Statistics
in Medicine

Research Article

Received: 3 May 2016, Accepted: 9 November 2016 Published online 25 November 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7188

The median hazard ratio: a useful measure of variance and general contextual effects in multilevel survival analysis

Peter C. Austin,[a,b,c*†] Philippe Wagner[d,e] and Juan Merlo[d,f]

Multilevel data occurs frequently in many research areas like health services research and epidemiology. A suitable way to analyze such data is through the use of multilevel regression models (MLRM). MLRM incorporate cluster-specific random effects which allow one to partition the total individual variance into between-cluster variation and between-individual variation. Statistically, MLRM account for the dependency of the data within clusters and provide correct estimates of uncertainty around regression coefficients. Substantively, the magnitude of the effect of clustering provides a measure of the General Contextual Effect (GCE). When outcomes are binary, the GCE can also be quantified by measures of heterogeneity like the Median Odds Ratio (MOR) calculated from a multilevel logistic regression model. Time-to-event outcomes within a multilevel structure occur commonly in epidemiological and medical research. However, the Median Hazard Ratio (MHR) that corresponds to the MOR in multilevel (i.e., 'frailty') Cox proportional hazards regression is rarely used. Analogously to the MOR, the MHR is the median relative change in the hazard of the occurrence of the outcome when comparing identical subjects from two randomly selected different clusters that are ordered by risk. We illustrate the application and interpretation of the MHR in a case study analyzing the hazard of mortality in patients hospitalized for acute myocardial infarction at hospitals in Ontario, Canada. We provide R code for computing the MHR. The MHR is a useful and intuitive measure for expressing cluster heterogeneity in the outcome and, thereby, estimating general contextual effects in multilevel survival analysis.

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: Median Hazard Ratio; Median Odds Ratio; clustered data; multilevel analysis; frailty models; survival analysis

1. Introduction

Data with a multilevel nature occur frequently in public health, health services research, behavioral research, and in epidemiology. Examples include residents nested or clustered within neighborhoods or regions, patients nested within hospitals, and students nested within schools. The existence of multilevel information is relevant in the practice of epidemiology for both formal statistical and substantive epidemiological reasons, and a suitable way of analyzing these kinds of data is by using multilevel regression models (MLRM) [1–4].

From a statistical perspective, a condition for performing conventional regression analyses is the independence of the observations. Therefore, conventional analyses are inappropriate in the presence of clustering because subjects within the same cluster are more likely to have similar outcomes compared

[a] *Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*
[b] *Institute of Health Management, Policy, and Evaluation, University of Toronto, Toronto, Ontario, Canada*
[c] *Schulich Heart Research Program, Sunnybrook Research Institute, Toronto, Canada*
[d] *Unit for Social Epidemiology, Faculty of Medicine, Lund University, Malmö, Sweden*
[e] *Centre for Clinical Research Västmanland, Uppsala University, Uppsala, Sweden*
[f] *Center for Primary Health Care Research, Region Skåne, Malmö, Sweden*
**Correspondence to: Peter Austin, Institute for Clinical Evaluative Sciences, G106, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5, Canada.*
[†] *E-mail: peter.austin@ices.on.ca*

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

to randomly selected subjects from different clusters, and thus subjects within the same cluster are not statistically independent of one another. This lack of independence decreases the effective sample size, so that failure to account for clustering falsely increases the precision of the estimates. MLRM have been developed to fit regression models to data that have a multilevel structure [1,2,4,5]. Such models incorporate cluster-specific random effects that account for the dependency of the data by partitioning the total individual variance into variation due to the clusters or 'higher-level units' and the individual-level variation that remains [6].

It is important to remember that MLRM have been used for the analysis of repeated measurements within individuals. In this case, the between-cluster (i.e., between individuals) component of variance is typically a large proportion of the total observed variation in the response variable. This is to be expected as the human body is a natural and homogenous system with well delimited boundaries [7], so the intra-individual correlation of the repeated measurement is anticipated to be high. Therefore, when performing regression analyses of repeated measurements in a sample of individuals, the existence of a strong intra-individual residual correlation can be seen as a nuisance that needs be taken into account for correct estimation of uncertainty around fixed effects. However, the situation is very different when MLRM are applied for investigating contextual effects on individual health. From an epidemiological perspective, knowing the share of the total individual variance that is attributable to the cluster level or the 'variance partition coefficient' (VPC) is relevant information on its own. Obviously, the higher the VPC, the more relevant the context appears to be for understanding individual health or disease outcomes [5,6]. The VPC informs on the existence of a *general contextual effect* (GCE), which is called 'general' because it reflects the influence of the cluster context as a whole, without specifying any contextual characteristic other than the very boundaries that delimit the cluster.

As we have discussed elsewhere [8], quantifying the GCE is actually an implicit goal in all studies evaluating hospital performance, even in those studies that do not apply MLRM. Many such studies are based on single-level models analyzing individual patient data with dummy variables for the hospitals or using information aggregated at the hospital level. Even when information is available at the patient, institutional, and geographical levels, assessments are typically performed at a single level by using, for instance, funnel plots, health league tables, or similar methods to compare hospitals or small geographical area averages. This single-level approach is used extensively, but may provide misleading information for decision makers as it does not provide information on the size of the GCE. For evaluating hospital performance, what matters most is not the existence of differences between hospital averages in some quality indicator, but rather the share of the differences in patient outcomes that are due to the hospital level. That is, the GCE. Obviously, the higher that this share is, the more relevant the hospital context is.

In health services research, and particularly within publicly funded systems, a fundamental task is the assessment of equity in health care. Equity means that health care is provided on equal terms and according to needs, regardless of the location of the patient or of the hospital in which they are treated. In other words, after adjusting for possible differences in case-mix between health care providers, an equitable and well-functioning healthcare system should result in comparable outcomes across different healthcare providers following provision of medical or surgical care. Using this perspective, we have recently suggested [8,9] that for the evaluation of health care performance in a health care system (e.g., Canada or Sweden) it is necessary to consider at least two fundamental parameters: the national average of the quality indicator under analysis (e.g., mortality after hospitalization for acute myocardial infarction (AMI or heart attack)), and the size of the GCE. The use of these two parameters allows us to consider four different categories of health system performance.

The ideal scenario (A) is when the national average denotes high quality (e.g., low mortality after hospitalization for AMI) and the GCE is very small (that is, a very low VPC). Under this scenario, all hospitals are performing homogenously well. That is, there is a well-functioning healthcare system without disparities between hospitals. The worse scenario (B) is when the national average denotes low quality (e.g., high mortality after hospitalization for AMI) and the GCE is also small (that is, a very low VPC). Under this scenario, all hospitals are performing homogenously poorly. That is, there is an overall poorly functioning healthcare system. In both scenarios A and B, any intervention to improve quality (scenario B) or to maintain the high quality of the health care system (scenario A) should not be directed at specific hospitals, but to all hospitals in the state, province, or country. The next scenario (C) is when the national average denotes high quality (e.g., low mortality after hospitalization for AMI) but the GCE is large (that is, a high VPC). This scenario indicates that even if the quality of the whole health care system is, on average, good, there is heterogeneity in hospital performance, so that the prognosis after AMI is consistently poor for the patients in some hospitals but consistently good for

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

patients in some other hospitals. In this scenario, an intervention to improve health care quality should be focused on the hospitals with poor quality (i.e., those with high average mortality after hospitalization for AMI). The final scenario (D) is when the national average denotes poor quality (e.g., high average mortality after hospitalization for AMI in the country) and the GCE is large (that is, a high ICC). This scenario indicates that even if the quality of the health care system is poor on average, there is heterogeneity in hospital performance so that the prognosis after myocardial infarction is consistently bad for the patients in some hospitals but consistently better for patients in other hospitals. In scenario D, an intervention to improve health care quality should be focused on all hospitals in the country and especially on those with the highest average mortality after hospitalization for AMI.

This simplified approach of four scenarios can be applied together with classical funnel plots, health league tables, or similar method commonly used to compare hospitals or small geographical area averages [8–11]. It can also be used for the non-randomized evaluation of interventions in health care [12,13]. When using MLRM with more than two levels (e.g., patient, wards, and hospitals [8], or patient, physicians, and hospitals [13]), the level-specific GCE may provide information on the relative important or relevance of the different levels. This approach may be used to complement the information provided by initiatives like ‘hospital report cards’, in which patients’ outcomes are compared across hospitals within a given health care system. Such report cards have been published comparing mortality rates across hospitals for patients hospitalized with AMI in the Canadian province of Ontario [14], and in the American states of California [15] and Pennsylvania [16]. Similarly, outcomes have been compared across hospitals or surgeons for patients undergoing coronary artery bypass graft (CABG) surgery in Ontario [17] and in the American states of New York [18], New Jersey [19], Pennsylvania [20], and Massachusetts [21]. Similar analyses may be conducted in education (e.g., examining variation in student test scores across schools) and other areas in which it is important to quantify the GCE of the providers. Any researcher examining variation in outcomes across clusters in which subjects are nested (e.g., hospitals, schools, geographic regions) should be interested in formally describing and interpreting the observed variation in outcomes across clusters and the GCE.

A classical concern in health services research is the issue of how to interpret the magnitude of the estimated between-cluster variation. In doing so, one must consider Diehr *et al.*’s classical question, ‘What is too much variation?’ [22]. The simplest approach of directly interpreting the estimated cluster variance, as is classically done in single-level studies of hospital or small area variation, is insufficient. Our analytical framework, however, facilitates the answering of this classical question. Rather than focusing on cluster variance in isolation, we consider the existence of a multilevel continuum of individual variance that can be decomposed into between- and within-cluster components. Therefore, the cluster variance is large when it is a large share of the total individual variance. The advantage of using the VPC to quantify the GCE is that it provides a measure with clearly defined limits as it can be expressed as a percentage that extends from 0% to 100%.

The VPC is easy to calculate and interpret in multilevel linear regression models with continuous outcomes [23]. In the case of discrete responses, the calculation and interpretation of the VPC are more complicated because, among other issues, the individual and cluster components of variance are on different scales. There are, however, several alternative approaches for computing the VPC with discrete outcomes. These include using a normal response approximation, the simulation method, Taylor series linearization, and the latent response method [6]. Also, one could use random effects-based predictions and their corresponding area under the ROC curve [8,9]. A problem is that many epidemiological practitioners and physicians are not familiar with concepts like underlying latent variables and constant individual variances of $\frac{\pi^2}{3}$, which may explain the reluctance to use this approach.

To avoid the interpretative technicalities of the VPC, Larsen et al. introduced the concept of the median odds ratio (MOR) that can be used in the interpretation of GCE when fitting a random effects logistic regression model [24,25]. The MOR indicates the median value of the odds ratios obtained when comparing the odds of the occurrence of the outcome in an individual from a randomly selected clusters with another individual with identical covariates but randomly selected from a different cluster when the clusters are ordered by risk. In other words, to calculate the MOR we should first measure the odds of the occurrence of the outcome for all randomly taken pairs of individuals from different clusters and, thereafter, compute the odds ratio for each pair of individuals having the individual from the cluster with the higher odds in the numerator and the individual from the cluster with the lower odds in the denominator. This would produce a distribution of odds ratios that are always equal to or higher than 1 and the MOR is the median OR of this distribution [26]. The MOR can be thought

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

of as the median increase in the odds of the occurrence of the outcome that would arise when an individual moves from a lower-risk cluster to a higher-risk cluster. An advantage to the MOR is that it permits the analyst to present the between-cluster variation as a measure of association (i.e., an odds ratio) and thereby allows the comparison of the GCE with the fixed effect of the covariates in the model.

Unlike the VPC which is a measure of homogeneity and intra-cluster correlation within hierarchical data structures, the MOR is strictly a measure of heterogeneity between clusters. Also, the MOR is a probabilistic measure of association that extends from 1 to +∞ rather than from 0 to 1 as does the VPC. When interpreting the MOR, we need to consider that both the MOR and the VPC measures are simply functions of the cluster variance and both express the same GCE. For instance, a VPC as low as 2% corresponds to a MOR of 1.28, which some epidemiologist may interpret as a 28% increased risk, which is actually a low MOR.

An important, but neglected, issue concerns the calculation of the GCE for survival or time-to-event outcomes that occur frequently in the medical and epidemiological literature [27]. In fact, the quantification of the GCE is infrequent with time-to-event outcomes even if the Median Hazard Ratio (MHR), as an extension of the MOR, is available for this purpose. While the MHR was empirically applied in 2007 to examine geographic variation in ischemic heart disease mortality in Sweden [28], a formal derivation and interpretation of the MHR was first published by Lanke in 2010 as a short appendix written in Historical Methods [29] to quantify the impact of the family-specific frailty in southern Sweden during the period 1766–1895 [30].

The MOR, described by Larsen in the year 2000 in Biometrics [24], only started being applied regularly in the medical and epidemiological research after 2005 when the concept was introduced and its utility demonstrated for an epidemiological audience [25,26]. This highlights the relevance and importance of translational studies such as this one to introduce to an epidemiologic audience methods developed elsewhere. Introducing advanced statistical techniques and concepts like MLRA to everyday epidemiological practice is a relevant task, not only for improving the validity of the epidemiological analysis, but also because statistical ideas may transform the way medical epidemiologists interpret information.

The objective to the current paper was two-fold. First, to introduce the MHR to researchers in epidemiology and biostatistics. Second, to illustrate its utility for measuring the general impact of the hospital context (i.e., general contextual effects) on the hazard of death in patients hospitalized for an AMI in Canada. The paper is structured as follows. In Section 2, we briefly review frailty models for analyzing clustered survival data and define the MHR. In Section 3, we provide a case study in which we illustrate the utility of this metric for assessing the magnitude of the general contextual (i.e., hospital) effects when analyzing survival data. Finally, in Section 4, we summarize our report and place it in the context of the existing literature.

2. The median hazard ratio

In this section we first formally define the MOR. We then define the MHR.

2.1. The median odds ratio

The MOR was defined by Larsen et al. for use with a random effects logistic regression model [24], and was subsequently popularized in the epidemiological literature [25,26]. Assume that the following random effects logistic regression model had been fit:

$$\operatorname{logit}\left(p_{ij}\right) = \alpha_0 + \beta \mathbf{X}_{ij} + \alpha_j \tag{1}$$

where, p_{ij} is the probability of the occurrence of the binary outcome for the ith subject in the jth cluster, $\mathbf{X}_{ij}$ denotes a vector of explanatory variables, β denotes the vector of associated regression coefficients, and α_j denotes the cluster-specific random effects. The assumption is typically made that the random effects follow a normal distribution: $\alpha_j \sim \mathrm{N}(0, \sigma^2)$. For this random effects logistic regression model, the MOR may be calculated as $\mathrm{MOR} = \exp\left(\sqrt{2\sigma^2}\Phi^{-1}(0.75)\right)$, where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function.

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

2.2. The median hazard ratio

Lanke, in an appendix [29] to a paper [30] published in the history literature, extended the concept of the MOR for use with survival or time-to-event outcomes when Cox frailty models are fit to account for the clustered nature of the data. We assume a Cox proportional hazards regression model that has incorporated cluster-specific random effects:

$$\log\left(h_{ij}(t)\right) = \log(h_0(t)) + \beta\mathbf{X}_{ij} + \alpha_j \tag{2}$$

where $h_{ij}(t)$ denotes the hazard function for the ith subject within the jth cluster, while $h_0(t)$ denotes the baseline hazard function (i.e., the hazard function for a subject whose covariates are all equal to zero). Furthermore, the vector $\mathbf{X}_{ij}$ denotes a vector of predictor or explanatory variables, while β denotes the vector of associated regression coefficients. The α_j denotes the cluster-specific random effects. The model can also be written in multiplicative form:

$$h_{ij}(t) = h_0(t)e^{\beta\mathbf{X}_{ij}}e^{\alpha_j}. \tag{3}$$

When using the multiplicative formulation, the term e^{α_j} is referred to as a frailty term [31,32]. The frailty terms have a multiplicative effect on the hazard function.

Frailty models are described by the distribution of the frailty terms. When the distribution of the random effects is normal, the frailty terms will have a log-normal distribution. We refer to such as a model as a Cox log-normal frailty model. When the distribution of the frailty terms follows a Gamma distribution, we refer to the resultant model as a Cox Gamma frailty model. These are the most common Cox frailty models, and both can be implemented in popular statistical software such as R or SAS (while Stata only permits estimation of the Cox Gamma frailty model).

Lanke demonstrated that when the frailty terms followed a log-normal distribution, then the MHR is evaluated as

$$\text{MHR} = \exp\left(\sqrt{2\sigma^2}\Phi^{-1}(0.75)\right) \tag{4}$$

where σ^2 is the variance of the random effects (i.e., $\alpha_j \sim \text{N}(0,\sigma^2)$) and Φ^{-1} denotes the inverse of the standard normal cumulative distribution function [29]. Thus, in the case of normally distributed random effects, the MHR is evaluated in a method that is identical to that which is used for evaluating the MOR for the logistic-normal hierarchical regression model. The above result arises from the fact that when the random effects follow a normal distribution, then the distribution of $|\alpha_i - \alpha_j|$ follows a half-normal distribution with variance equal to $2\sigma^2$. The median of this half-normal distribution is given by $\sqrt{2\sigma^2}\Phi^{-1}(0.75)$.

When the frailty terms follow a Gamma distribution with variance σ^2 (i.e., $\exp(\alpha_i) \sim \Gamma(\sigma^{-2}, \sigma^2)$, under the convention that $\text{E}[\exp(\alpha_j)] = 1$), then the MHR is evaluated as the upper quantile of an $F(2\sigma^{-2}, 2\sigma^{-2})$ distribution [30].

R code for computing the MHR and MOR is provided in the appendix.

3. Case study

We provide a case study to illustrate the utility of the MHR for evaluating the hospital GCE on the hazard of death subsequent to hospitalization for AMI. Typically, studies assessing hospital performance in survival after hospitalization for AMI assume that over and above patient characteristics, the hospital context exerts a general, shared effect on all patients at the same hospital. This concept is analogous to the frailty effect in multilevel survival regression. However, as explained in Section 1, this frailty effect cannot be properly assessed by interpreting between-hospital variation only, but rather by a measure that, like the MHR, explicitly operationalizes hospital general contextual effects.

3.1. Data

We used data from the Ontario Myocardial Infarction Database, which contains data on patients hospitalized with an AMI at Ontario hospitals between 1992 and 2013 [33]. For this case study, we used hospital separations (occurring because of patient discharge or of in-hospital death) that occurred in the 12-month period between April 1, 2006 and March 31, 2007. The data have a multilevel structure, with

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

patients nested within hospitals. The study sample consisted of 17 243 patients treated at 157 hospitals. Due to the study inclusion and exclusion criteria, no patient had more than one hospital discharge during the one year time frame of the study.

Eleven variables, consisting of the variables in the Ontario AMI Mortality Prediction model (age, sex, congestive heart failure, cardiogenic shock, arrhythmia, pulmonary edema, diabetes mellitus with complications, stroke, acute renal disease, chronic renal disease, and malignancy), were measured on each patient [34]. The one continuous explanatory variables (age) was centered around the sample average.

The outcome for the case study was the time from hospital admission to the occurrence of death due to any cause. Patients were followed for up to one year from the time of hospital admission, and were censored after 365 days of follow-up if they were still alive. Death within one year of hospital admission occurred for 3758 (21.8%) patients in the sample.

3.2. Statistical analysis

We fit two different Cox frailty models. First, we fit the null model, which included only hospital-specific random effects:

$$\log\left(h_{ij}(t)\right) = \log(h_0(t)) + \alpha_j. \tag{5}$$

Second, we fit a frailty model which comprised the 11 variables in the Ontario AMI mortality prediction model and the hospital-specific random effects:

$$\log\left(h_{ij}(t)\right) = \log(h_0(t)) + \beta_{\text{patient}}\mathbf{X}_{ij} + \alpha_j \tag{6}$$

where the vector $\mathbf{X}_{ij}$ denotes the vector of patient-level characteristics, and β_{patient} denotes the vector of associated regression coefficients.

We first fit these two models assuming that the distribution of the random effects was normal: $\alpha_j \sim N(0, \sigma^2)$ (equivalent to assuming that the frailties follow a log-normal distribution). We then fit these two models assuming that the distribution of the frailty terms followed a Gamma distribution: $\exp(\alpha_j) \sim \Gamma(\theta^{-1}, \theta)$, so that $E[\exp(\alpha_j)] = 1$ and $\mathrm{Var}[\exp(\alpha_j)] = \theta$.

For each of the sets of two models, we computed the MHR using the methods described in Section 2.

Statistical analyses were conducted using PROC PHREG in SAS (SAS/STAT version 13.1) (Cary, NC) unless otherwise noted (in which case, the stcox function in Stata (version 13.1) (College Station, TX) was used).

3.3. Results

Estimated hazard ratios and associated 95% confidence intervals obtained from the two frailty models are reported in Table I. The estimated hazard ratios and associated 95% confidence intervals were essentially identical between the two frailty models.

Table I. Hazard ratios and 95% confidence intervals for the two frailty models

Variable	Cox log-normal model	Cox Gamma model
Age (per 10-year increase)	1.85 (1.79,1.91)	1.85 (1.79,1.91)
Female (vs. male)	0.96 (0.90,1.03)	0.96 (0.90,1.03)
Congestive heart failure	1.60 (1.49,1.72)	1.60 (1.49,1.71)
Cerebrovascular disease	1.57 (1.35,1.82)	1.57 (1.35,1.82)
Pulmonary edema	1.70 (1.26,2.31)	1.70 (1.26,2.31)
Diabetes with complications	1.27 (1.18,1.37)	1.27 (1.18,1.37)
Malignancy	2.88 (2.54,3.27)	2.88 (2.54,3.26)
Chronic renal failure	1.38 (1.25,1.52)	1.38 (1.25,1.52)
Acute renal failure	1.51 (1.35,1.69)	1.51 (1.35,1.69)
Cardiogenic shock	6.37 (5.54,7.32)	6.36 (5.53,7.31)
Cardiac arrhythmia	1.21 (1.12,1.31)	1.21 (1.12,1.31)

Each cell contains the estimated hazard ratio and the associated 95% confidence interval.

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

For the two log-normal frailty models, the estimated variance of the distribution of the random effects (i.e., the variance of the underlying normal distribution) were 0.06218 (null model) and 0.02391 (model with patient characteristics). The MHR for these two models were 1.27 and 1.16, respectively. For the null model, the estimated standard error of this estimated variance was 0.01360. A modified Wald test which is equal to the estimated variance divided by an estimate of its standard error allows one to test whether the variance is significantly different than zero [35]. The modified Wald test statistic for the null model is 4.57, which can be compared to the critical value of 1.64 for a normal one-sided test. Thus, we would reject the null hypothesis of no between-hospital variation with a highly significant p-value ($P < 0.001$). For the adjusted model, the estimated standard error of the estimated variance was 0.007666, resulting in a modified Wald test statistic of 3.11. Thus, even after adjustment for patient characteristics, we would reject the null hypothesis of no between-hospital variation in the hazard of death ($P < 0.001$).

For the two gamma frailty models, the estimated variance of the distribution of the frailty terms (i.e., the variance of the Gamma distribution) were 0.05592 (null model) and 0.02105 (model with patient characteristics). Using a modified Wald test, we rejected the null hypothesis of no between-hospital variation in both the null model and the model that adjusted for patient characteristics ($P < 0.001$) (the components for computing the Wald test for the gamma frailty models were obtained from models fit using Stata, as SAS does not provide an estimate of the standard error of the estimated variance of the frailty distribution for the gamma frailty model). The MHR for these two models were 1.26 and 1.15, respectively. Thus, when using the gamma frailty model, prior to adjusting for patient characteristics, the median increase in the hazard of mortality when comparing a patient at a hospital with higher mortality to a patient at a hospital with lower mortality was 26%. After accounting for patient characteristics, the median increase in the hazard of mortality when comparing a patient at a hospital with higher mortality to a patient at a hospital with lower mortality was 15%. Comparable interpretations are drawn from the log-normal frailty model.

One can better understand the magnitude of between-hospital variation in mortality by comparing the MHR for the full model with the hazard ratios for the patient-level characteristics. The MHR for the gamma frailty model was 1.15 (the reciprocal of this MHR is 1/1.15 = 0.87). Only one patient-level characteristics (female sex) had a hazard ratio that lay between 0.87 and 1.15. Thus, the median effect of clustering on mortality was less than the effect of 10 of the 11 patient characteristics.

The estimated distributions of the frailty terms are described in Figure 1. The left panel contains the two distributions of the frailty terms under the null models, while the right panel contains the two distributions under the model that adjusted for the 11 patient characteristics. For a given model (null

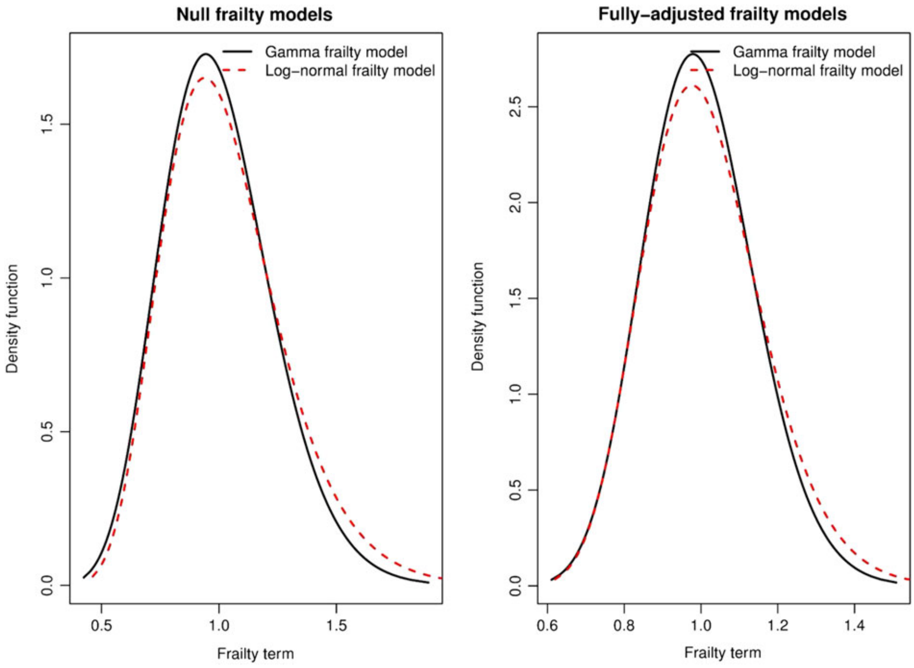

Figure 1. Distribution of frailty terms.

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

vs. adjusted), the choice of distribution had only a marginal effect on the shape of the distribution. The right tail was slightly heavier under the log-normal distribution than under the Gamma distribution. Thus, under the lognormal model, the proportion of hospitals that have a very elevated risk of death compared to the average hospital is higher than under the Gamma model. Through a comparison of the estimated hazard ratios (Table I), the MHRs, and the distribution of the frailty terms (Figure 1), one notes that the choice of frailty distribution had at most a minor impact on the conclusions that would be drawn from the data.

As a sensitivity analysis to examine the effect of duration of follow-up and number of events on the MHR, we repeated the above analyses allowing each patient to be followed for up to five years from the time of hospital admission. In this secondary analysis, 6904 (39.4%) patients died within five years of hospital admission. The MHRs for the models with normally distributed random effects were 1.23 (null model) and 1.15 (model with patient characteristics). The MHRs for the models with the gamma-distributed frailty terms were 1.22 (null model) and 1.14 (model with patient characteristics). These MHRs were qualitatively comparable to those obtained when subjects were followed for one year after hospital admission.

3.4. *Other measures of dependence*

For comparative purposes, we examined other measures of dependence for use with frailty models. These measures provide alternative methods to quantify the effect of clustering. These measures were derived in the context of bivariate survival data (i.e., when the clusters consist of two observed survival times) [36]. However, they can be used in the general setting with multiple survival times observed per cluster. In that case, these measures refer to the bivariate marginal distribution. Closed-form expressions exist for these measures under the Gamma frailty model, but not under the log-normal frailty model.

Kendall's τ denotes the correlation of subjects' outcomes within groups [31,36,37]. A closed-form expression exists for Kendall's τ under the Gamma frailty model, but not under the log-normal frailty model [31,32]. Under the Gamma frailty model, $\tau = \frac{\theta}{\theta+2}$, where θ denotes the variance of the frailty distribution. Under the Gamma frailty model estimated above, τ is equal to 0.027 for the null model and 0.010 for the model with patient characteristics. Thus, prior to adjustment for patient characteristics, 2.7% of the variation in survival times is due to variation between hospitals, while after adjustment for these 11 covariates, 1.0% of the variation in survival times is due to variation between hospitals.

Spearman's correlation coefficient for bivariate survival data can be defined as $\rho = 12\int_0^1\int_0^1 S(u,v)dudv - 3$ [36]. Under the Gamma frailty model, this can be evaluated as $\rho = \frac{12\theta^{-1}(\theta^{-1}+1)}{(1+2\theta^{-1})^2}\,{}_3F_2\left(\theta^{-1}+1,1,1,2\left(\theta^{-1}+1\right),2\left(\theta^{-1}+1\right),1\right)$, where ${}_3\mathrm{F}_2$ is a hypergeometric function defined by ${}_3F_2(\alpha,\beta,\gamma,\delta,\varepsilon,x) = \sum_{m=0}^{\infty}\frac{\Gamma(\alpha+m)\Gamma(\beta+m)\Gamma(\gamma+m)\Gamma(\delta)\Gamma(\varepsilon)x^m}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)\Gamma(\delta+m)\Gamma(\varepsilon+m)m!}$. Under the Gamma frailty model estimated above, Spearman's correlation coefficient is 1.014 for the null model and 1.005 for the model with patient characteristics.

Median concordance is the concordance of a single observation (T_1,T_2) in relation to a fixed point (Median(T_1),Median(T_2)). It is defined by $\kappa = E\mathrm{sign}\{(T_1 - \mathrm{median}(T_1))(T_2 - \mathrm{median}(T_2))\}$ [36]. Under the Gamma frailty model, it can be evaluated as $\kappa = 4\left(2^{1+\theta^{-1}} - 1\right)^{-\theta^{-1}} - 1$. Under the Gamma frailty model estimated above, median concordance was 0.026 for the null model and 0.010 for the model with patient characteristics.

Using each of these measures of dependence, one would conclude that the magnitude of the effect of clustering was weak to modest, both before and after adjustment for patient characteristics. However, unlike the MHR, each of these measures lacks the ability to compare the magnitude of the contextual effect to that of the effects of individual patient characteristics. Furthermore, closed-form expressions for these dependence measures exist for the Gamma frailty model, but not the log-normal frailty model. In contrast to this, the MHR can be evaluated for both of these families of frailty models.

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

4. Discussion

The objective of the current paper was to introduce researchers in epidemiology and medical research to the concept of the MHR for use with multilevel analysis of clustered survival data. The MHR allows one to quantify the magnitude of the general contextual effects (i.e., the ‘frailty’) on the hazard of the occurrence of the outcome on the hazard ratio scale. Furthermore, it also permits a comparison of the magnitude of this general contextual effect with that of model covariates.

After an empirical application of the MHR in epidemiology in 2007 [28] and its subsequent formal mathematic derivation in 2010 in the history literature [29], only two papers have applied the MHR in the peer-reviewed literature (Source: Science Citation Index, April 28, 2016). The first used the MHR to quantify the family effect on infant and child mortality in Sweden during the period 1766–1895 [30] and the second examined the effect of the mother on fertility in a nineteenth century alpine village [38]. Therefore, despite the frequency with which multilevel survival or time-to-event outcomes occur in the medical, epidemiological, and health services research literature, investigators in these fields appear to be unaware of the MHR. The purpose of our brief article was to introduce researchers in these fields to the existence and utility of the MHR.

In our case study, we used the MHR to quantify the magnitude of the effect of clustering within hospitals, that is the ‘frailty’ or general contextual effect, on the hazard of mortality subsequent to hospitalization for an AMI. We found that the MHR for the gamma frailty model that contained 11 patient characteristics was 1.15, indicating that, for 50% of possible pair-wise comparisons, the hazard of death for a reference patient was less than 15% greater when comparing a hospital with higher mortality to a hospital with lower mortality. Furthermore, the MHR, which measures the median effect of ‘frailty’ on the hazard ratio scale, was smaller in magnitude than the hazard ratios for 10 of the 11 patient characteristics. Reporting the MHR complements the reporting of the variance of the frailty distribution (and possibly a plot of the density function of the frailty distribution). The MHR provides a characterization of the magnitude of the effect of clustering that would not have been possible had we simply reported the variances of the frailty distributions. From a description of the frailty distribution on its own, it is difficult to summarize the effect of context on the hazard of outcomes. In contrast to this, the MHR provides a summary measure of the contextual effect on the hazard of outcome.

In conclusion, the MHR allows one to determine the median relative change in the hazard of the outcome between a subject in a cluster at a higher risk for the outcome and an identical subject in a cluster at a lower risk for the outcome. Such a measure permits for an intuitive description of the magnitude of the impact of the hospital ‘frailty’ or general contextual effects when analyzing clustered survival data.

Appendix: R code for computing the MOR and MHR

```
# We assume that the appropriate multilevel logistic regression model or multilevel
# survival regression model has been fit in the user’s statistical software package of
# choice (e.g. R, SAS, or Stata).

# MOR: for use with the multilevel logistic regression model and
# MHR: for use with the Cox log-normal frailty model.
# Let var.re denote the estimate variance of the random effects.

sd.re < - sqrt(var.re)
# The standard deviation of the distribution of the random effects. The random effects
# follow a normal distribution.

MOR < - exp(sd.re * qnorm(0.75) * sqrt(2))
MHR < - exp(sd.re * qnorm(0.75) * sqrt(2))

# MHR for the Cox Gamma frailty model.
# Let var.frailty denote the estimated variance of the frailty terms (which follow
# a Gamma distribution).

df.F < - 2/var.frailty
MHR < - qf(0.75,df.F,df.F)
# The MHR for a Cox Gamma frailty model
```

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Acknowledgements

This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported in-part by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508) and in-part by the Swedish Research Council (VR #2013-2484; PI: Juan Merlo). Dr. Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation. This study was approved by the institutional review board at Sunnybrook Health Sciences Centre, Toronto, Canada. These datasets were linked using unique encoded identifiers and analyzed at the Institute for Clinical Evaluative Sciences (ICES). Parts of this material are based on data and information compiled and provided by CIHI. However, the analyses, conclusions, opinions, and statements expressed herein are those of the author, and not necessarily those of CIHI.

References

1. Snijders T, Bosker R. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage Publications: London, 1999.
2. Goldstein H. Multilevel Statistical Models. John Wiley & Sons Ltd.: West Sussex, 2011.
3. Singer JD, Willett JB. Applied Longitudinal Data Analysis. Oxford University Press: New York, NY, 2003.
4. Raudenbush SW, Bryk AS. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage Publications: Thousand Oaks, 2002.
5. Merlo J. Multilevel analytic approaches in social epidemiology: measures of health variation compared with traditional measures of association. *Journal of Epidemiology & Community Health* 2003; **57**(8):550–552.
6. Goldstein H, Browne W, Rasbash J. Partitioning variation in generalised linear multilevel models. *Understanding Statistics* 2002; **1**:223–232.
7. Merlo J, Ohlsson H, Lynch KF, Chaix B, Subramanian SV. Individual and collective bodies: using measures of variance and association in contextual epidemiology. *Journal of Epidemiology and Community Heath* 2009; **63**(12):1043–1048.
8. Ghith N, Wagner P, Frolich A, Merlo J. Short term survival after admission for heart failure in Sweden: applying multilevel analyses of discriminatory accuracy to evaluate institutional performance. *PLoS One* 2016; **11**(2e0148187). doi:10.1371/journal.pone.0148187.
9. Merlo J, Wagner P, Ghith N, Leckie G. An original stepwise multilevel logistic regression analysis of discriminatory accuracy: the case of neighbourhoods and health. *PLoS One* 2016; **11**(4e0153778). doi:10.1371/journal.pone.0153778.
10. Merlo J, Viciana-Fernandez FJ, Ramiro-Farinas D. Bringing the individual back to small-area variation studies: a multilevel analysis of all-cause mortality in Andalusia. *Spain. Social Science & Medicine* 2012; **75**(8):1477–1487.
11. Ohlsson H, Librero J, Sundquist J, Sundquist K, Merlo J. Performance evaluations and league tables: do they capture variation between organizational units? An analysis of 5 Swedish pharmacological performance indicators. *Medical Care* 2011; **49**(3):327–331.
12. Ohlsson H, Merlo J. Understanding the effects of a decentralized budget on physicians compliance with guidelines for statin prescription; a multilevel methodological approach. *BMC Health Services Reseach* 2007; **7**(1):68.
13. Hjerpe P, Ohlsson H, Lindblad U, Bostrom KB, Merlo J. Understanding adherence to therapeutic guidelines: a multilevel analysis of statin prescription in the Skaraborg Primary Care Database. *European Journal of Clinical Pharmacology* 2011; **67**(4):415–423.
14. Tu JV, Austin PC, Naylor CD, Iron K, Zhang H. Acute myocardial infarction outcomes in Ontario. In Cardiovascular Health and Services in Ontario: An ICES Atlas, Naylor CD, Slaughter PM (eds). Institute for Clinical Evaluative Sciences: Toronto, 1999; 83–110.
15. Luft, H. S., Romano, P. S., Remy, L. L., and Rainwater, J. *Annual Report of the California Hospital Outcomes Project*. Sacramento, CA, California Office of Statewide Health Planning and Development, 1993.
16. Pennsylvania Health Care Cost Containment Council. *Focus on heart attack in Pennsylvania: research methods and results*. Harrisburg, PA, Pennsylvania Health Care Cost Containment Council, 1996.
17. Naylor CD, Rothwell DM, Tu JV, Austin PC, the Cardiac Care Network Steering Committee. Outcomes of coronary artery bypass surgery in Ontario. In Cardiovascular Health and Services in Ontario: An ICES Atlas, Naylor CD, Slaughter PM (eds). Institute for Clinical Evaluative Sciences: Toronto, 1999; 189–198.
18. Coronary artery bypass graft surgery in New York State 1989–1991. Albany, NY, New York State Department of Health, 1992.
19. Jacobs, F. M. *Cardiac surgery in New Jersey in 2002: a consumer report*. Trenton, NJ, Department of Health and Senior Services, 2005.
20. Pennsylvania Health Care Cost Containment Council. *Consumer guide to coronary artery bypass graft surgery*, volume 4. Harrisburg, PA, Pennsylvania Health Care Cost Containment Council, 1995.
21. Massachusetts Data Analysis Center. *Adult coronary artery bypass graft surgery in the commonwealth of Massachusetts: fiscal year 2010 report*. Boston, MA, Department of Health Care Policy, Harvard Medical School, 2012.
22. Diehr P, Cain K, Connell F, Volinn E. What is too much variation? The null hypothesis in small-area analysis. *Health Services Research* 1990; **24**(6):741–771.
23. Merlo J, Chaix B, Yang M, Lynch J, Rastam L. A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. *Journal of Epidemiology & Community Health* 2005; **59**(6):443–449.

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

24. Larsen K, Petersen JH, Budtz-Jorgensen E, Endahl L. Interpreting parameters in the logistic regression model with random effects. *Biometrics* 2000; **56**(3):909–914.
25. Larsen K, Merlo J. Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *Am.J.Epidemiol.* 2005; **161**(1):81–88.
26. Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Rastam L, Larsen K. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology and Community Heath* 2006; **60**(4):290–297.
27. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology* 2010; **63**(2):142–153.
28. Chaix B, Rosvall M, Merlo J. Assessment of the magnitude of geographical variations and socioeconomic contextual effects on ischaemic heart disease mortality: a multilevel survival analysis of a large Swedish cohort. *Journal of Epidemiology and Community Heath* 2007; **61**(4):349–355.
29. Lanke J. How to describe the impact of the family-specific frailty (appendix). *Historical Methods* 2010; **43**(1):26–27.
30. Bengtsson T, Dribe M. Quantifying the family frailty effect in infant and child mortality by using median hazard ratio (MHR). *Historical Methods* 2010; **43**(1):15–27.
31. Duchateau L, Janssen P. The Frailty Model. Springer: New York, NY, 2008.
32. Wienke A. Frailty Models in Survival Analysis. Chapman & Hall/CRC: Boca Raton, FL, 2011.
33. Tu JV, Austin P, Naylor CD. Temporal changes in the outcomes of acute myocardial infarction in Ontario, 1992–96. *Canadian Medical Association Journal* 1999; **161**(10):1257–1261.
34. Tu JV, Austin PC, Walld R, Roos L, Agras J, McDonald KM. Development and validation of the Ontario acute myocardial infarction mortality prediction rules. *Journal of the American College of Cardiology* 2001; **37**(4):992–997.
35. Rondeau V, Mazroui Y, Gonzalez JR. frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software* 2012; **47**(4). doi:10.18637/jss.v047.i04.
36. Hougaard P. Analysis of Multivariate Survival Data. Springer-Verlag: New York, 2000.
37. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Springer-Verlag: New York, 2000.
38. Quaranta L. Agency of change: fertility and seasonal migration in a nineteenth century alpine community. *European Journal of Population* 2011; **27**(4):457–485.

© 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.