Statistics in Medicine

Research Article

Received 20 December 2013, Accepted 16 December 2014 Published online 15 January 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6409

Evaluating hospital performance based on excess cause-specific incidence

Bart Van Rompaye,[a,b*†] Marie Eriksson[a] and Els Goetghebeur[b]

Formal evaluation of hospital performance in specific types of care is becoming an indispensable tool for quality assurance in the health care system. When the prime concern lies in reducing the risk of a cause-specific event, we propose to evaluate performance in terms of an average excess cumulative incidence, referring to the center's observed patient mix. Its intuitive interpretation helps give meaning to the evaluation results and facilitates the determination of important benchmarks for hospital performance. We apply it to the evaluation of cerebrovascular deaths after stroke in Swedish stroke centers, using data from Riksstroke, the Swedish stroke registry.

Keywords: competing risks; excess cumulative incidence; hospital performance evaluation; quality-of-care; stroke

1. Introduction

1.1. Health care provider evaluations

As health records are gathered on an ever larger scale, the quality of care provided can be monitored more systematically [1]. This allows the identification of best practices as well as weaknesses in the care process and hence points to room for improvement. It has been shown to improve health care [2, 3] and reduce costs [4], guiding policy makers towards a targeted distribution of funds. With public dissemination, it could inform patients on their choice of provider [5].

Acceptance of such assessments depends critically on their credibility, including perceived fairness and transparency [4]. The principal issue here is an adequate and well-understood risk adjustment, which avoids patient characteristics (such as age, gender, comorbidity, and severity of illness at hospital admission) confounding the hospital's outcome evaluation. The need for such adjustment is well recognized, for example, for hospitals with a specific demographic composition. When specialized facilities attract severely ill patients, without adequate risk adjustment, the more negative outlook of the high-risk patients might erroneously be ascribed to lower center effectiveness, seemingly contradicting the expert status.

For benchmarking, objective bounds are required, on which meaningful consensus is best reached if the metric used has a clear and practical interpretation. This challenge grows when results are publicized to a broad audience. Standardized mortality ratios (SMRs) or risk ratios emphasize relative differences, with unclear practical importance [6]. Directly standardized rates evaluate different centers as if they acted on the same (hypothetical) patient population, thereby allowing for direct comparisons. Unfortunately, they may not play a straightforward practical role – counterfactual or otherwise – and could firmly extrapolate the data: there is no guarantee that a center serving its low-risk patients well will perform equally well in a more diverse standard population. Indirectly standardized rates on the other hand evaluate each center

[a]*Department of Statistics, School of Business and Economics, Umeå University, SE-901 87 Umeå, Sweden*
[b]*Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, S9, 9000 Ghent, Belgium*
[*]*Correspondence to: Bart Van Rompaye, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, S9, 9000 Ghent, Belgium.*
[†]*E-mail: bart.vanrompaye@ugent.be*
The copyright line for this article was changed on 17 September 2015 after original online publication.

on its own patient population and compare its own outcome to how a typical (average) center would perform on the same patients. This also entails some extrapolation, but the question answered may be more relevant to policy makers pondering the impact of possible interventions.

While the previously mentioned metrics have a proven track record, more specific, targeted summary measures may better guide the process improvement. Therefore, we will work towards an indirectly standardized outcome on an additive scale, referring to a specific death cause of interest, and with clear, practical implications. We thus hope to promote both acceptance of the method and the inclination to act upon the results.

1.2. The competing risks setting

When a non-negligible proportion of patients dies from causes unrelated to the disease under study, all-cause mortality analyses (using e.g. SMRs or the Cox model) suffer diminished discriminatory power [7]. Through disease-specific event analyses, one may not only recover power but also gain insight [8]. In what follows, we refer to the various competing risks as 'deaths from different causes', although in general they need not refer to mortality endpoints.

With competing risks, one often assumes the cause-specific hazards (i.e., the rate of occurrence of a specific event at a given time point t, conditional on survival up to t) to be proportional, similar to the Cox model for overall survival. Pitfalls exist when interpreting a (cause-specific) hazard ratio as a causal effect or as the sole output of a survival model [9]. Nevertheless, the cause-specific hazard, which represents an observable instantaneous risk among patients who survived so far, is closely linked to the practical clinical experience. It sheds light on the dynamics underlying the force of mortality within a population, even if it does not readily imply observables over an extended period, as the direct competition with other causes intervenes. Surprisingly, even when good treatment postpones a cause-specific event, its incidence may increase if at the same time fewer competing events occur, thus leaving more patients at risk [8]. Vice versa, a cardiology unit that neglects basic hygienic measures may see fewer patients die from heart disease, simply because many of them succumb to nosocomial infections first. Cumulative incidences (i.e., the probability of dying from a given cause prior to a given time t) will however reflect this and may thus be more relevant when one is interested in the occurrence of events over time.

1.3. An average excess cumulative incidence measure

We will build on the proportional cause-specific hazards assumptions to obtain cumulative incidence predictions based on patients' measured baseline covariates. Using these expected patient risks, we will contrast for each patient the outcome predicted in his own center with the expected outcome should he be treated at a randomly chosen center. Averaging this contrast over all patients in a given center then yields an 'excess' outcome for that center: How much higher is their population's risk than what it could expect across all centers? By directly referring to the patient mix of the center under study, the excess represents the predicted additional percentage of deaths from the specific disease by time t in the center, because of being treated in that center. As such, it may be used to compute 'observable' quantities such as expected reduction in (cause-specific) mortality or expected cost reduction.

This excess cannot be seen in isolation of the competing cause incidence and is best presented with a careful interpretation in this light. These issues (and others) are discussed in Sections 3 and 4.

In Section 2, we set up the theoretical framework for the evaluation and develop an estimator for the center-specific excess outcome. While we use proportional cause-specific hazards models, other models (like the Fine and Gray model, [10]) leading to cumulative incidence estimates can be used (see the discussion), without fundamentally changing the construction and interpretation of our proposed outcome measure. We discuss the meaning of the excess outcome and propose an extension to evaluate its behavior over time. Technical results on variance estimation are deferred to the Supporting information, together with a small simulation exercise.

In Section 3, the approach is applied to evaluate quality of care in Swedish stroke care centers, based on a database containing information on 187 264 patients from the Swedish stroke registry, Riksstroke. Additional output for the Riksstroke data is presented in the Supporting information. The final section provides conclusions and a discussion.

2. Developing an excess measure in the competing risks setting

2.1. Notation and model assumptions

We consider a sample of n patients treated in m health care centers, as indicated by center indicators $X_{i,j}$ ($X_{i,j} = 1$ for patient i in center j, $X_{i,j} = 0$ elsewhere, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$), with $\mathbf{X_i} = (X_{i,1}, \ldots, X_{i,m-1})$. The reference center m is typically chosen to be a large-enough center. Without loss of generality, we consider only two possible failure types F_i: one from the cause of interest $F_i = dfd$ (for *death from disease*) and one from all other causes combined $F_i = oc$ (for *other causes*). The time after admission until failure of either type, T_i, may be censored at time C_i yielding observed time $\tilde{T}_i = min(T_i, C_i)$ until failure or censoring.

We further assume that model covariates include the necessary confounders for the relation between hospital choice and event occurrence. In practice, information on many of the strongest prognostic factors (typically gender, age, some measure of severity of illness at admission, and so on) tends to be available from hospital records or registries. The prognostic factors for the event of interest are denoted $\mathbf{Z}_{i,dfd}$, those for the other causes $\mathbf{Z}_{i,oc}$, where both sets may overlap. Although our method technically allows for time-varying covariates, prediction of the cumulative incidence function at time t then requires knowledge of the entire covariate history up to t, which is not available for each patient (e.g., those who died or were lost to follow-up before t, or with only partial measurements of clinical predictors like blood pressure). One should also avoid to naively correct for covariates whose time evolution may be influenced by the choice of hospital and the care received there, as this removes part of the total hospital effect [11].

For each cause, we assume proportionality at the level of the individual cause-specific hazards

$$\lambda_f(t; \mathbf{X}, \mathbf{Z}_{dfd}, \mathbf{Z}_{oc}) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(T < t + \Delta t, F = f | T \geqslant t; \mathbf{X}, \mathbf{Z}_{dfd}, \mathbf{Z}_{oc}) \qquad \text{with } f \in \{dfd, oc\},$$

so that

$$\begin{aligned} \lambda_{dfd}(t; \mathbf{X}, \mathbf{Z}_{dfd}, \mathbf{Z}_{oc}) &= \lambda_{dfd,0}(t) \exp\left(\boldsymbol{\beta}_{dfd}^T \mathbf{Z}_{dfd} + \boldsymbol{\gamma}_{dfd}^T \mathbf{X}\right) \\ \lambda_{oc}(t; \mathbf{X}, \mathbf{Z}_{dfd}, \mathbf{Z}_{oc}) &= \lambda_{oc,0}(t) \exp\left(\boldsymbol{\beta}_{oc}^T \mathbf{Z}_{oc} + \boldsymbol{\gamma}_{oc}^T \mathbf{X}\right). \end{aligned} \tag{1}$$

The covariate effects on the cause-specific hazard are thus represented by the column vector of regression coefficients $\boldsymbol{\beta}_{dfd}$, while $\boldsymbol{\beta}_{oc}$ represents the effects on the competing risks hazard. Center effects are given by $\boldsymbol{\gamma}_{dfd}$ and $\boldsymbol{\gamma}_{oc}$, respectively.

Censoring is assumed to be independent (conditional on the model covariates, including center choice) in the sense that the compensators of both cause-specific counting processes are assumed to be unchanged by the presence of censoring (as clarified in [12]).

2.2. Estimating the excess cumulative incidence and its variance

The predicted death-from-disease-specific cumulative incidence at time t for given covariate values $(\mathbf{x}, \mathbf{z}_{dfd}, \mathbf{z}_{oc})$ is

$$\hat{F}_{dfd}(t; \mathbf{x}, \mathbf{z}_{dfd}, \mathbf{z}_{oc}) = \hat{P}(T \leqslant t, F = dfd | \mathbf{x}, \mathbf{z}_{dfd}, \mathbf{z}_{oc}) = \int_0^t \hat{S}(u; \mathbf{x}, \mathbf{z}_{dfd}, \mathbf{z}_{oc}) \hat{\lambda}_{dfd}(u; \mathbf{x}, \mathbf{z}_{dfd}, \mathbf{z}_{oc}) du, \tag{2}$$

with the estimated overall survival

$$\hat{S}(u; \mathbf{x}, \mathbf{z}_{dfd}, \mathbf{z}_{oc}) = \exp\left\{ -\exp\left(\hat{\boldsymbol{\beta}}_{dfd}^T \mathbf{z}_{dfd} + \hat{\boldsymbol{\gamma}}_{dfd}^T \mathbf{x}\right) \int_0^u \hat{\lambda}_{dfd,0}(s) ds - \exp\left(\hat{\boldsymbol{\beta}}_{oc}^T \mathbf{Z}_{oc} + \hat{\boldsymbol{\gamma}}_{oc}^T \mathbf{x}\right) \int_0^u \hat{\lambda}_{oc,0}(s) ds \right\}. \tag{3}$$

Equations (2) and (3) use the effect estimates $(\hat{\boldsymbol{\gamma}}_{dfd}, \hat{\boldsymbol{\gamma}}_{oc}, \hat{\boldsymbol{\beta}}_{dfd}, \hat{\boldsymbol{\beta}}_{oc})$ together with estimates for the baseline cumulative cause-specific hazards (e.g., Breslow estimates).

We then define the excess cumulative incidence $E_{j,i}$ for patient i in center j as the difference between the expected cumulative incidence for patient i's profile in the actually visited center j and the average of his expected cumulative incidences over all centers:

$$\hat{E}_{j,i}(t) = \hat{F}_{dfd}(t; \mathbf{x}_i, \mathbf{z}_{dfd,i}, \mathbf{z}_{oc,i}) - \frac{1}{m} \sum_{c=1}^{m} \hat{F}_{dfd}(t; x_c \equiv 1, \mathbf{z}_{dfd,i}, \mathbf{z}_{oc,i}), \tag{4}$$

where $x_c \equiv 1$ is to be interpreted as $x_c = 1$, *and* $x_{c'} = 0$ *for* $c' \neq c$, thus indicating that we plug in the effect of center c, regardless of whether the patient was treated there.

The excess cause-specific cumulative incidence (which we call the *ECSCI*) for center j is then the average excess over all n_j patients in center j:

$$\hat{E}_j(t) = \frac{\sum_{i \in \text{center } j} \hat{E}_{j,i}(t)}{n_j}. \tag{5}$$

This compares the expected risk of the patients under current conditions with what is estimated to take place under a reference level of care, which could be interpreted as the level of care expected if center choice was random, among all centers under study.

A pointwise 95% confidence interval for $E_j(t)$ can be based on asymptotic normality, using a variance $V_j(t)$ developed along the lines of [13]. This variance incorporates the variances of, and the covariances between, the patient-specific predicted cumulative incidences. Because of the elaborate formulae, we defer the expression for $\hat{V}_j(t)$ to the Supporting information. Implementing the complex expression for $\hat{V}_j(t)$ is intricate, however, and asymptotic approximations may not hold in finite samples (as indicated by a small simulation in the Supporting information). Variance estimates may therefore rely on alternative techniques such as a wild bootstrap [14] or a jackknife estimate [15, 16]. We apply a parametric bootstrap [17], essentially by fitting our semi-parametric model to the observed data, followed by repeated simulations from the estimated cumulative baseline cause-specific hazards and fitted model parameters, and refitting of the model. The parameters of interest (i.e., the patient-level or hospital-level ECSCIs) can each time be recomputed based on the refitted models, and the empirical covariance matrix over all simulations estimates the true covariance matrix. Under the model assumptions, this approach is relatively simple, fast, and reliable. When model validity is an issue, nonparametric alternatives may be needed, for example, bootstrapping complete observations or paired cause-specific martingale residuals, or alternatively perhaps pseudo-values when using the Fine and Gray model (along the lines of [18]).

More detail on how to perform simulations in the competing risks setting can be found in [19, 20], or [21]. While we used self-written functions switching between SAS and R, implementation of the entire analysis in standard software packages can use for example the `mstate` package in `R` [22] or the `CumInc` and `CumIncv` macros in `SAS` [23].

2.3. *Interpretation and alternative excess definitions*

$\hat{E}_j(t)$ gives the absolute percentage of additional events of interest that are expected to occur prior to t in center j, over and above to the risk center j's population would experience on average over all centers. It is intuitively interpreted from an intervention perspective as the proportion of a center's patients that would be saved, were they to be directed to any of the available treatment centers at random. Therefore, a positive ECSCI indicates an increased cause-specific mortality.

This interpretation is illustrated by the toy example in Table I. In a population of three centers, we study center 1 that has seen two patients with profiles $(\mathbf{z}_{dfd}, \mathbf{z}_{oc}, \mathbf{x})_1$ and $(\mathbf{z}_{dfd}, \mathbf{z}_{oc}, \mathbf{x})_2$. The evaluation of center 1 is based on the predicted incidence for its two patients; in center 1 on the one hand (0.10 and 0.16, respectively) and averaged over the three centers on the other (0.09 and 0.14, respectively). The ECSCI of center 1 is 1.5%, indicating that the patients in center 1 are expected to have a 1-year *dfd*-specific cumulative incidence that is 1.5 percentage points higher than what is expected if they were to visit a randomly chosen center.

Table I. Numerical example of the composition of an excess cumulative incidence.

		Patient 1	Patient 2
Center 1	$\hat{F}_{dfd}(t=1; \mathbf{z}_{dfd,i}, \mathbf{z}_{oc,i}, x_1=1)$	0.10	0.16
Center 2	$\hat{F}_{dfd}(t=1; \mathbf{z}_{dfd,i}, \mathbf{z}_{oc,i}, x_2=1)$	0.09	0.11
Center 3	$\hat{F}_{dfd}(t=1; \mathbf{z}_{dfd,i}, \mathbf{z}_{oc,i}, x_1=x_2=0)$	0.08	0.15
	$\hat{E}_{x_1=1,i}(t=1)$	$0.1 - \frac{0.1+0.09+0.08}{3} = \mathbf{0.01}$	$0.16 - \frac{0.16+0.11+0.15}{3} = \mathbf{0.02}$
		$\hat{E}_1(t=1) = \frac{0.01+0.02}{2} = \mathbf{0.015}$	

Definition (4) contrasts the performance of the hospital under study to that of the complete population of hospitals. The excess can be redefined by changing the comparator term to reflect the difference with an external benchmark, a specific reference center, or another 'average' of centers (e.g., weighted by number of patients to reflect the higher chance of patients of turning to the larger center).

Alternatively, one could easily compare with what is expected in 'a hospital' that lies centered in the observed log cause-specific hazard ratio space by using effect-coded hospital dummies (as shown in the Supporting information). It is however harder to imagine what kind of center this reference would be. Alternatively, a known target benchmark could provide a natural comparator. However, patient profile-specific benchmarks in incidence terms are typically not available, prohibiting effective risk adjustment. Therefore, the incidence based on some prespecified quantile (e.g., the 75th percentile, as in [24]) of the log cause-specific hazard ratios is sometimes used.

As an indirectly standardized outcome, the ECSCIs for different centers relate to different underlying populations (those actually observed in each center). Therefore, direct comparisons of excesses, or the comparison of one excess against a quantile of the overall excess distribution, should be treated with care. Alternatively, one could sum over the whole patient population in Equation (5), leading to a directly standardized excess. Comparisons then seem more readily made, as parameters refer to the same population, although there is no guarantee that a hospital's performance carries over to a different patient mix. This extrapolates the expected performance to the full study population, with likely different compositions than the originally treated one. Changing the summation in Equation (5) further can address a center's potential performance in a specific patient subset, for example, in older patients and patients with diabetes.

A 'deleted' excess, averaging over all centers except for the one under study, may better distinguish excesses that are truly different from 0. This could then be interpreted as the expectation should the current center be closed and patients randomly distributed over the remaining centers. To reflect such closure more realistically, one could use weights reflecting the propensity of patients for redistribution to nearby centers. In such case, one should account for possible effects of distance to hospital in the risk adjustment and always correct the value of this risk factor in the various terms of the summation in Equation (4).

Finally, ceiling effects may occur in extremely high-risk or low-risk populations. When virtually everyone (no one) experiences the event of interest, the cumulative incidence quickly approaches 1 (0), leaving little room for an additional increase (decrease) in incidence due to hospital performance. This gives a practical upper (lower) bound close to zero on the excess. Differences in excess will then be small, reflecting that the choice of hospital matters little in this respect. The direct comparison of excesses should be carefully interpreted in this light when populations are extremely diverse. Furthermore, this emphasizes the importance of discriminatory power when choosing an appropriate outcome to evaluate quality of care. Ceiling effects may be mitigated by using a relative measure of attributable incidence, for example, along the lines of the excess fraction of Greenland and Robins [25].

2.4. The excess risk building up over time

In many diseases and for many routinely used clinical parameters, consensus exists about the time over which information is to be accrued, for example, a 5-year survival among breast cancer patients or a 30-day survival in intensive care units. This choice of evaluation time is important when using the ECSCI. Enough time should elapse to allow the quality of treatment to play its role, but it should also not stretch too far into the follow-up, as center differences may disappear under the natural course of disease. An optimal evaluation time depends on the expected cause-specific hazards and their evolution over time, interpreted in light of the character of the disease and its treatment. For this choice, it may be instructive to assess the buildup of the ECSCI $\hat{E}_j(t)$ over an extended period of follow-up.

Because cumulative incidences build up over time, the ECSCIs (and any differences between them) will change with t, even if the (relative) quality of care remains constant. This comes in part from the interplay of the baseline cause-specific hazards changing over time while being scaled by the appropriate hazard ratios. Also, with different baseline patient populations between centers, the differential selection of less frail patients with t matters.

Capturing real center-specific changes in performance over the course of disease (e.g., because of the provided care trajectory) requires more flexible models than model (1). With enough data on all centers,

one could use models stratified by center. However, in many realistic settings, small centers appear, where the performance of such analysis may suffer from a limited number of events. Including parametric time-varying center effects may then offer an alternative:

$$\lambda_f(t;\mathbf{X},\mathbf{Z}_f) = \lambda_{f,0}(t)\exp\left(\boldsymbol{\beta}_f^T\mathbf{Z}_f + \boldsymbol{\gamma}_f^T(t)\mathbf{X}\right) \qquad f \in \{dfd, oc\}\,. \tag{6}$$

In practice, simple parametric models (such as piecewise constant models) for the center effect as a function of time often provide sufficient flexibility while at least approximating the real changes over patient time. Selection of the number and position of change points may use both subject knowledge (e.g., when inpatient care only lasts for a fixed period) and formal model selection strategies (e.g., by minimizing a cross-validated partial likelihood [26]). More flexible options (e.g., through cubic splines) may require more events in each center, which may be infeasible.

While being slightly more involved, the computation of the ECSCI function can largely proceed as before. The resulting plots of the various $\hat{E}_j(t)$ should be interpreted with care. Even if the underlying cause-specific hazard ratios suddenly change, changes in the cumulative incidence will lag behind as event rates need to accrue over time. Furthermore, many diseases evolve slowly, and the cause-specific hazards may only react to detrimental therapeutic events years after the fact.

2.5. The presence of small centers

In general screening or surveillance of hospitals, centers with only few patients may appear. This may lead to extreme, and unstable, estimates or even fitting problems as monotonicity of the partial likelihood becomes an issue. This may be avoided with an appropriate likelihood penalization, such as the Firth correction [27, 28]. While this reduces the bias for the cause-specific log hazard ratios, such reduction for the derived ECSCI is less clear.

The shrinkage that results from the Firth correction, from other penalizations [29], or with popular hierarchical models, such as random effects models [30, 31], will typically reduce the power to detect abnormal excesses [32]. This is most severe in smaller centers, which may be especially problematic if volume effects are suspected. Recent research suggests that the Firth correction stabilizes the estimates, while shrinking them less than for example normal random effects models do [24, 33]. Other attempts to restore the power have been proposed, for example, by setting up a whole investigative framework [34].

In the Riksstroke application of Section 3, these issues do not play, as all centers there are large enough. Therefore, we do not apply any regularization technique.

3. The Swedish Riksstroke registry

3.1. Introduction to the data

Our method is applied to analyze data from the Swedish stroke registry, Riksstroke (http://www.riksstroke.org/eng/), which contains information from all Swedish hospitals admitting patients with acute stroke, with the aim of monitoring and supporting the quality of care for stroke patients [35]. Our goal here is to evaluate the centers' performance in terms of the newly proposed ECSCI measure, evaluated at 1 year after admission for acute stroke, and to compare this with other, more commonly used, approaches.

The original data set contained both patient-level and center-level covariates for 219 573 acute stroke admissions to 90 hospitals, conveying information on patient mix, treatment received, and hospital type. Consciousness at admission was used as a proxy for disease severity. After restricting ourselves to first admissions between January 1, 2001 and December 31, 2009 (removing 32 103 admissions) and removing a limited number (206) of observations with clear data entry errors, the final database-contained information on 187 264 patients was distributed evenly over time (although slightly fewer patients are registered during the summer months). Among these, 14 464 patients had one or more covariate values missing, mainly in the smoking indicator (14.1%). No pattern was found in the distribution of missingness across centers, although the missingness rates for some variables decreased over time. We added covariate-specific missingness indicators as separate predictors in the model, as missingness was sometimes associated to unconsciousness at admission and thus appeared to be explained in these cases. Tables II and III offer a summary description of the categorical data, and Figure 1 shows the age distribution. Lindmark *et al*. [36] provide more background information on the Riksstroke registry.

Table II. Distribution of the binary covariates in the Riksstroke data set, in percent.

	Coding	0	1	Missing
Sex	Women/men	49.7	50.3	0
Previous stroke	No/yes	79.9	18.2	2.8
Diabetes	No/yes	79.0	19.7	1.2
Atrial fibrillation	No/yes	72.0	25.9	2.1
Dependence in personal activities of daily living	No/yes	88.9	9.3	1.8
Treatment for high blood pressure	No/yes	46.6	51.2	2.3
Current smoker	No/yes	72.5	13.4	14.1
Institutional living	No/yes	90.8	8.6	0.7
Living alone	No/yes	50.1	49.0	0.9

Table III. Summary of the categorical covariates in the Riksstroke data set, in percent.

Stroke subtype	ICD I61	ICD I63	ICD I64	Missing
	11.9	83.6	4.6	0
Consciousness at admission	Alert	Drowsy	Unconscious	Missing
	81.0	12.4	5.4	1.1
Education	Primary	Secondary	University	Missing
	41.0	26.5	10.3	22.2

ICD, International Classification of Diseases.

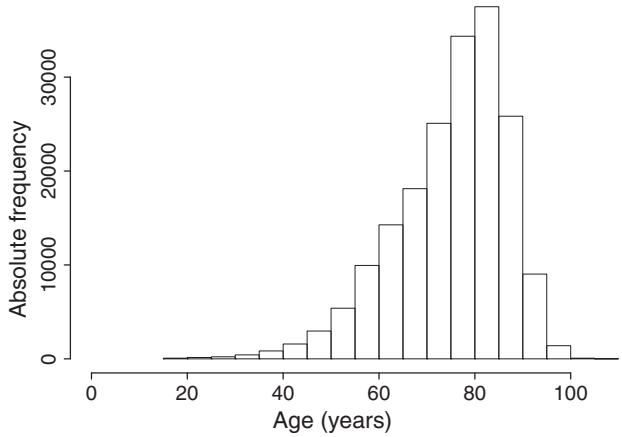


Figure 1. The distribution of the ages in the data set.

We consider two types of death cause: cerebrovascular diseases (International Classification of Diseases coding I60–69) and all others combined, referred to as death from disease (*dfd*) and other cause (*oc*), respectively. Follow-up ended on December 31, 2010, with a median of 5.5 years, for a total of 1 030 312 years of follow-up, during which 36 578 patients died from cerebrovascular disease and 59 076 died from other causes. The overall survival was estimated at 75.0%, 61.4%, and 49.7%, after 1, 3, and 5 years, respectively (standard error $< 0.26\%$ for all).

Only administrative censoring occurs (in 91 610 observations), as the use of the personal identification number allows linkage to the Swedish cause-of-death register, managed by the Swedish National Board of Health and Welfare. However, in other settings, linkage to a death register may not be possible or may only be performed irregularly. Using models that cannot deal with censoring would effectively prohibit yearly reporting in such settings. By evaluating at 1 year after admission in this paper, the death status of all patients is known. For the purpose of this paper, this allows the calculation of quality measures that cannot easily cope with censoring (e.g., based on logistic regression), for comparison with our ECSCI outcome.

3.2. Model building

When risk adjustment was needed, we applied forward model building referring to the two-sided 5% significance level. For the continuous age and income variables, various modeling options were investigated: linear effects, piecewise constant effects with various cut points (inspired by percentiles and residuals), and piecewise linear models with attached knots.

Eventually, for all models (logistic and cause-specific proportional hazards), all categorical covariates of Tables II and III were included (with missing as a separate category, if applicable). Separate dummies for each year were included, as were health care center indicators. The month of admission was dichotomized to indicate winter months (defined as November, December, January, and February), but was not retained in the proportional other cause-specific hazard model because of a lack of significance.

Income (adjusted to the 2009 consumer price index) was consistently modeled through a piecewise constant log hazard ratio at 81 500, 123 600, and 224 200 Swedish Kronor (SEK) (the approximate 10th, 50th, and 90th percentiles). Missing and negative incomes constituted separate categories with constant risk levels. Age was modeled by a piecewise linear model with attached knots at ages 50, 60, and 70 years (the 3rd, 11th, and 27th percentiles). At age 80 years (59th percentile), a disjoint knot allowed for an additional effect of the missingness of education level above this age (as described in [36]), and a final attached knot was applied at age 90 years (88th percentile). The Supporting information shows a schematic of the age dependence allowed by the model.

3.3. Two standard analyses

A first impression of the data per hospital is given by two often-used outcomes: first, the observed mortality at 1 year after admission, and second, a risk-adjusted (indirectly) SMR (RSMR). In the presence of censoring, the former would require a Kaplan–Meier estimate. For the latter, risk adjustment proceeds through a logistic regression model for the death status at 1 year, with the RSMR calculated as

$$RSMR = \frac{\sum_i \frac{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z})}{1+\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z})}}{\sum_i \frac{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z}_0)}{1+\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z}_0)}},$$

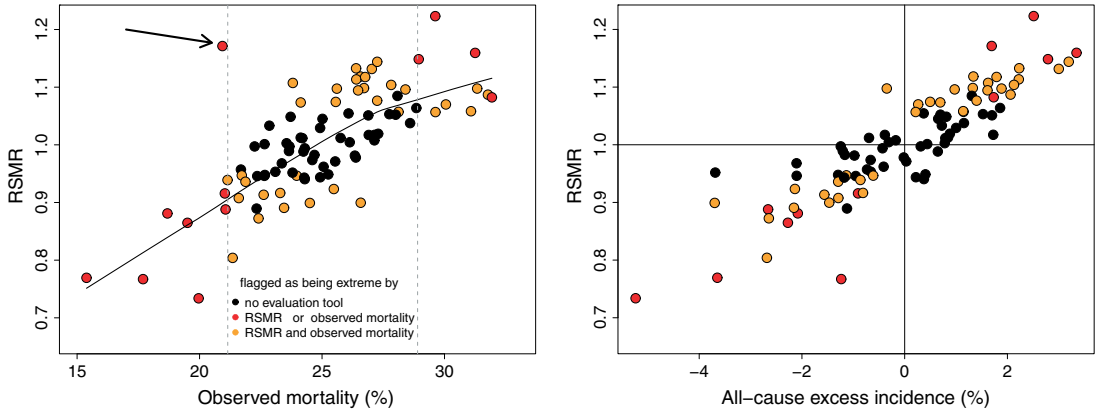
where summations run over all patients in the center under study, $\mathbf{z}$ generically represents the complete set of predictors (including hospital indicators), $\hat{\boldsymbol{\beta}}$ the parameter estimates from the logistic regression, and $\mathbf{z}_0$ the set of predictors, with hospital indicators changed in such a way that the hospital log odds ratio becomes the average over all hospitals of all estimated log odds ratios. In this way, the numerator reflects the number of predicted deaths in the observed hospital, under its current care level, while the denominator reflects an expected number of deaths in the observed hospital, but under the average care level observed over all hospitals. We estimate the uncertainty on the RSMR using a parametric bootstrap with 100 replications.

Often, RSMRs are based on hierarchical logistic regression models [37], mainly to overcome fitting problems related to small centers. This can however decrease the power to detect signals and introduces additional assumptions. Because sample size is not an issue in our data, we simply avoid these issues by using plain logistic regression.

Figure 2(a) shows that the two analyses are clearly correlated, reflecting that the difference between a crude and a risk-adjusted view of the mortality is often small. This is due to stroke being an acute illness for which patients rush to the nearest hospital. This leads to limited selectivity, thus to largely similar case mixes across hospitals, and to summations over roughly similar populations in the RSMRs for all hospitals. This would also lead to some correspondence between direct and indirect standardizations results.

For the crude mortality, a typical analysis flags the top and bottom 10th percentile centers. For the RSMR, results often focus on significance. Figure 2 shows both approaches, where a center is seen in orange if it is either one of the top 9 (mortality under 21.15%) or bottom 9 (mortality above 28.9%) centers in terms of observed mortality, or if its RSMR is significantly different from 1. If both criteria are met, it is represented in red.

In one center (indicated by an arrow in Figure 2(a)), both outcomes clearly differ: the observed mortality is very low, while the RSMR is very high. This is due to a very specific case mix, with more young patients with high socioeconomic status, who are mostly conscious at admission. This results in a low



(a) RSMR vs. observed mortality (b) RSMR vs. all-cause excess incidence

Figure 2. Comparison of overall mortality measures: (a) observed mortality and RSMR, and (b) excess all-cause incidence at 1 year and RSMR.

observed mortality, although it is higher than what could be expected in a similar population receiving average care.

3.4. An additional quality measure: the all-cause excess

Before analyzing the ECSCIs, we present an all-cause excess for a given center j, defined as

$$E_j(t) = \frac{\sum_{i \in \text{center } j} \left(F(t; \mathbf{x}_i, \mathbf{z}_i) - \frac{1}{m} \sum_{c=1}^{m} F(t; x_{i,c} \equiv 1, \mathbf{z}_i) \right)}{n_j},$$

where $\mathbf{z}$ holds the risk factors for all-cause mortality, $x_{i,c} \equiv 1$ is to be interpreted as '$x_{i,c} = 1$, *and all other* $x_{i,c'} = 0$', and where the all-cause cumulative incidence at time t, $F(t, \mathbf{x}_i, \mathbf{z}_i)$ can be derived from an all-cause proportional hazards model (as in our Riksstroke analysis later) or from separate cause-specific proportional hazards models. While this all-cause excess is considerably less complex than the ECSCIs, variance expressions are still difficult and are most easily obtained through bootstrapping.

Arguably, such an excess all-cause incidence is more of direct interest to patients concerned with their lives than the ECSCIs. Furthermore, in some sense, it bridges the gap between the RSMR on the one hand (being a relative measure of all-cause outcomes) and the two ECSCIs on the other (measuring absolute cause-specific differences).

For the Riksstroke data, a relatively strong correlation exists between this all-cause excess (calculated at 1 year of follow-up) and the RSMR, as shown in Figure 2(b). Even so, some large RSMRs lead to limited expected excess deaths and vice versa. For patients, the all-cause excess may be more intuitive: It gives the number of deaths expected above or below the expected average, in 100 treated patients, while the RSMR reflects a proportional change in an unspecified base level mortality.

3.5. Analysis of the ECSCIs

Figure 3 shows the observed cumulative incidences for both causes in the first year after stroke, revealing the relative importance of both death causes. Initially, cerebrovascular deaths occur more often than other-cause deaths, but after a few months, the rate of cerebrovascular deaths drops, and after 1 year, the cumulative incidence of both death types is roughly similar (13.1% for cerebrovascular deaths and 11.8% for the other causes).

The two cause-specific proportional hazards models were built as described earlier. Time-varying covariate effects were not pursued after informal assessment of the Schoenfeld residuals. The two models were identical apart from the omission of the winter month indicator in the other cause model. Results are summarized in Tables 3, 4, 5, and 6 in the Supporting information. For some covariates that are self-reported at admission (e.g., smoking), the hazard in the 'missing' category is much higher. The missingness there may in part be related to the severity of the stroke or even the patients' early death.

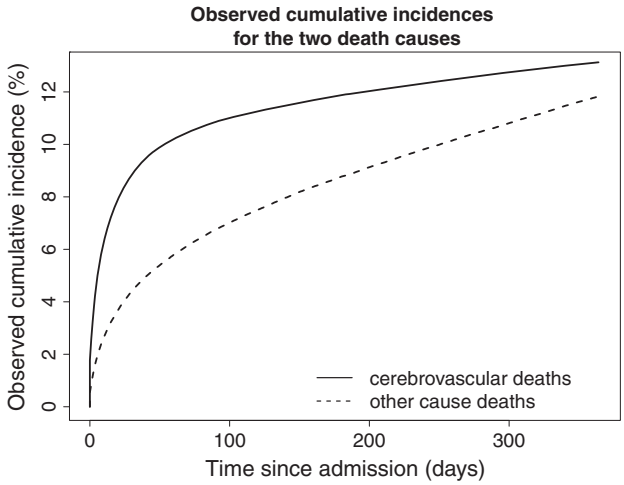


Figure 3. Comparison of both observed cumulative incidences. Full line, cerebrovascular deaths; striped line, other cause deaths.

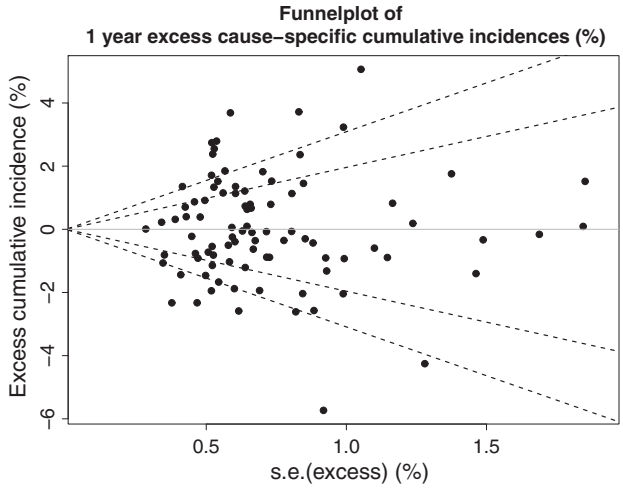


Figure 4. Funnel plot of the stroke-related ECSCIs. The inner lines represent 5% significance tests, and the outer ones 0.2%.

Excess cause-specific cumulative incidences are estimated at 1 year after admission. Their uncertainty is estimated using 100 parametric bootstrap samples, which appeared to be sufficient to yield stable variance estimates.

When objective benchmarks for performance are desired, results are often presented in a funnel plot (Figure 4). Lacking an interpretable parameter that unambiguously drives the variance (which would typically be the center size), we plot the excesses against their bootstrap-estimated standard errors. This reverses the direction of the x-axis, compared with a typical funnel plot, and leads to straight 95% and 99.8% reference lines in Figure 4. By not referring directly to the size of centers on the x-axis, this further protects their anonymity, should this be an issue. More stringent reference levels can be used to protect against inflated family-wise error rates or allow for overdispersion.

With a 1-year cerebrovascular death incidence of approximately 13%, the spread in ECSCIs from plus to minus 5% (with prime concern for high positive values showing increased stroke-related mortality) represents substantial variation between centers. The funnel plot further shows that many more estimates fall outside the boundaries than expected under their respective significance test levels. This is not uncommon for health care evaluation outcomes [38] and reflects the existence of a substantial between-center variation component, even after the initial risk adjustment. Parallel to the difference between statistical and practical significance, the question now becomes, which deviations are acceptable and which are not? Here, the interpretability of the excess outcome proves useful and helps in achieving consensus on

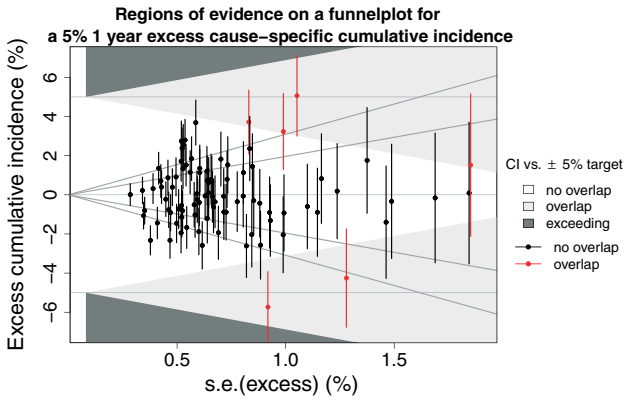


Figure 5. Funnel plot showing the possible overlap between the 95% confidence intervals for the ECSCIs and a target excess level of 5%, and the different regions of evidence.

a well-understood target excess level, deemed important enough to lead to a (provisional) labeling of the center. This level could be added to the funnel plot as a horizontal line.

The choice of such an appropriate level typically depends on many factors, including disease under study, outcome of interest, baseline risk, and so on. With an observed population risk of 13%, an intuitive argument with practical implications could be that a surplus incidence below 5% is acceptable; otherwise, immediate action is needed. Another argument could use the health economic impact of certain excess incidence levels.

Standard hypothesis tests (as in e.g. [39]) exclusively control the type I error. Nevertheless, the cost of not detecting suboptimal performance might be high, a concern that becomes more pressing when low volume centers are at risk of lesser performance. Also, for very large centers, very small observed effects may become significant but may still be far from clinically relevant. One simplified way to address this checks whether 95% confidence intervals fall below a prespecified level of clinically important excess incidence, overlap with such level, or surpass it completely, as already carried out in the context of clinical trials [40]. This leads to different areas of performance on the funnel plot in Figure 5, where 5% ECSCIs are targeted. Confidence intervals lie completely between the two target levels of 5% and −5% in the white area, overlap with at least one of these two levels in the light gray area (red intervals), or lie completely beyond one target level in the dark gray area. Although the confidence intervals add little information to the plot, they clarify that the confidence intervals of significant ECSCIs do not always overlap with the target 5% level. Conversely, for the outer right excess (highest standard error), there is an overlap, reflecting that, with this substantial uncertainty, the true ECSCI value may actually be of the clinically interesting 5% level. This approach may be particularly appropriate for screening purposes, where multiple testing issues [38] force one to consider target rates for different errors explicitly. The discussion further points to a method that can provide a chosen balance between power and type I error in this setting.

3.6. Comparing different-cause excesses

As with cumulative incidences, one should be careful in drawing conclusions from the ECSCI for one death cause in isolation. An increased ECSCI of one event type may indeed result from poor treatment to avoid this event, but could also result from a more efficient treatment against other causes. Therefore, competing risks should be inspected, either through their separate excess, their cumulative incidence, or perhaps even at the level of the cause-specific hazard ratios (and accounting for the relative importance of both cause-specific hazards).

Figure 6 shows the ECSCI of interest as a function of the other cause ECSCI. Most concern lies with the upper right corner where the incidence of both causes is increased, while it is decreased for both causes in the lower left corner. The figure reveals that the magnitude of the excesses of other cause deaths is substantially smaller than that of the cerebrovascular deaths. This suggests that treatment quality differs between hospitals, affecting the probability of (primarily early) stroke death, while the short-term hospitalization has less impact on the long run (primarily other cause) death probability. While some hospitals see a negative excess on one cause, together with a positive excess on the other,

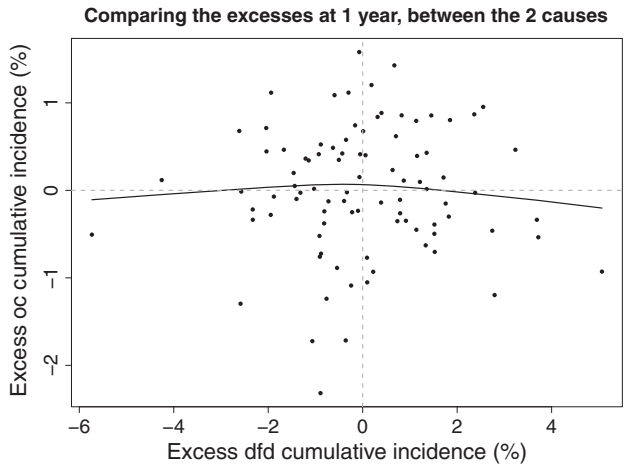


Figure 6. Comparison of the excess measures for both death types, with indication of the local average (full line) and null effects (striped lines).

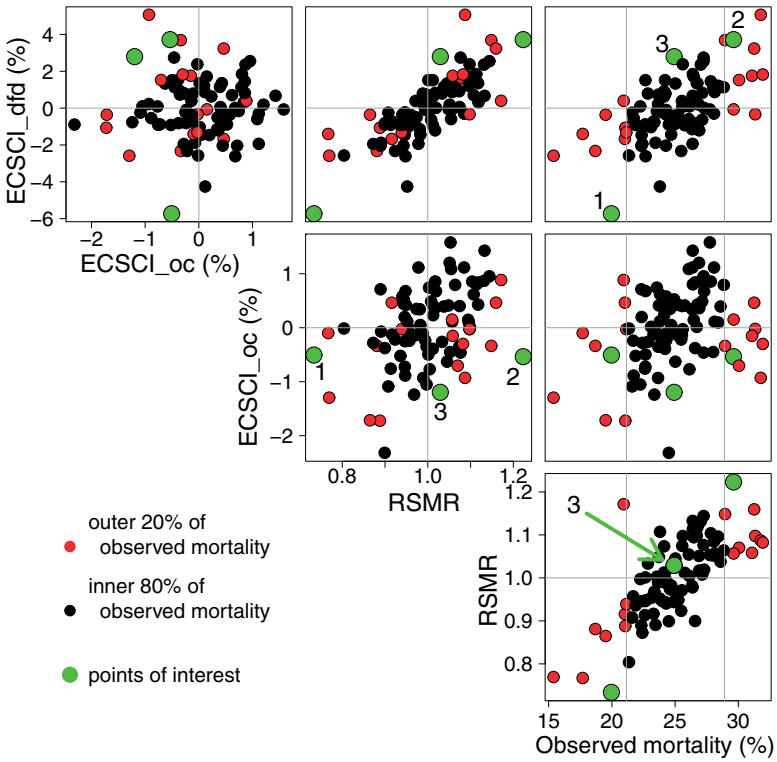


Figure 7. Comparison of the various analyses. Reference lines are drawn at 0% for the two ECSCIs, at 1 for the RSMR and at the 10th and 90th percentile for the observed mortality. The nine centers with the highest, and the nine with the lowest, observed mortality are shown in red; the centers discussed in the text in green.

this is not systematically so. Ergo, there are centers whose improved stroke care also results in higher overall survival.

3.7. Comparison of the ECSCIs and the two standard analyses

Figure 7 shows no overwhelming correlation between the risk-adjusted excesses (named ECSCI_dfd and ECSCI_oc here) that only reflect cause-specific events and the crude overall mortality. Because the other-cause excesses vary less than the stroke-related death excesses, they form the smaller part of the observed variation in crude mortality, leading to a smaller correlation, also with the RSMR.

The added value of the ECSCIs is further illustrated in Figure 7. Hospital 1 (big green dot, lower left corner of the RSMR versus crude mortality panel) scores well both on observed mortality (19.96%) and RSMR (0.734). The ECSCIs suggest that this is mainly due to excellent performance in terms of stroke-related mortality (ECSCI_dfd=-5.73%), while it is more average for the other causes, with an excess of only −0.5%. Hospital 2 (big green dot, upper right corner of the RSMR versus crude mortality panel) has a very high observed mortality (29.62%) and the highest RSMR of all hospitals (1.22). The ECSCIs reveal that it does actually better than average for other causes (with an excess of −0.54%), while the number of stroke-related deaths is strongly increased: with an excess of 3.72%, per 100 patients treated for acute stroke almost four more are expected to have suffered stroke-related deaths after 1 year than under average care.

Apart from specifying which part of the care process is likely running exceptionally well or poor, the ECSCIs' values are also more directly informative than the RSMRs. Hospital 3 lies central in the RSMR versus crude mortality panel and is not significant in terms of RSMR nor is it in either of the outer 10th percentiles of crude mortality. However, looking at the two ECSCIs, it shows both one of the largest stroke-related excesses (2.79%) and one of the largest negative other-cause excesses (−1.20%). These largely offset each other when recombined into an all-cause analysis. While ECSCIs allow such direct comparison of the effect sizes on the two causes, this is more difficult with for example two cause-specific SMRs. Still, evaluations like these, distinguishing among deaths causes, are important in guiding processes evaluations in a hospital.

3.8. Comparison of the ECSCIs and the all-cause excess incidence

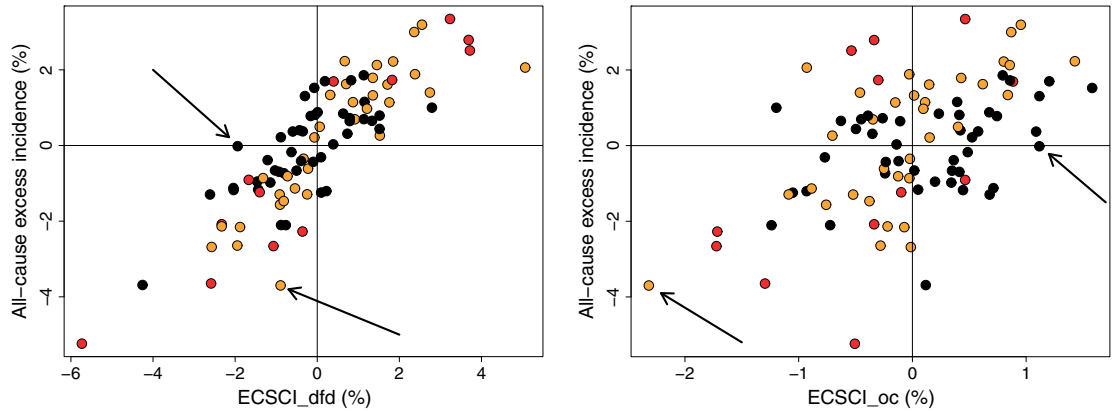
The combined information in the cause-specific ECSCIs matches that available in an all-cause excess, as evidenced by the correlation of 97.7% between the all-cause excess and the sum of the two ECSCIs. Furthermore, the values of the two are very close (as seen in Figure 5 in the Supporting information). Small deviations arise from differences in the underlying models.

While the all-cause excess incidence is appealing to patients, for policy makers, a decomposition into cause-specific excesses may be useful, as it localizes problems with delivered care more precisely. The comparison between the two separate ECSCIs and the all-cause excess in Figure 8 shows this.

For example, Figure 8(a) shows one center (indicated by an arrow) with very low all-cause excess (indicating decreased mortality), while its dfd-specific ECSCI is rather average. This indicates that it is mainly the general care in this facility that is excellent (as panel (b) confirms), while the stroke specific care is merely up to standard. In another indicated center, the all-cause excess is practically 0, masking the fact that the stroke-related mortality is in fact one of the lowest, while it is the other-cause mortality that is increased, possibly indicating problems with the general care.

3.9. Using the cause-specific hazard ratios as evaluation criteria

Contrasting the ECSCIs with the log cause-specific hazard ratios (the γs) may be interesting, because they are alternative quality measures that are much easier to calculate. Still, while these hazard ratios



(a) death from disease (b) other causes

Figure 8. The decomposition of the all-cause excess into two parts reflected by the ECSCIs.

offer a purer perspective on underlying effects playing primarily on one type of outcome, their meaning is hard to operationalize and often leads to confusion [41–43].

The dependence between both measures is clear in our setting, as illustrated by a correlation of 99.1% between the excess stroke-specific incidence and the log stroke-specific hazard ratios (γ_{dfd}) associated to the various centers, and a correlation of 90.9% between the excess other cause-specific incidences and the log other cause-specific hazard ratios (γ_{oc}). The latter correlation is weaker, as the other-cause effects are less variable. The comparison of these various quantities is shown in Figure 7 of the Supporting information.

Because the agreement between both approaches is not perfect and depends on the setting, and because the interpretation of the excesses is more directly useful by relating to an observable risk, we feel that it is worth reporting the ECSCI, instead of (or at least next to) the cause-specific hazard ratios.

4. Discussion of the methodology and concluding remarks

We have presented a novel approach to the evaluation of disease-specific outcomes among health care providers, based on a cumulative incidence in excess of the expected incidence for the same population, under the average care level across all hospitals. The contribution of this new measure (which we call an ECSCI) lies in its intuitive interpretation, grounded in the clinical experience with a hospital's observed patient mix: it is directly relevant to clinical practice and provides a basis for meaningful discussion among stakeholders, for example, on appropriate benchmarks and standards to strive for. Building the measure on standard cause-specific hazards models facilitates the statistical approach to case-mix adjustment.

Compared with SMRs, a considerable advantage of our measure is its additive nature on the probability scale. In the relative SMR measure, large effects may still refer to a minor impact on numbers of deaths, when the baseline is low. As such, these do not necessarily indicate clinically important differences.

Generally speaking, an excess incidence does not necessarily indicate a number of 'avoidable deaths', an issue that may sometimes be misunderstood [44]. There is always the caveat that baseline risk adjustment may not be adequate; for example, socioeconomic factors and comorbidities are rarely fully accounted for. Even with adequate risk adjustment, observed death numbers are subject to random fluctuations. Furthermore, a careful follow-up analysis (that goes beyond a purely statistical investigation) should identify factors involved when a high excess is observed. While some of these factors may be influenced through hospital policy, this is not always the case, and there may not always be room for improvement. Nevertheless, we feel that from a technical, statistical perspective, naming our outcome an 'excess' seems appropriate, and warn against overly simplified and far-reaching interpretations.

In risk adjustment models, covariate-by-center interactions locate care improvement needs in specific patient groups, but their inclusion also quickly leads to computational problems. Very recent research [45] shows that, when patient mixes are roughly similar between centers, omitting existing interactions from a risk adjustment model has little impact on derived standardized mortalities (be it direct or indirect standardization). We expect that ECSCIs behave similarly, allaying concerns about possible unmodeled interactions in the Riksstroke setting, where patient mixes are similar because of the acute need for treatment after stroke. When interactions are present in the model, separate ECSCIs for different patient groups illustrate that, while excellent care saves lives in one patient group, lives are lost by suboptimal care in another. Such ECSCIs are obtained by restricting the summation in Equation (5) to the relevant patients. Similarly, the individual patient ECSCIs from Equation (4) can be plotted against continuous risk factors (e.g., age).

Although our presentation focused on the use of proportional cause-specific hazards models, the proposed excess measure is generic and can use any model to obtain the cumulative incidence. When the proportional cause-specific hazards assumption is deemed incorrect, one could use extensions of the Cox model (e.g., including time-varying effects) or apply an additive hazards model. One obvious alternative is the Fine and Gray model [10], which assumes proportionality on the sub-distribution hazard scale. Vertical modeling also allows the modeling of the cumulative incidence in a more direct way, by explicitly separating the modeling of event time and event type [46]. When focusing on the evaluation at one specific time point t, one could fit models by minimizing the prediction error at t [47]. A decisive criterion to choose between options may be their predictive strength at time t. Finally, having reliable diagnoses of death causes is essential to cause-specific analyses. In practice, however, administrative and diagnostic

errors happen. When their prevalence is known, the excess evaluation can proceed using the correction method in [48].

The excess measure forms a useful basis for health economic evaluations. When the one-time cost to patient, hospital, or society linked to the occurrence of a specific event is known, multiplying the ECSCI with that expected cost and the number of patients in a center translates into monetary value lost (or gained) due to center performance. When the repercussion of the occurrence of an event is accumulating over time, viewing the excess as a function of time allows one to estimate a total number of event-free years gained and the associated cost benefits (as in [49]). Similarly, with an appropriate weighting scheme, the quality-adjusted life years gained can be estimated. However, as the single-cause excess does not translate exactly into life years gained, one will need to integrate over both cause-specific excess measures.

While we have restricted ourselves to a health care setting, quality evaluations with similar issues arise in other settings. Education evaluations may, for example, study the time to reach a specific educational attainment as a measure of quality. When dropout for various reasons comes into play (e.g., [50]), the developed methodology could serve there as well.

Future work will focus on the most appropriate way to track hospital performance over calendar time. Choosing an appropriate benchmark is challenging, as the comparison of excesses over different years requires a relatively stable reference. Depending on the choice made (e.g., an average over all centers admitting patients in a 5-year window), correlations between excesses may complicate their formal comparison. One may base the comparator on the past performance record of the center under study, rather than on the average contemporary performance over all other centers. Again, the covariances appearing in such a longitudinal development need to be studied in more detail.

The formal benchmarking of health care quality is another open challenge. The current focus lies on standard hypothesis tests (e.g., [51] or [52]), which are primarily concerned with the protection against type I errors, that is, not wrongly labeling a center as 'different from the norm'. However, missing poor quality of care is at least as important when seeking care improvement, highlighting a concern for type II errors. The relative importance of both types of error also depends on their impact in the setting at hand, for example, depending on the resulting action (either confidential feedback or public dissemination of results), or on the outcome under study (e.g., weight loss after a gastric bypass or mortality after cardiac surgery). Furthermore, a p-value-driven approach may not always target outcomes with practical importance, either by selecting small effects measured with great precision or by not picking up large effects measured in small centers. A so-called *balanced* test [53] can address such issues in a flexible way, by striking a desired balance between type I and type II errors for a meaningful effect size. This would use the more informative approach to specify in advance what amount of t-year excess risk is reasonably acceptable for a given disease in a given setting and then targets detection of 'significant' deviations from this. Because of its intuitive interpretation, the ECSCI presented here is well suited for the specification of such effect size of interest.

The excess measure presented in this paper hopes to provide a useful addition to the currently used methodology. The promise of care improvement through evaluations is enhanced when clear and intuitive outcomes are used, allowing for meaningful discussions on relevant deviations from the norm.

Acknowledgements

The authors thank the members of the Riksstroke Collaboration (http://www.riksstroke.org) for their support. B. V. R. was supported through a research fellowship from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Vlaanderen). B. V. R. and E. G. acknowledge support from IAP research network grant no. P07/06 from the Belgian government (Belgian Science Policy). E. G. and M. E. acknowledge support from the Swedish Research Council (grant no. 2012-5934).

References

1. Bird S, Cox D, Farewell V, Goldstein H, Holt T, Smith P. Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society - Series A* 2005; **168**:1–27.
2. Lamb G, Smith M, Weeks W, Queram C. Publicly reported quality-of-care measures influenced Wisconsin physician groups to improve performance. *Health Affairs* 2013; **32**:536–543.
3. Mehta R, Peterson E, Califf R. Performance measures have a major effect on cardiovascular outcomes: a review. *American Journal of Medicine* 2007; **120**:398–402.

4. Shahian D, Edwards F, Jacobs J, Prager R, Normand S, Shewan C, O'Brien S, Peterson E, Grover F. Public reporting of cardiac surgery performance: part 1 – history, rationale, consequences. *Annals of Thoracic Surgery* 2011; **92**:S2–S11.
5. Berwick D, James B, Coye M. Connections between quality measurement and improvement. *Medical Care* 2003; **41**: 130–138.
6. Moore L, Hanley J, Turgeon A, Lavoie A, Eric B. A new method for evaluating trauma centre outcome performance: TRAM-adjusted mortality estimates. *Annals of Surgery* 2010; **251**:952–958.
7. Van Rompaye B, Goetghebeur E, Jaffar S. Design and testing for clinical trials faced with misclassified causes of death. *Biostatistics* 2010; **11**:546–558.
8. Putter H, Fiocco M, Geskus R. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 2007; **26**:2389–2430.
9. Hernán M. The hazards of hazard ratios. *Epidemiology* 2010; **21**:13–15.
10. Fine J, Gray R. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 1999; **94**:496–509.
11. Bekaert M, Vansteelandt S, Mertens K. Adjusting for time-varying confounding in the subdistribution analysis of a competing risk. *Lifetime Data Analysis* 2010; **16**:45–70.
12. Andersen P, Borgan O, Gill R, Keiding N. *Statistical Models Based on Counting Processes*. Springer-Verlag: New York, 1993.
13. Cheng S, Fine J, Wei L. Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* 1998; **54**:219–228.
14. Beyersmann J, di Termini S, Pauly M. Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics* 2012; **40**(3):387–402.
15. Lipsitz S, Dear K, Zhao L. Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* 1994; **50**:842–846.
16. Shao J, Tu D. *The Jackknife and Bootstrap*. Springer-Verlag: New York, 1995.
17. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
18. Klein J, Andersen P. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function, 2005.
19. Allignol A, Schumacher M, Wanner C, Drechsler C, Beyersmann J. Understanding competing risks: a simulation point of view. *BMC Medical Research Methodology* 2011; **11**:86.
20. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Statistics in Medicine* 2009; **28**:956–971.
21. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**:1713–1723.
22. de Wreede L, Fiocco M, Putter H. mstate: an R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software* 2012; **38**:1–30.
23. Rosthø j S, Andersen P, Abildstrom S. SAS macros for estimation of the cumulative incidence functions based on a Cox regression model for competing risks survival data. *Computer Methods and Programs in Biomedicine* 2004; **74**:69–75.
24. Goetghebeur E, Van Rossem R, Baert K, Vanhoutte K, Boterberg T, Demetter P, De Ridder M, Harrington D, Peeters M, Storme G, Verhulst J, Vlayen J, Vrijens F, Vansteelandt S, Ceelen W. Kwaliteit van rectale kankerzorg - fase 3: statistische methoden om centra te benchmarken met een set van kwaliteitsindicatoren. Good clinical practice (GCP), Technical Report *KCE Report 161A*, Federaal Kenniscentrum voor de Gezondheidszorg (KCE). D/2011/10.273/38, 2011.
25. Greenland S, Robins J. Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology* 1988; **128**:1185–1197.
26. Verweij P, van Houwelingen H. Cross-validation in survival analysis. *Statistics in Medicine* 1993; **12**:2305–2314.
27. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**:27–38.
28. Heinze G, Schemper M. A solution to the problem of monotone likelihood in Cox regression. *Biometrics* 2001; **7**:114–119.
29. Ambler G, Seaman S, Omar R. An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in Medicine* 2012; **31**:1150–1161.
30. Normand S, Glickman M, Gatsonis C. Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association* 1997; **92**:803–814.
31. DeLong E, Peterson E, DeLong D, Muhlbaier L, Hackett S, Mark D. Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* 1997; **16**:2645–2664.
32. Ash A, Fienberg S, Louis T, Normand S, Stukel T, Utts J. Statistical issues in assessing hospital performance. Commissioned by the Committee of Presidents of Statistical Societies, 2012. (Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf) [Accessed on 17 November 2014].
33. Varewyck M, Els Goetghebeur E, Eriksson M, Stijn Vansteelandt S. On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics* 2014; **15**(4):651–664.
34. Ohlssen D, Sharples L, Spiegelhalter D. A hierarchical modelling framework for identifying unusual performance in health care providers. *Journal of the Royal Statistical Society - Series A* 2007; **170**:865–890.
35. Asplund K, Hulter Asberg K, Appelros P, Bjarne D, Eriksson M, Johansson A, Jonsson F, Norrving B, Stegmayr B, Terént A, Wallin S, Wester P. The Riks-Stroke story: building a sustainable national register for quality assessment of stroke care. *International Journal of Stroke* 2011; **6**:99–108.
36. Lindmark A, Glader E, Asplund K, Norrving B, Eriksson M. Socioeconomic disparities in stroke case fatality – observations from Riks-Stroke, the Swedish stroke register. *International Journal of Stroke* 2012; **9**(4):429–436.
37. Shahian D, Normand S. Comparison of 'risk-adjusted' hospital outcomes. *Circulation* 2008; **117**:1955–1963.
38. Spiegelhalter D. Funnel plots for comparing institutional performance. *Statistics in Medicine* 2005; **24**:1185–1202.

39. Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, Blunt I, Wood C, Grigg O. Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society - Series A* 2012; **175**:1–25.
40. Man-Son-Hing M, Laupacis A, O'Rourke K, Molnar F, Mahon J, Chan K, Wells G. Determination of the clinical importance of study results – a review. *Journal of General Internal Medicine* 2002; **17**:469–476.
41. Pintilie M. Analysing and interpreting competing risk data. *Statistics in Medicine* 2007; **26**:1360–1367.
42. Latouche A, Beyersmann J, Fine J. Comments on analysing and interpreting competing risk data. *Statistics in Medicine* 2007; **26**:3676–3680.
43. Wolbers M, Koller M. Comments on analysing and interpreting competing risk data (original article and authors reply). *Statistics in Medicine* 2007; **26**:3518–3523.
44. Spiegelhalter D. Have there been 13 000 needless deaths at 14 NHS trusts? *BMJ* 2013; **347**:f4893.
45. Varewyck M, Els Goetghebeur E, Eriksson M, Stijn Vansteelandt S. The impact of (ignoring) interactions in evaluating clinical center performance. *Presented at the Joint Statistical Meetings 2014*, Boston, U.S., 2014, 55–56. (Available from: http://www.amstat.org/meetings/jsm/2014/onlineprogram/AbstractDetails.cfm?abstractid=311941) [Accessed on 17 November 2014].
46. Nicolaie M, van Houwelingen H, Putter H. Vertical modeling: a pattern mixture approach for competing risks modeling. *Statistics in Medicine* 2010; **29**:1190–1205.
47. Uno H, Cai T, Tian L, Wei L. Evaluating prediction rules for *t*-year survivors with censored regression models. *Journal of the American Statistical Association* 2007; **478**:527–537.
48. Van Rompaye B, Jaffar S, Goetghebeur E. Estimation with Cox models cause-specific survival analysis with misclassified cause of failure. *Epidemiology* 2012; **23**:194–202.
49. Rapsomaniki E, White I, Wood A, Thompson S. The Emerging Risk Factors Collaboration. A framework for quantifying net benefits of alternative prognostic models. *Statistics in Medicine* 2012; **31**:114–130.
50. Zhao M, Glewwe P. What determines basic school attainment in developing countries? Evidence from rural China. *Economics of Education Review* 2010; **29**:451–460.
51. Ferris T, Torchiana D. Public release of clinical outcomes data – online CABG report cards. *New England Journal of Medicine* 1999; **363**:1593–1595.
52. New York State Department of Health. Adult cardiac surgery in New York State 1998–2000, Technical Report, New York State Department of Health Albany, 2004.
53. Moerkerke B, Goetghebeur E. Selecting "significant" differentially expressed genes from the combined perspective of the null and the alternative. *Journal of Computational Biology* 2006; **13**:1513–1531.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

