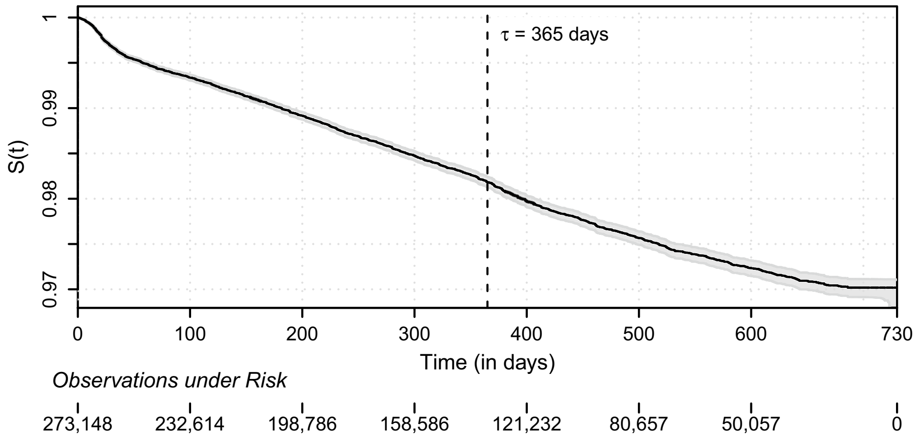

Fig. 3 Kaplan–Meier curve for knee replacement operations with corresponding pointwise 95% CIs

near the hip joint. Figure 4 displays the nationwide Kaplan–Meier curves, stratified by type of operation.

The probabilities of revisions substantially vary by type of operation, with nationwide survival estimates of $\hat{S}_{\text{elective}}(\tau = 730) = 0.973$ for elective operations and $\hat{S}_{\text{fracture}}(\tau = 730) = 0.965$ for the treatment of femoral fractures. Therefore, we compared indirectly standardized hospital results by calculating observed and expected numbers of revisions under consideration of the type of initial implantation operation as explained above. For calculating the SMR, we included all operations within the observed time period of 2 years. The nationwide SMR is 8289/8290 = 1.00 (95% CI 0.98–1.02), with hospital-specific estimates ranging from 0 to 77.6. Of all hospitals, 76 are classified as performance outliers due to their estimate θ_i lying above the threshold value $\rho^{\theta} = 2.39$.

4.2.1 Sensitivity analyses

As an additional analysis, we estimated the expected number of revisions on the basis of a Cox proportional hazards model instead of strata-specific Kaplan–Meier curves. Furthermore, we also estimated a Cox model with gamma-distributed frailty term $\omega_{p(i,j)}$, shared across operations of the same patient p:

$$\lambda_{ij}(t|Z_{ij}, \omega_{p(i,j)}) = \lambda_0(t) \cdot \exp(\beta' Z_{ij} + \omega_{p(i,j)}),$$

as well as a parametric accelerated failure time model on the basis of the GG distribution as specified in Sect. 3 (without shared frailty). From all four methods, we calculated the expected number of cases without revision by time as displayed in Fig. 5. Estimates from the Cox and the Cox frailty model are almost identical, indicating that ignoring the bilaterality of operations does not lead to substantially different estimates, likely because the number of bilateral operations at the same patient is relatively small (about 3.8% of all implantations are carried out on patients that already had an implantation operation at the other side during the data collection period for hip replacement operations; for knee implantations, this is the case for about 4.5% of the implantation operations). For similar results see Robertsson and Ranstam (2003).

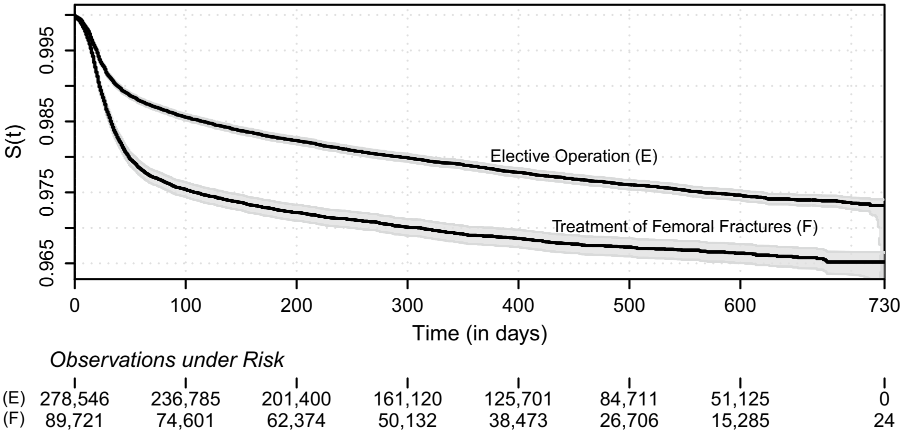

Fig. 4 Kaplan–Meier curves for hip replacement operations with corresponding pointwise 95% CIs

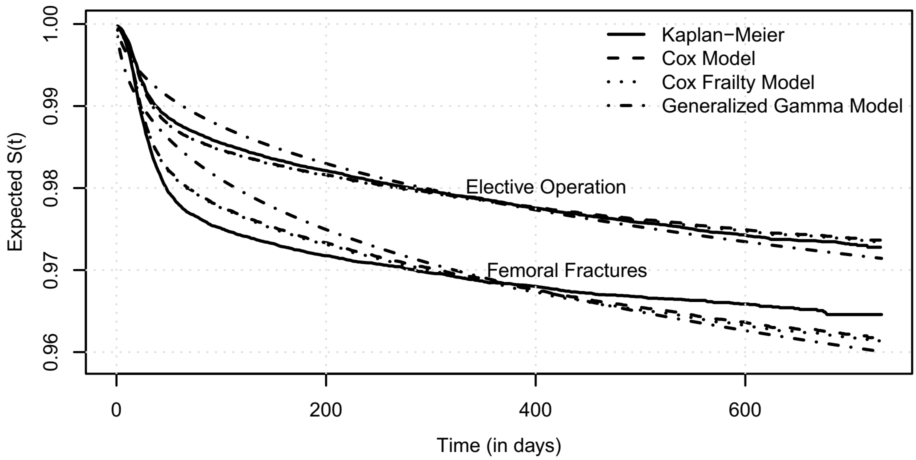

Fig. 5 Comparison of the expected proportions of cases without of subsequent operations

The difference between the Kaplan–Meier estimates and the semi-parametric Cox model indicates a deviation from the proportional hazards assumption, which we assessed as small on the basis of plotting the complementary log–log transformed survival curves against time (Fig. 5). Due to the small deviation from the proportional hazards assumption and the fact that the number of strata was small in our case, we decided to calculate the expected number of revisions based on the Kaplan–Meier estimator. We also checked the assumption of time-constant hospital effects of the multiplicative hazards model by analyzing time periods of varying length (by censoring observations after 90, 180 and 365 days, respectively). Because the hospitals classified as performance outliers varied only to a small extent, we concluded that time-varying effects are a minor concern in our analysis setting.

5 Discussion

Revision rates are an important measure of quality, both for product quality as well as for the initial implantation operation in the areas of arthroplasty and cardiac pacemakers. While registry studies usually trace implants over a long period of time and, thus, often make use of survival analysis methods, hospital-specific and nationwide revision rates were usually analyzed by calculating simple proportions within a fixed follow-up period in the context of German quality assurance. Instead of using rates or logistic regression, we show that these measures can be improved by using standard tools from survival analysis. The proposed methods allow the inclusion of all observed cases including those with only incompletely observed follow-up times and, thus, allow a more timely reaction to increasing revision rates. In addition, they provide a straightforward framework for adjusting for varying patient case-mixes. Previous studies have shown that various factors influence the probability of revisions of joint implants or cardiac pacemakers, which should thus be taken into account when classifying hospitals as performance outliers. In our case, we restricted the adjustment for risk factors to the type of initial hip operation and calculated unadjusted indicator results for knee replacement operations, though factors like age or comorbidities are known to be of influence (Jamsen et al. 2013; Bayliss et al. 2017). This matter is further complicated by the fact that information on the occurrence of competing events, like the death of patients, is unfortunately not available in our context because only operational events are currently transmitted as part of the mandatory reporting. Factors which increase the risk of such events may thus appear to have preventive effects on the risk of revisions. We hope that the data setting of future regulations allow such information to be included. If this is the case, the proposed survival methodology provides a framework in which the analysis of competing events is well established (Tsiatis 2005; Beyersmann et al. 2012). Incorporating competing risks into hazard based models, for instance, is relatively straightforward, as cause-specific hazard ratios can be estimated by simply censoring for competing events. These cause-specific hazards also provide a fair basis for comparing hospital effects on the event of interest, as the (potentially hospital-specific) presence and magnitude of competing risks does not affect the comparison of cause-specific hazards. For estimating cumulative revision probabilities, however, the proposed Kaplan–Meier estimator is not well suited in a setting with competing risks. Alternatively, the cumulative incidence function could be used to compare absolute event probabilities between different hospitals (Latouche et al. 2013).

Our evaluation of the performance of the analyzed methods was restricted to one specific simulation design, which, though, was chosen to have close resemblance with one of the application cases. Yet, the performance of semi- or fully parametric models always depends upon the validity of their inherent assumptions given the data at hand. In most cases, however, survival analysis methods should be preferable over traditional proportions or logistic regression. In addition to a more precise classification of performance outliers, an additional advantage of survival methods is their ability of dealing not only with right-censored but also left-truncated data (Klein and Moeschberger 2003). Despite being slightly more complex, this provides the possibility to left-truncate observation times that have already been qualitatively scrutinized in previous annual evaluations. This way, each annual evaluation could be based on observation times that have not been previously analyzed, avoiding that hospitals are repeatedly classified as outliers due to their performance in earlier years. The same can be achieved using traditional methods, though not with the efficiency of the methods advocated above.

For future evaluations of the quality of hospital care based on follow-up data, additional questions concerning the inclusion of observations into the analyses arise. If, as in the first analysis of follow-up data presented above, initial operations from a period of multiple years are included, hospital results cannot be attributed to a single year. This differs from the evaluation of other indicators in the German context of quality assurance and may hamper the identification of the underlying causes of the observed results. Additionally, improvements as well as deteriorations in the provided quality of hospital care are harder to detect if operations from multiple years are pooled. However, it allows more timely reactions, both regarding the quantitative assessment as well as the qualitative investigation of observations in the structured dialogue with hospitals. It is thus clear that for future analyses, developments in terms of communicable, adequate, simple and timely methods will be needed. However, survival analysis is a sound statistical framework for analyzing the data at hand. As easy as this claim is to make, it takes thorough discussions and specific display of added value to implement and communicate such methods into the quality assurance of 1353 hospitals. To summarize, the proposed statistical methodology fits well into the existing context of quality assurance in Germany while advancing the use of adequate statistical methods for the problem. In addition, the methodology meets the requirements of the regulatory context, e.g. the exact reproducibility of results, and the interpretation of the derived estimates closely resembles that of previously established measures in this context, easing the interpretation of results for a non-statistical audience.

Acknowledgements We thank the many colleagues from the IQTIG involved in this project for fruitful discussions.

Author contributions JH and MH analyzed the data, interpreted the results and wrote the manuscript. All authors read and approved the final manuscript.

Availability of data and materials The data are collected and analysed as part of the routine quality assurance in Germany based on §136ff SGB V. The restricted possibilities to perform analysis on these data is described in https://iqtig.org/datenerfassung/sekundaere-datennutzung/

Compliance with ethical standards

Conflict of interest Both authors declare that they have no conflict of interests.

Ethics approval and consent to participate The data are collected and analysed as part of the routine quality assurance in Germany based on §136ff SGB V. Thus, no ethical approval of the study was necessary.

Appendix: Exact mid-p confidence intervals for a survival function

As before, let $t_1 < t_2 < \cdots < t_D$ be the distinct event times, with $t_0 = 0$ and $t_{D+1} = \infty$, and let $l, l = 1, \dots, D+1$, denote the intervals $(0, t_1], (t_1, t_2], \dots, (t_D, \infty)$. Also, let a^- and b^- be vectors of length $D+2$, defined as $a_l^- = 1$ and $b_l^- = 0$ if $d_l = 0$ and $a_l^- = y_l - d_l + 1$ and $b_l^- = d_l$ if $d_l > 0$. Furthermore, let $a_0^- = 1$ and $b_0^- = 0$. A beta-product-variable with parameter vectors $\mathrm{a} = (a_0, \dots, a_{D+1})'$ and $\mathrm{b} = (b_0, \dots, b_{D+1})'$ is defined as:

$$\mathrm{BP}(\mathrm{a}, \mathrm{b}) = \prod_{l=0}^{D+1} B(a_l, b_l),$$

with $B(a, b)$ as beta-distributed random variable with parameters $a, b > 0$. Additionally, let $B(1, 0) \equiv 1$ and $B(0, 1) \equiv 0$ and define

$$W^-(t_l) = \prod_{h=0}^{l} B(a_h^-, b_h^-),$$

i.e. as a beta-product-variable based on the previously defined parameter vectors up to index l. Thus, $W^-(t_l) \sim \mathrm{BP}(\mathrm{a}_l^-, \mathrm{b}_l^-)$, with $\mathrm{a}_l^- = (a_0^-, \dots, a_l^-)'$ and $\mathrm{b}_l^- = (b_0^-, \dots, b_l^-)'$. In addition to that, let

$$\begin{aligned} W^+(t_l) &= W^-(t_l) B(y_{l+1}, 1) \\ &= \mathrm{BP}\big((a_0^-, \dots, a_l^-, y_{l+1})', (b_0^-, \dots, b_l^-, 1)'\big) \\ &= \mathrm{BP}(\mathrm{a}_l^+, \mathrm{b}_l^+), \end{aligned}$$

with $\mathrm{a}_l^+ = (a_0^-, \dots, a_l^-, y_{l+1})'$ and $\mathrm{b}_l^+ = (b_0^-, \dots, b_l^-, 1)'$. For $t_{l-1} \leq t < t_l$, define

$$W^*(t) = UW^-(t_{l-1}) + (1-U)W^+(t_l),$$

with $U \sim B(\frac{1}{2})$ as Bernoulli-distributed random variable with probability 1/2. A two-sided $(1-\alpha) \cdot 100\%$ exact mid-p confidence interval for $\overline{S}(t) = P(T \geq t)$ with $t_{l-1} \leq t < t_l$ can be defined as

$$\left(Q\left\{\frac{\alpha}{2}, W^*(t)\right\}, Q\left\{1-\frac{\alpha}{2}, W^*(t)\right\}\right),$$

with $Q\left\{\frac{\alpha}{2}, W^*(t)\right\}$ as $\alpha/2$-quantile of $W^*(t)$. As suggested by Fay and Brittain (2016), we approximate W^- and W^+ with beta-variables with matching first and second moments and obtain an exact mid-p interval by combining the quantiles of the respective beta distributions. Fay and Brittain (2016) show that, given non-informative censoring, the exact mid-p confidence intervals guarantee (on average) nominal level coverage independently from sample size. Without censoring, the confidence intervals reduce to exact mid-p confidence intervals for a binomial parameter.

References

Agency for Healthcare Research and Quality: Technical Notes on the Standardized Mortality Ratio (SMR). Technical report, Agency for Healthcare Research and Quality, Rockville, MD (2017)

Andersen, P.: Decomposition of number of life years lost according to causes of death. Stat. Med. **32**, 5278–85 (2013)

Bayliss, L., Culliford, D., Monk, A., Glyn-Jones, S., Prieto-Alhambra, D., Judge, A., Cooper, C., Carr, A., Arden, N., Beard, D., Price, A.: The effect of patient age at intervention on risk of implant revision after total replacement of the hip or knee: a population-based cohort study. Lancet **389**, 1424–30 (2017)

Benbassat, J., Taragin, M.: Hospital readmissions as a measure of quality of health care. Arch. Intern. Med. **160**, 1074–1081 (2000)

Bernatz, J., Tueting, J., Anderson, P.: Thirty-day readmission rates in orthopedics: a systematic review and meta-analysis. PLoS ONE **10**, e0123593 (2015)

Berry, G.: The analysis of mortality by the subject-years method. Biometrics **39**, 173–184 (1983)

Beyersmann, J., Schumacher, M., Allignol, A.: Competing Risks and Multistate Models with R. Springer, New York (2012)

Borkowf, C.: A simple hybrid variance estimator for the Kaplan–Meier survival function. Stat. Med. **26**, 827–851 (2005)

Bosco, J., Karkenny, A., Hutzler, L., Slover, J., Iorio, R.: Cost burden of 30-day readmissions following medicare total hip and knee arthroplasty. J. Arthroplasty **29**, 903–905 (2014)

Breslow, N.: Analysis of survival data under the proportional hazards model. Int. Stat. Rev. **43**, 45–57 (1975)

Briggs, A., Sculpher, M., Dawson, J., Fitzpatrick, R., Murray, D., Malchau, H.: The use of probabilistic dicision models in technology assessment: the case of total hip replacement. Appl. Health Econ. Health Policy **3**, 79–89 (2004)

Cox, C., Chu, H., Schneider, M., Muñoz, A.: Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Stat. Med. **26**, 4352–74 (2007)

Delaunay, C.: Registries in orthopaedics. Orthop. Traumatol. Surg. Res. **101**, S69–S75 (2015)

Dunkler, D., Heinze, G., Schemper, M.: Gene selection in microarray survival studies under possibly non-proportional hazards. Bioinformatics **26**(6), 784–790 (2010)

Fay, M.: Two-sided exact tests and matching confidence intervals for discrete data. R Journal **2**, 53–58 (2010)

Fay, M., Brittain, E.: Finite sample pointwise confidence intervals for a survival distribution with right-censored data. Stat. Med. **35**, 2726–2740 (2016)

Federal Joint Committee: Richtlinie über Maßnahmen der Qualitätssicherung in Krankenhäusern. Technical report (2016)

Gandy, A., Kvaløy, J.T., Bottle, A., Zhou, F.: Risk-adjusted monitoring of time to event. Biometrika **97**, 375–388 (2010)

Goldstein, H., Spiegelhalter, D.: League tables and their limitations: statistical issues in comparisons of institutional performance. J. R. Stat. Soc. Ser. A **159**, 358–443 (1996)

Gwinnutt, J., Symmons, D., MacGregor, A., Chipping, J., Lapraik, C., Mashall, T., Lunt, M., Verstappen, S.: Predictors of and outcomes following orthopaedic joint survery in patients with early rheumatoid arthritis followed for 20 years. Rheumatology **56**, 1510–1517 (2017)

He, K., Schaubel, D.: Methods for comparing center-specific survival outcomes using direct standardization. Stat. Med. **33**, 2048–2061 (2014)

Iezzoni, L.: Risk Adjustment for Measuring Healthcare Outcome. Health Administration Press, Chicago (2013)

IQTIG: Planungsrelevante Qualitätsindikatoren. Abschlussbericht zur Auswahl und Umsetzung. Stand: 31. August 2016. Technical report, Berlin: Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (2016)

IQTIG: Methodische Grundlagen V1.0s. Entwurf für das Stellungnahmeverfahren. Stand: 31. January 2017. Technical report, Berlin: Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (2017)

Jamsen, E., Peltola, M., Eskelinen, A., Lehto, M.: Comorbid diseases as predictors of survival of primary total hip and knee replacements: a nationwide register-based study of 96754 operations on patients with primary osteoarthritis. Ann. Rheum. Dis. **72**, 1975–82 (2013)

Junnila, M., Laaksonen, I., Eskelinen, A., Pulkkinen, P., Havelin, L., Furnes, O., Fenstad, A., Pedersen, A., Overgaard, S., Kärrholm, J., Garellick, G., Malchau, H., Mäkelä, K.: Implant survival of the most common cemented total hip devices from the nordic arthroplasty register association database. Acta Orthop. **87**, 546–553 (2016)

Kaplan, E., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**, 457–481 (1958)

Klein, J., Moeschberger, M.: Survival Analysis: Techniques for Censored and Truncated Data, 2nd edn. Statistics for Biology and Health, Springer (2003)

Kristoffersen, D., Helgeland, J., Waage, H., Thalamus, J., Clemens, D., Lindman, A., Rygh, L., Tjomsland, O.: Survival curves to support quality improvement in hospitals with excess 30-day mortality after acute myocardial infarction, cerebral stroke and hip fracture: a before–after study. BMJ Open **5**, e006741 (2015)

Latouche, A., Allignol, A., Beyersmann, J., Labopin, M., Fine, J.P.: A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. J. Clin. Epidemiol. **66**(6), 648–653 (2013)

Lim, J., Ng, G., Jenkins, R., Ridley, D., Jariwala, A., Sripada, S.: Total hip replacement for neck of femur fracture: comparing outcomes with matched elective cohort. Injury **47**, 2144–2148 (2016)

McGrory, B., Caryn, D., Lewallen, D.: Comparing contemporary revision burden among hip and knee joint replacement registries. Arthroplasty Today **2**, 83–86 (2016)

Mehrotra, A., Sloss, E., Hussey, P., Adams, J., Lovejoy, S., SooHoo, N.: Evaluation of a centers of excellence program for knee and hip replacement. Med. Care **51**, 28–36 (2014)

Oliveira, J., Valenca, D., Medeiros, P., Marcula, M.: Risk-adjusted monitoring of time to event in the presence of long-term survivors. Biom. J. **58**, 1485–1505 (2016)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2017)

Ranstam, J., Robertsson, O.: The cox model is better than the fine and gray model when estimating relative revision risks from arthroplasty register data. Acta Orthop. **88**, 578–580 (2017)

Robertsson, O., Ranstam, J.: No bias of ignored bilaterality when analysing the revision risk of knee prostheses: analysis of a population based sample of 44,590 patients with 55,298 knee prostheses from the national swedish knee arthroplasty register. BMC Musculoskelet. Disord. **4**, 1–4 (2003)

Shehla, R., Khan, A.: Reliability analysis using an exponential power model with bathtub-shaped failure rate function: a bayes study. SpringerPlus **5**, 1076 (2016)

Tarasevicius, S., Cebatorius, A., Valaviciene, R., Stucinskas, J., Leonas, L., Robertsson, O.: First outcome results after total knee and hip replacement from the lithuanian arthroplasty register. Medicina **50**, 87–91 (2014)

Therneau, T.: A Package for Survival Analysis in S. version 2.38 (2015)

Tsiatis, A.A.: Competing risks. In: Armitage, P., Colton, T. (eds.) Encyclopedia of Biostatistics, 2nd edn, pp. 824–835. Wiley, New York (2005)

Xie, J., Liu, C.: Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Stat. Med. **24**, 3089–3110 (2005)