

The measurement model

The reviewers rated preventability on a 0-100 scale representing the probability of survival had care been optimal. Their ratings were highly skewed toward higher ratings of preventability, but normalized by a log odds transformation suggesting that the reviewers estimated odds of survival on a multiplicative scale. We constructed a hierarchical model for the log odds of survival that can be represented mathematically as follows:

$$Y_{ij} = \beta_0 + u_i + e_{ij}$$

$$\text{with } u_i \sim \text{iid } N(0, \tau_{00}) \text{ and } e_{ij} \sim \text{iid } N(0, \sigma^2)$$

where:

Y_{ij} = the logodds of estimated survival with optimal care of the i^{th} patient by the j^{th} physician reviewer

β_0 = grand mean of Y (the log-odds of survival)

u_i = patient true log odds survival as deviations around the grand mean

e_{ij} = variation across the reviews within patient

τ_{00} and σ^2 are the variation in between-patient and between-review differences respectively where the

differences are independent and identically distributed (iid)

In the hierarchical model three parameters were estimated, the constant β_0 , τ_{00} , and σ^2 . A test for reviewer effect in a cross-classified hierarchical model did not show a significant reviewer effect. (See also Hofer et al for further details of these regression models examining the reliability of physician review.)²

Estimating the effect of unreliability and rating skew

It is well known that simply exponentiating log transformed model estimators ($E(\ln(Y))$) will lead to a biased posterior estimate of Y [$E(Y)$]. Therefore the posterior or shrunken means for each patient or u_i 's were calculated¹⁻³ and in a Monte-Carlo simulation 100 Y_{ij} 's per patient were generated by drawing from the estimated distributions of β_0 and e_{ij} keeping the u_i 's fixed by patient. The Y_{ij} 's were then back transformed to the 0-100 probability scale by the inverse of the log-odds transformation. (This is an alternative method to the "smear" estimate⁴ and this type of post-model-estimation simulation technique is described accessibly in King et al for more complex models.)¹ This allows us to examine the effect of using 100 reviews per patient, based on the measurement characteristics of implicit physician review, to estimate preventability and to take either the mean or the median of a simulated 100 reviewers.

1. King G, Tomz M, Wittenberg J. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*. 2000 Apr; 44(2):341-355

2. Hofer, T. P.; Bernstein, S. J.; DeMonner, S., and Hayward, R. A. Discussion between reviewers does not improve reliability of peer review of hospital quality. *Med Care*. 2000 Feb; 38(2):152-61.

3. Goldstein, Harvey. *Multilevel Statistical Models*. 2nd ed. New York: Halstead Press; 1995.

4. Duan N, Manning WG. A comparison of alternative models for the demand for medical care. *J Econ Bus Stat* 1983;1:115-126.s