DOI: 10.1111/j.1475-6773.2006.00661.x

Overestimating Outcome Rates: Statistical Estimation When Reliability Is Suboptimal

Rodney A. Hayward, Michele Heisler, John Adams, R. Adams Dudley, and Timothy P. Hofer

Objective. To demonstrate how failure to account for measurement error in an outcome (dependent) variable can lead to significant estimation errors and to illustrate ways to recognize and avoid these errors.

Data Sources. Medical literature and simulation models.

Study Design/Data Collection. Systematic review of the published and unpublished epidemiological literature on the rate of preventable hospital deaths and statistical simulation of potential estimation errors based on data from these studies.

Principal Findings. Most estimates of the rate of preventable deaths in U.S. hospitals rely upon classifying cases using one to three physician reviewers (implicit review). Because this method has low to moderate reliability, estimates based on statistical methods that do not account for error in the measurement of a "preventable death" can result in significant overestimation. For example, relying on a majority rule rating with three reviewers per case (reliability ~ 0.45 for the average of three reviewers) can result in a 50–100 percent overestimation compared with an estimate based upon a reliably measured outcome (e.g., by using 50 reviewers per case). However, there are statistical methods that account for measurement error that can produce much more accurate estimates of outcome rates without requiring a large number of measurements per case.

Conclusion. The statistical principles discussed in this case study are critically important whenever one seeks to estimate the proportion of cases belonging to specific categories (such as estimating how many patients have inadequate blood pressure control or identifying high-cost or low-quality physicians). When the true outcome rate is low (<20 percent), using an outcome measure that has low-to-moderate reliability will generally result in substantially overestimating the proportion of the population having the outcome *unless* statistical methods that adjust for measurement error are used.

Key Words. Reliability, statistical estimation, measurement error, medical errors, preventable deaths, adverse events

Measurement error is an inescapable part of scientific inquiry. The conventional wisdom is that random measurement error in an outcome (i.e., the dependent variable) does not affect your point estimates but only the standard errors and thus, while a nuisance, will only bias your results toward the null. (Carmines and Zeller 1979; Information Bias 2005) While true when estimating the overall population mean, an increasing body of literature illustrates substantial biases can occur due to overestimating the amount of "true" variation across groups of observations like physicians, hospitals or geographic areas (see Glossary; Diehr and Grembowski 1990; Diehr et al. 1990; Gatsonis et al. 1993, 1995; Hayward et al. 1994; Hofer and Hayward 1996; Hofer et al. 1999; Oppenheimer and Kher 1999; Krein et al. 2002). There has been considerably less discussion, however, of how measurement error can also result in inaccurate prevalence and incidence estimates when classifying cases into categories (especially dichotomies), such as when we set a specific test value (e.g., low-density lipoprotein [LDL] cholesterol ≥ 130 mg/dl) as the treatment threshold for a medical intervention (Hofer and Weissfeld 1994), designate a threshold for labeling a provider as an outlier (e.g., physician "report cards") (Hofer et al. 1999), or classify adverse events (AEs) as either "preventable" or "not preventable"(Hayward and Hofer 2001). Under such circumstances, even moderate measurement error in the dependent variable can result in substantial inaccuracies in estimating outcome rates.

In this paper, we examine this phenomenon through a case study of the widely quoted statistics that up to 100,000 Americans die each year in U.S. hospitals due to medical errors (Institute of Medicine 1999). While these estimates have been controversial, most criticisms of these numbers have focused on the method used to measure preventability: physician implicit review (trained physicians review medical records and estimate the likelihood that the death was due to medical error). Many criticisms of these implicit reviews have focused on whether the medical record provides adequate access to the information necessary to make comprehensive judgments about

---

Address correspondence to Rodney A. Hayward, M.D., Department of Veterans Affairs, VA Center for Practice Management & Outcomes Research, P.O. Box 130170, Ann Arbor, MI 48113-0170. Dr. Hayward, Michele Heisler, MD, MPA, and Timothy P. Hofer, MD, MSc, are with the Department of Veterans Affairs, VA Center for Practice Management & Outcomes Research, VA Ann Arbor Healthcare System, Ann Arbor, MI and the Departments of Internal Medicine & Health Management & Policy, University of Michigan Schools of Medicine & Public Health, Ann Arbor, MI. R. Adams Dudley, M.D., M.B.A., is with the Department of Internal Medicine, Pulmonary & Critical Care Medicine, University of California, San Francisco, CA. John Adams, Ph.D., is with The Rand Health Program, Santa Monica, CA.

medical errors, a shortcoming that might produce estimates that are too low as well as too high (Brennan 2000; Hofer, Kerr, and Hayward 2000; Leape 2000; McDonald et al. 2000; Sox and Woloshin 2000; Hayward and Hofer 2001; Hofer and Hayward 2002; Hofer, Asch, and Hayward, 2004). We wish to set aside the debate on the merits of physician implicit review in order to focus on an overlooked statistical issue. We use this case example to: (1) demonstrate how and why ignoring measurement error can result in large bias in estimating the prevalence of an outcome; and (2) outline some ways to recognize and avoid such bias in future work.

METHODS

Review of the Epidemiological Literature on Preventable Major AEs

We conducted a comprehensive evaluation of the published and unpublished epidemiological research evaluating the frequency of preventable *major* AEs (injuries resulting in death or substantial disability). Our inclusion criteria were: (1) the study assessed the proportion of *major* AEs that were preventable by better medical care based on information from direct observation, detailed investigation (such as interviewing people involved), or the medical record; (2) the sampling method and study population were adequate to determine that the estimates were representative of an identifiable patient or community population; and (3) the estimates of the reliability of preventability measures were obtainable. We reviewed all articles cited in the 1999 *To Err Is Human* report (Institute of Medicine 1999) and updated this review by conducting a search restricted to the years 1998–2003 using PubMed (www.pubmed.gov) and the search terms *medical errors, medication errors, preventable deaths*, and *preventable adverse events*. We also contacted over 30 experts from five different countries to solicit their suggestions of any additional epidemiological studies addressing this issue (see Appendix A).

We ultimately limited our study to estimates of preventable deaths as no study meeting our inclusion criteria reported the needed information on patients with nonfatal major AEs. Because only one of the four identified studies (Hayward and Hofer 2001) used a statistical method that accounted for measurement error (Table 1), we sought to obtain and reanalyze the original data for the other three studies. However, we were informed that the original data of the HMPS and the RAND Mortality Study are no longer available (Dubois and Brook 1988; Brennan et al. 1991; Leape, Brennan, and Laird 1991), and that almost all assessments of deaths in the UTCOS study had only a single

Table 1: Literature on Physician Assessments of the Preventability of Hospital Deaths

Study Patient Population (Reference)	*Interrater Reliability (κ or ICC of a Single Review) for Preventability or Negligence*	*Preventability Ratings*	*Sources Used for Review*	*Distribution of Reviewer Ratings*
Rand mortality rate study Dubois and Brook (1988), Dubois and Brook (1987) 182 inpatient deaths for patients admitted for stroke, myocardial infarction, or pneumonia	Average κ 0.2–0.3	546 reviews (3 reviews for each of the 182 deaths) 33 percent of cases were rated as "possibly preventable" based on a single review	Medical records and dictated hospital summaries	Results were dichotomized and original data are no longer in the possession of the authors (personal communication, March 1, 2004)
Harvard medical practice study Brennan et al. (1991), Leape *et al.* (1991) ~175 hospital deaths reviewed, with 154 being rated as adverse event deaths	κ ~ 0.14* (Estimated based upon a reported κ in the study of 0.24 for two reviews)	~9 percent of deaths were rated as due to negligence (rating of >"3" by 2 reviewers on the 6-point "due to negligence" scale)	Medical records	Negligence determined based upon a score of >"3" on a 6-point scale of self-reported confidence that medical management caused the AE's: 1, little or no evidence; 2, slight evidence; 3, not quite likely; 4, more likely than not; 5, strong evidence; 6, virtually certain Results were dichotomized and original data are no longer retrievable (personal communication with the authors [December 2003])
Utah/Colorado studyThomas et al. (2000, 2002)	κ = 0.19–0.23* 3 reviewers per case for a subgroup of 500 cases.	~6 percent of deaths were rated as due to negligence (rating of >"3" by a review	Medical records	Same 6-point scale as the HMPS. Required a confidence score of >3 for determination

continued

Table 1: *Continued*

Study Patient Population (Reference)	*Interrater Reliability (κ or ICC of a Single Review) for Preventability or Negligence*	*Preventability Ratings*	*Sources Used for Review*	*Distribution of Reviewer Ratings*
~ 55 hospital deaths reviewed with 38 being rated as adverse event deaths		on the 6-point "due to negligence" scale)		of negligence. Almost all deaths had only a single review. Only 2 percent of ratings were "6"*
VA preventable death study Hayward and Hofer (2001) 179 hospital deaths. After excluding 68 deaths that were receiving comfort care only, 111 hospital deaths underwent full review	ICC = 0.22 (ICC was 0.34 for 2 reviewers)	23 percent of active care deaths were rated as at least "possibly" preventability & 6 percent were rated as at least "probably" preventable (based upon a single review)	Medical records	Probability that the death was preventable by optimal care was rated in two ways: On a 5-point scale (1 = definitely; 2 = probably; 3 = uncertain/possibly; 4 = probably not; 5 = definitely not) And on a 0–100 percent scale

ICC, Intraclass correlation coefficient.
* Intra-class correlation not cited in the article (see Glossary).

reviewer (Thomas et al. 2000), so direct reanalyses accounting for measurement error were not possible for these studies. Therefore, we obtained parameter estimates from summary data in the published literature. We then developed mathematical models to assess how much bias there would be in estimates of the prevalence of preventable deaths if measurement error is ignored.

An Analytic Approach to Quantifying the Impact of Measurement Error on Outcome Rate Estimates

We formulated the problem of identifying preventable deaths using a classical test theory framework (Fleiss 1986; Oppenheimer and Kher 1999), dichotomizing a continuous assessment of the preventability of death into two classes, preventable versus not preventable. We stipulate that each case has an underlying "true" rating T, and an observed rating X that is measured with an additive random error term. The "true" score T and the error terms are independent and normally distributed. Finally we assume that there is a threshold A for the "true" score T, above which a death is "preventable" and below which a death is "not preventable." Starting with these assumptions, it is possible to write an equation for the false negative rate and false positive rate (see Appendix A; Oppenheimer and Kher 1999).

Simulations of the Impact of Measurement Error on Outcome Rate Estimates

We used standard statistical simulation techniques to determine whether the analytic calculations of estimation bias outlined above are robust to different assumptions and situations not amenable to an analytic solution, such as highly skewed or bimodal distributions of the outcome measure and nonnormality or heterogeneity in the error terms (Concato and Feinstein 1997; Feiveson 2002). We began by generating populations with known amounts of between-case variation (i.e., the amount of variance between the cases' "true" ratings) and within-case variance (i.e., the amount of variance in case ratings due to measurement error [in this case limited to random sampling variation]). We report on two case examples, one in which the "true" distribution of preventability ratings in the population is normally distributed with a mean and median rating of 0.3 (SD = 0.1), and one in which the "true" distribution of ratings are heavily skewed—half of a normal distribution bounded at the lower end by zero (mean = 0.14, SD = 0.12). We then generated 1,000 cases with each case having a known "true" result (i.e., the preventability rating that would be achieved from the universe of potential qualified reviewers) and then

randomly generated 100 reviews per case. (The statistical code for the simulations is given in Appendix A available in the online version of this paper and can be obtained directly from RAH upon request.) Interrater reliability (IRR) was assessed by obtaining the intraclass correlation coefficient (ICC) from random-effects analysis of variance of the 100,000 reviews of the 1,000 simulated cases (100 reviews per case). Results given different numbers of reviewers per case were obtained by random bootstrap resampling of 3, 15, and 50 reviews per case (2,000 iterations for each; Concato and Feinstein 1997; Feiveson 2002).

RESULTS

Overestimation When Using a Diagnostic Test—A Simple Thought Experiment

Let's begin with a hypothetical example that makes an analogy to a well-known phenomenon when using a diagnostic test. Imagine that the following is true: (1) one in 200 deaths (0.5 percent) are truly preventable (based upon a hypothetical gold standard determination); and (2) two out of two reviewers rating a death as preventable (a nongold standard test) has a sensitivity and specificity of 90 percent. As is demonstrated in Figure 1, under these

Figure 1: Dramatic Overestimation of Preventable Deaths (PDs) Even if Sensitivity and Specificity Are Good—A Thought Experiment

Assumptions:

- The accuracy of 2 of 2 reviewers rating a death as preventable is
  Sensitivity = 90 % & Specificity = 90%
- The true rate of preventable deaths is 0.5% (1 in 200 people)

Therefore, out of every 10,000 deaths, on average:

50 deaths are truly preventable & 9,950 are not, resulting in:

45 True Positives (TPs) = (50 * 0.9), since TPs = # of PDs * sensitivity
995 False Positives (FPs) = (9,950 * (1 - 0.9)),
since FPs = # of non-PDs * (1 – specificity)

$$\text{Proportion of deaths that are "rated" as PD} = \frac{995 + 45}{10{,}000} = 10\%$$

$$\text{Proportion of deaths that are truly PDs} = \frac{50}{10{,}000} = 0.5\%$$

circumstances we would overestimate the proportion of deaths that are preventable by 20-fold if we classified deaths as "preventable" based upon two implicit reviews. This is because, by definition, a specificity of 90 percent results in 10 percent of people having a positive test result even when the prevalence of the disease in the study population is zero. This well-known epidemiological phenomenon is why it takes a test with near-perfect specificity to avoid substantially overestimating the prevalence of a rare disease.

Overestimation Due to Random Measurement Error—An Analytic Approach

Of course, we do not have a gold-standard for detecting preventable deaths so there is no way to know the sensitivity and specificity of implicit review. However, for now let us assume that the "true" mean implicit review rating is a gold standard for identifying preventable deaths, but that in practice the rating produced by a single implicit review has low reliability (its reliability was ≈ 0.2–0.3 in the two studies used to generate the 100,000 preventable death statistics). Table 2 shows how random measurement error alone can result in dramatic estimation errors. For example, if the "true" prevalence of preventable deaths is 1 percent, not adjusting for measurement error in an otherwise perfect test (if averaged across enough repeat measurements) would result in a 12-fold overestimation of preventable deaths. Just as in the case of diagnostic testing, the degree of overestimation increases when the true prevalence is very low. Although throughout most of this paper we discuss this phenomenon with respect to overestimating rare outcome events, Table 2 demonstrates how this effect is symmetric (resulting in under-estimation due to false negative findings when the true prevalence is very high).

Review of the Epidemiological Literature on Preventable Major AEs

How do the general phenomena described above relate to how estimates of 100,000 preventable deaths were actually made? We found four studies meeting our inclusion criteria that estimated the number of preventable deaths, all of which used implicit review. In these studies (see Table 1), one to three trained physician raters examined the medical record of hospital patients who had died, and the reviewers rated the probability that the death was "due to negligence" or "was preventable by optimal care." Based upon the rating of these one to three reviewers, they would classify cases as preventable versus not preventable (generally using majority rule when more than 1 reviewer reviewed the case).

Table 2: Effect of Random Measurement Error (Reliability = 0.25) on Outcome Estimation Based on Classical Test Theory*

"True" Proportion of Deaths Meeting the Threshold for Being Rated "Preventable"	*Frequency of "True Positives"*	*Frequency of "False Positives"*	*"Estimated" Proportion of Deaths Meeting the Threshold for Being Rated "Preventable"*
0.99	0.88	<0.01	0.88
0.95	0.78	0.02	0.79
0.70	0.49	0.11	0.60
0.50	0.33	0.17	0.50
0.30	0.19	0.21	0.40
0.05	0.02	0.18	0.21
0.01	<0.01	0.12	0.12
0.005	<0.01	0.10	0.10

*These calculations are based upon methods described in Oppenheimer and Kher (1999) and assume that both the underlying measure of preventability and the error term (random measurement error) are independent and normally distributed.

The main findings of all four studies are strikingly similar (Table 1). As reported previously by Thomas et al., the number of cases classified as "due to negligence" can vary considerably depending upon: (1) where you draw your cutoff between what is "preventable" versus what is "not preventable"; and (2) how many reviewers you use to classify a case (Thomas et al. 2002). Most of the apparent differences among studies in the percentage of events that are classified as "preventable" or "due to negligence" appear to be due to differences in cutoff points and the number of reviewers used—not due to true differences in the underlying distributions of individual reviewer ratings in these studies (Thomas et al. 2000). Despite differences among the four studies (in the wording of the questions, the measurement scales and the patient populations), if you classify cases as preventable based upon the ratings of one to three reviewers, all four studies in Table 1 would classify at least 6–10 percent of deaths as being preventable. The estimates that 40,000–100,000 preventable deaths occur in U.S. hospitals each year were obtained by multiplying the estimates obtained from the HMPS and UTCOS by the number of hospital deaths that occur each year in the United States (Sox and Woloshin 2000).

Overestimation of Preventable Deaths—A Simulation Approach

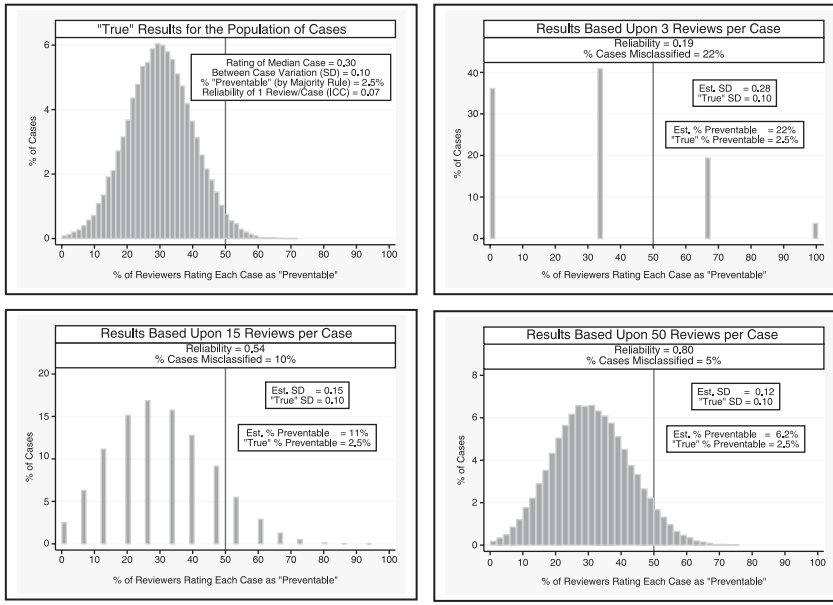
We have argued previously that dichotomizing cases as "preventable" versus "not preventable" is artificial as the concept of preventability more naturally

resides on a continuous 0–100 percent probability scale (e.g., "What is the probability that the patient would have lived if care had been optimal?"; Hayward and Hofer 2001). However, as only the VA Mortality Study asked reviewers to estimate a continuous measure of preventability, in this paper we restrict ourselves to exploring the categorical measurement approach used in the other three preventable death studies. In this instance, each case still has a 0–100 percent "true" rating, but the underlying "true" rating represents the percentage of a very large number of qualified raters that would rate the case as "preventable" versus "not preventable" (i.e., if you had thousands of reviewers rate this case, would 0, 10, 20 percent, etc. rate the case as "preventable?"). By a majority rules criterion, the mean rating would have to be above 50 percent for the case to be designated as preventable. When there are many reviewers, the mean rating of a case will fall near the "true" rating so misclassification is unlikely. However, with just a small number of raters, you should have much less confidence that the average rating of those few reviewers will necessarily be close to the "true" rating, resulting in both misclassification and overestimating the "true" between-case variance ($\sigma^2_{(\text{true score})}$; see Glossary).

This simple phenomenon is what causes complex problems for estimating outcome rates when using an outcome measure that has low or moderate reliability random measurement error is added to that of the "true" variance, thus increasing the total observed variance between your units of observation ($\sigma^2_{(\text{total observed})}$). Therefore, the variance that you see (the observed variance) is an overestimate of the "true" between-case variation ($\sigma^2_{(\text{true score})}$), which means that you are overestimating how far apart your groups/observations are from each other and how far they are from the population mean.

The examples in Figure 2 visually demonstrate this phenomenon. In each instance, we know the "true" proportion of variance due to within-case variance (in this instance, random measurement error that is solely due to sampling error) versus between-case variance (the differences between the cases' "true" ratings). For illustrative purposes, in Figure 2A we have arbitrarily set the "true" distribution of preventability ratings to be normally distributed with a mean and median rating of 0.3 (i.e., for the median case 30 percent of reviewers will rate the death as being "preventable") and a standard deviation of 0.1 (resulting in 95 percent of cases having between a 10–50 percent probability of being rated as "preventable" by the average of a very large sample of reviewers randomly selected from the universe of potential reviewers). In this example the IRR of a single review is quite poor (ICC = 0.07). The low reliability does not bias our estimate of the population

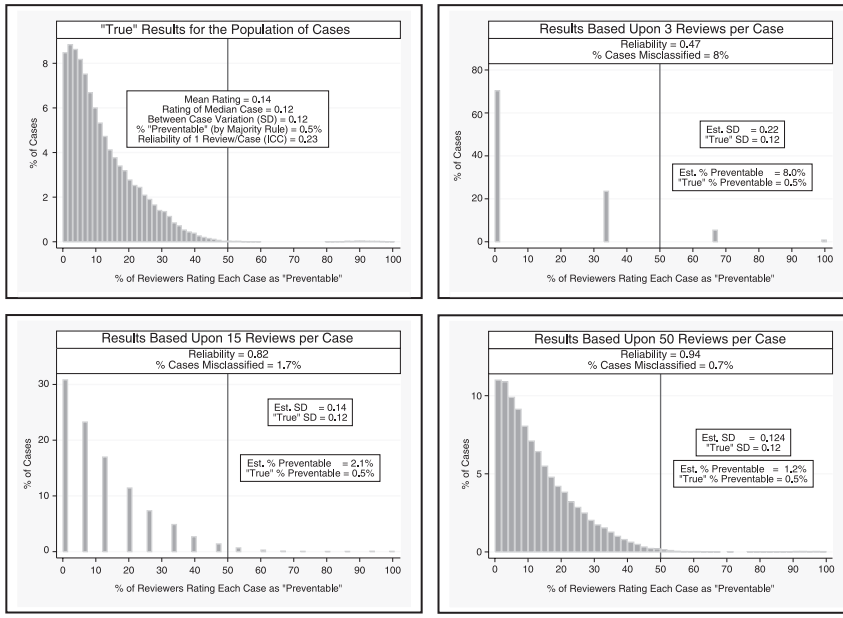
Figure 2A: Example A: How Low Reliability Can Result in Overestimation



mean rating (as the measurement error is random), but the random measurement error is added to the "true" variability between cases, thus resulting in an overestimate of between-case variation and the number of cases that fall above the preventability threshold. For example, the estimated between-case SD based on three reviews per case is 0.28 (almost three-times greater than the "true" between-case SD of 0.1) which in turn results in almost a 10-fold overestimation of the percentage of deaths above the "preventability" criterion (22 versus 2.5 percent, see Figure 2A). Even with 15 reviewers per case (reliability = 0.54) we overestimate the number of deaths above our preventability criterion by over fourfold (11 versus 2.5 percent). When categorizing cases based upon a continuous measure, even minimal measurement error can result in substantial misclassification when a substantial proportion of the population have "true" ratings that are close to the categorization threshold criterion. In our example in Figure 2A, modest measurement error results in dramatic overestimation because a fair proportion of the population have "true" preventability ratings between 30–50 percent, which are only slightly lower than our 50 percent threshold criterion.

In Figure 2B we show results that follow the general distribution found in our literature review of preventable deaths. In this example, we would

Figure 2B: Example B: Results Consistent with the Published Epidemiological Literature



The above examples demonstrate how low reliability results in overestimating the between-case variance, which in turn can dramatically overestimate the percentage of cases above the classification threshold (the threshold in this case is >50 percent of reviewers rating the death as being "preventable"). For instance, in example a, even when reliability is 0.54 (15 reviewers per case), the estimated standard deviation is 50 percent greater than the "true" standard deviation (0.15 versus 0.10, respectively) and the estimated percentage of deaths above the "preventable death" threshold criterion is over four times greater than the "true" percentage (11 versus 2.5 percent, respectively).

estimate that 14 percent of cases are "preventable" based upon one review per case (reliability [ICC] = 0.23), 8.0 percent are "preventable" based upon three reviews (reliability = 0.47), and 2.1 percent based upon 15 reviews (reliability = 0.82; the "true" value = 0.5 percent). Just as in Example A, the low reliability resulted in an overestimation of the "true" between-case variance, and this led to dramatic overestimation of the percentage of cases truly meeting the "preventability" criterion.

How Many Preventable Deaths Are There Really?

Although Example B shows an example *consistent with* the epidemiological literature on preventable deaths, there are other distributions consistent with this literature (particularly bimodal distributions) that would result in substantially less shrinkage (some distributions only show 50 percent shrinkage in estimates). In other words, if there are a small number of outlier cases with preventability ratings slightly above the 50 percent threshold criteria (instead of a more continuous single distribution of cases' preventability ratings) there could be much less shrinkage. It is not possible to further resolve this issue (i.e., whether the 100,000 preventable deaths estimate would have shrunk by 50 versus over 99.9 percent) in the absence of the original data (which is only available for the VA Mortality Study, which found a 75–85 percent shrinkage after reliability-adjusting a continuous [as opposed to a dichotomous] assessment of preventability; Hayward and Hofer 2001). However, resolving this specific issue may not be especially important for the other three studies as Example B also shows a more fundamental problem in past research on this topic—how using a majority rules criterion and a dichotomized outcome ("preventable" versus "not preventable") can be misleading regardless of the statistics used. After all, counting almost all cases as "not preventable" simply because few cases meet the majority rules criterion would obscure the fact that for many cases there is substantial disagreement about whether the deaths are "preventable" and we cannot determine who is correct.

Accordingly, we believe that a more appropriate summary of the preventable deaths literature is that implicit review finds very few clear-cut "preventable deaths" in which a majority of reviewers would rate the case as "preventable," *but* there are many deaths in which a substantial proportion of reviewers would rate the death as "preventable" (Hayward and Hofer 2001). Those who believe that preventable hospital deaths are common can therefore argue that many errors may not be evident from the medical record and that the physician reviewers may be reluctant to criticize fellow physicians (Leape 2000). Alternatively, those who believe that few hospital deaths are preventable can counter that there is no clear evidence suggesting that preventable deaths cannot be detected from the medical record (Brennan et al. 1990) and that the outlier opinions (those who rate the deaths as preventable) are simply second-guessing reasonable care using hindsight (McDonald et al. 2000). We thus recommend that the health policy and health services research communities acknowledge that there is not strong epidemiological

evidence to support either position and that we should keep an open mind while awaiting more rigorous evidence on this topic (Hayward and Hofer 2001).

Reliability-Adjusting Outcome Measures

We have emphasized that the fundamental phenomenon underlying these estimation errors is that random measurement error results in an overestimation of the "true" between-case variation. Consequently, the key take home point is that whenever reliability is suboptimal, you should adjust for measurement error in order to better estimate the "true" distribution of cases in your study population and that this adjustment should always be done before you assign cases to categories. A full discussion of the different statistical approaches for adjusting for measurement error is a complex topic that is well beyond the scope of this paper. The optimal choice of statistical approach may in part depend upon whether you are trying to improve: (1) a specific probability estimate; (2) your estimate of the overall population distribution; or (3) the rank order of individual cases (Shen and Louis 2000). The metric and hypothesized distribution of the outcome measure should also influence the choice of statistical method. The statistical literature already contains a detailed discussion of alternative statistical techniques (Clayton 1991; Holt, McDonald, and Skinner 1991; Schulzer, Anderson, and Drance 1991; Gatsonis et al. 1993, 1995; Caroll, Ruppert, and Stefanski 1995; Coory and Gibberd 1998; Hofer et al. 1999; Shen and Louis 2000; Hayward and Hofer 2001; Skrondal and Rabe-Hesketh 2004), and we believe that most health services researchers will be best served by consulting an experienced statistician regarding which analytic approach is best given their study's specific circumstances. However, we briefly discuss below the general principles underlying most commonly used statistical approaches.

Key points
1. Low to moderate reliability of an outcome measure results in an overestimation of "true" between-case/group differences
2. Under such circumstances, failure to account for measurement error will usually result in substantial overestimation of prevalence when the "true" prevalence of the outcome is low
3. Examining the distribution of reliability-adjusted measures of the outcome variable should always be done *before* making classification decisions or conducting further analyses (which will often require consultation with a statistician)

Of course, one could simply use a brute force method to improve the precision of your outcome estimates by taking an average of tens or even hundreds of measurements per observation (thereby reducing measurement error and directly improving your estimate of the "true" population variance). However, a reasonable estimate of the "true" population variance can usually be obtained without resorting to a large number of measures per case except in those rare instances when we need to make definitive judgments about individual cases (e.g., deciding about malpractice settlements; Cronbach 1990).

Most of the available statistical techniques for adjusting for measurement error in an outcome variable involve explicitly modeling the amount of variance due to measurement errorthus allowing "removal" of the measurement error from estimates of the "true" between-case variance (afterall, the "true" variance is defined as the observed variance after removal of all measurement error [see Glossary]; Clayton 1991; Holt, McDonald, and Skinner 1991; Schulzer, Anderson, and Drance 1991; Gatsonis et al. 1993, 1995; Coory and Gibberd 1998; Hofer et al. 1999; Shen and Louis 2000; Hayward and Hofer 2001; Skrondal and Rabe-Hesketh 2004). These reliability-adjusted results are therefore much better approximations of the "true" distribution of the outcome measure across the study population. However, in order to reliability-adjust estimates, you need to have a sufficient number of replicate measures (multiple measures by case/group) as that is the only way that you can estimate the amount of overall variance that is due to measurement error. Unfortunately, there is no simple rule of thumb that can be given for how many replicate measures are needed. Once again the optimal approach will depend upon the purpose of the specific study and the metric and hypothesized distribution of the outcome measure (Clayton 1991; Holt, McDonald, and Skinner 1991; Schulzer, Anderson, and Drance 1991; Caroll, Ruppert, and Stefanski 1995; Coory and Gibberd 1998; Shen and Louis 2000; Skrondal and Rabe-Hesketh 2004). Yet, while it is not possible to easily define the optimal number of replicate measures, even a moderate number of replicate measures will dramatically *improve* your estimate of "true" between-case/group variance, such as two to five replicate measures of 30–50 cases (Clayton 1991; Holt, McDonald, and Skinner 1991; Schulzer, Anderson, and Drance 1991; Gatsonis et al. 1993, 1995; Caroll, Ruppert, and Stefanski 1995; Coory and Gibberd 1998; Hofer et al. 1999; Shen and Louis 2000; Hayward and Hofer 2001; Skrondal and Rabe-Hesketh 2004). If problems with reliability in the outcome measure are anticipated, however, we strongly recommend that a statistician be consulted during the planning stages of the study regarding how best to measure reliability using replicate measures.

DISCUSSION

The importance of the principles discussed in this paper are not limited to medicine, but rather, are universal principles of measurement theory. For example, when NASA receives photos sent from their spacecrafts passing by distant planets, the original transmissions received on Earth are often fuzzy and difficult to interpret due to white noise from cosmic radiation (Lyon 2005). However, once the noise is modeled and removed, the resolution of the pictures can be excellent and highly accurate (the true signal hidden within the white noise). Measurement error must be dealt with explicitly whenever reliability is suboptimal if you want to obtain an accurate picture of what is really going on.

Still, the estimation errors discussed in this paper have been a recurrent problem in health services research. Over a decade ago, Diehr et al. (1990) demonstrated how not accounting for random variation resulted in overestimating the magnitude of small area practice variations, Hayward et al. (1994) demonstrated dramatic overestimation in designating high resource use physicians and Hofer and Hayward demonstrated how ignoring random measurement error led to substantial errors in identifying high-mortality-rate hospitals (Hofer and Hayward 1996; Hofer et al. 1999). Gatsonis and others also demonstrated in the mid-1990s how random-effects hierarchical regression methods could be used to adjust variance estimates for measurement error (Gatsonis et al. 1993, 1995). However, failure to eliminate the white noise of measurement error continues to result in both classification errors and an overestimate of the magnitude of difference between cases or groups in many situations, including evaluations of resource use variation, health plan and physician profiling, patient safety problems, disease prevalence/incidence, and levels of blood pressure or lipid control. For example, as many parts of the U.S. health care sector push for physician pay-for-performance, most performance measurement activities still do not reliability-adjust their performance profiles (Hayward et al. 1994; Hofer and Hayward 1996; Hofer et al. 1999; Krein et al. 2002; Hofer, Asch, and Hayward, 2004). The estimation errors discussed in this paper can be prevented or greatly reduced by remembering one important principle: if reliability is suboptimal, adjust outcome estimates to account for the level of reliability and examine the distribution of the reliability-adjusted outcome variable before making classification decisions or conducting further analysis. Most often, reliability-adjustment should be done in consultation with a statistician with experience with these methods. Our challenge now is to be vigilant in recognizing this

potential pitfall and to obtain a sufficient number of replicate measures to allow us to account for measurement error when our measurements have less than optimal reliability.

ACKNOWLEDGMENTS

The authors thank Pat Mault for assistance in preparing the manuscript, Judi Zemencuk for assistance with the literature review, and two anonymous reviewers for their comments on an earlier draft of this paper. This work was supported by the VA Health Services Research & Development Service Quality Enhancement Research Initiative (QUERI DIB 98-001). Dr. Heisler is a VA HSR&D Career Development awardee. Support was also provided by The Agency for Healthcare Research & Quality (P20-H511540-01) and The NIDDK of The National Institutes of Health (P60 DK-20572).

Glossary

Term	*Definition*
"True" rating or score	The "true" rating or score is a statistical term used to denote the result that would be found once all measurement error is removed (such as that obtained from an infinitely large sample of repeated measures). Therefore, this term only refers to perfect reliability, and should not be confused with a perfect "gold standard," which refers to perfect accuracy. For clarity, we use quotation marks to denote the use of the statistical term (e.g., "true" rating)
Gold standard	Refers to a test/measure that is believed to be so highly accurate that it can be used as the standard by which the accuracy of all other tests/measures can be judged
"True" between-case variation	The amount of variation in the study population when all groups/observations of interest are measured without error. When random measurement error is present (imperfect reliability), the observed between-case variance is an overestimate of the "true" variance (i.e., you overestimate how far the groups/observations are from each other and from the population mean)
Reliability	The proportion of total observed variance ($\sigma^2_{(\text{total observed})}$) in a set of measurements that is due to "true" differences between the groups/observations of interest:

continued

Table 2: *Continued*

Term	*Definition*
	$\text{Reliability} = \sigma^2_{(\text{“true” between-case scores})} / \sigma^2_{(\text{total observed scores})}$ Therefore, reliability is affected by both the precision of the measurement tool and the amount of “true” variance in the population being studied. A tool with a set amount of measurement error can have excellent reliability when used in a study population with substantial “true” between-case variance but have poor reliability when used in a study population with low “true” between-case variance
Interrater reliability (IRR)	The degree of consistency (reliability) between different observers’ ratings of a phenomenon in a population ($\sigma^2_{(\text{“true” between-case scores})} / \sigma^2_{(\text{total observed scores})}$). Usually expressed as the amount of agreement beyond that expected by chance alone, using the intraclass correlation coefficient (ICC) or the κ statistic
Intraclass correlation coefficient (ICC)	Currently, the ICC is generally considered the preferred statistic for estimating interrater reliability; Hofer et al. (2004), Cronbach (1990), Bravo and Potvin (1991) however, the κ statistic was often the only reliability statistic given in previous literature. Therefore in discussing results from previous studies, we sometimes are forced to report reliability results using the κ statistic, but the ICC is used in this paper whenever possible

REFERENCES

Bravo, G., and L. Potvin. 1991. “Estimating the Reliability of Continuous Measures with Cronbach’s Alpha or the Intraclass Correlation Coefficient: Toward the Integration of Two Traditions.” *Journal of Clinical Epidemiology* 44 (4–5): 381–90.

Brennan, T. A. 2000. “The Institute of Medicine Report on Medical Errors—Could It Do Harm?” *New England Journal of Medicine* 342: 1123–5.

Brennan, T. A., L. L. Leape, N. M. Laird, L. Hebert, A. R. Localio, A. G. Lawthers, J. P. Newhouse, P. C. Weiler, and H. H. Hiatt. 1991. “Incidence of Adverse Events

and Negligence in Hospitalized Patients." *New England Journal of Medicine* 324 (6): 370–6.

Brennan, T. A., A. R. Localio, L. L. Leape, N. M. Laird, L. Peterson, H. H. Hiatt, and B. A. Barnes. 1990. "Identification of Adverse Events Occurring during Hospitalization. A Cross-Sectional Study of Litigation, Quality Assurance, and Medical Records at Two Teaching Hospitals." *Annals of Internal Medicine* 112: 221–6.

Carmines, E. G., and R. A. Zeller. 1979. *Reliability and Validity Assessment.* Beverly Hills, CA: Sage Publications.

Caroll, R. J., D. Ruppert, and L. A. Stefanski. 1995. *Measurement Error in Nonlinear Models.* London: Chapman & Hall.

Clayton, D. G. 1991. *Models for the Analysis of Cohort and Case Control Studies with Inaccurately Measured Exposures. Statistical Models for Longitudinal Studies in Health.* New York: Oxford University Press.

Concato, J., and A. R. Feinstein. 1997. "Monte Carlo Methods in Clinical Research: Applications in Multivariable Analysis." *Journal of Investigative Medicine* 45: 394–400.

Coory, M., and R. Gibberd. 1998. "New Measures for Reporting the Magnitude of Small-Area Variation in Rates." *Statistical Medicine* 17: 2625–34.

Cronbach, L. J. 1990. *Essentials of Psychological Testing.* 5th Edition. New York: Harper and Row.

Diehr, P., K. Cain, F. Connell, and E. Volinn. 1990. "What Is Too Much Variation? The Null Hypothesis in Small-Area Analysis." *Health Services Research* 24: 741–71.

Diehr, P., and D. Grembowski. 1990. "A Small Area Simulation Approach to Determining Excess Variation in Dental Procedure Rates." *American Journal of Public Health* 80: 1343–8.

Dubois, R. W., and R. H. Brook. 1987. "Hospital Inpatient Mortality. Is It a Predictor of Quality?" *New England Journal of Medicine* 317 (26): 1674–80.

———. 1988. "Preventable Deaths: Who, How Often, and Why?" *Annals of Internal Medicine* 109: 582–9.

Feiveson, A. H. 2002. "Power by Simulation." *Stata Journal* 2: 107–24.

Fleiss, J. L. 1986. *The Design and Analysis of Clinical Experiments.* New York: Wiley.

Gatsonis, C. A., A. M. Epstein, J. P. Newhouse, S. L. Normand, and B. J. McNeil. 1995. "Variations in the Utilization of Coronary Angiography for Elderly Patients with an Acute Myocardial Infarction. An Analysis Using Hierarchical Logistic Regression." *Medical Care* 33: 625–42.

Gatsonis, C., S. L. Normand, C. Liu, and C. Morris. 1993. "Geographic Variation of Procedure Utilization. A Hierarchical Model Approach." *Medical Care* 31: 54–9.

Hayward, R. A., and T. P. Hofer. 2001. "Estimating Hospital Deaths Due to Medical Errors: Preventability Is in the Eye of the Reviewer." *Journal of the American Medical Association* 286: 415–20.

Hayward, R. A., W. G. Manning Jr., L.F McMahon, and A. M. Bernard. 1994. "Do Attending or Resident Physician Practice Styles Account for Variations in Hospital Resource Use?" *Medical Care* 32: 788–94.

Hofer, T. P., S. M. Asch, R. A. Hayward, L. V. Rubenstein, M. M. Hogan, J. Adams, and E. A. Kerr. 2004. "Profiling Quality of Care: Is There a Role for Peer Review?" *BMC Health Services Research* 4: 9.

Hofer, T. P., S. J. Bernstein, S. DeMonner, and R. A. Hayward. 2000. "Discussion between Reviewers Does Not Improve Reliability of Peer Review of Hospital Quality." *Medical Care* 38: 152–61.

Hofer, T. P., and R. A. Hayward. 1996. "Identifying Poor-Quality Hospitals. Can Hospital Mortality Rates Detect Quality Problems for Medical Diagnoses?" *Medical Care* 34: 737–53.

———. 2002. "Are Bad Outcomes from Questionable Clinical Decisions Preventable Medical Errors? A Case of Cascade Iatrogenesis." *Annals of Internal Medicine* 137: 327–33.

Hofer, T., R. Hayward, S. Greenfield, E. Wagner, S. H. Kaplan, and W. Manning. 1999. "The Unreliability of Individual Physician 'Report Cards' for Assessing the Costs and Quality of Care of A Chronic Disease." *Journal of the American Medical Association* 281 (22): 2098–105.

Hofer, T. P., E. A. Kerr, and R. A. Hayward. 2000. "What Is an Error?" *Effective Clinical Practice* 3: 261–9.

Hofer, T., and J. Weissfeld. 1994. "Designing A Simpler High Blood Cholesterol Case Detection Strategy: Are the Advantages of the NCEP Protocol Worth the Complexity?" *Medical Decision Making* 14: 357–68.

Holt, D., J. W. McDonald, and C. J. Skinner. 1991 The Effect of Measurement Error on Event History Analysis. *Measurement Errors in Surveys.* New York: Wiley.

Information Bias. Available at http://hsrd.durham.med.va.gov/eric/notebook/ERICIssue16.pdf. (accessed December 27, 2005)

Institute of Medicine. 1999. *To Err Is Human: Building a Safer Health System.* Washington, DC: National Academy Press.

Krein, S. L., T. P. Hofer, E. A. Kerr, and R. A. Hayward. 2002. "Who Should We Profile? Examining Diabetes Care Practice Variation among Primary Care Providers, Provider Teams and Healthcare Facilities." *Health Services Research* 37 (5): 1159–80.

Leape, L. L. 2000. "Institute of Medicine Medical Error Figures Are Not Exaggerated." *Journal of the American Medical AssociationA* 284: 95–7.

Leape, L. L., T. A. Brennan, N. Laird, A. G. Lawthers, A. R. Localio, B. A. Barnes, L. Hebert, J. P. Newhouse, P. C. Weiler, and H. Hiatt. 1991. "The Nature of Adverse Events in Hospitalized Patients. Results of the Harvard Medical Practice Study II." *New England Journal of Medicine* 324: 377–84.

Lyon, R. G. "Optical Systems Characterization and Analysis Research Project" [accessed December 27, 2005]. Available at http://satjournal.tcom.ohiou.edu/pdf/lyon.pdf

McDonald, C. J., M. Weiner, and S. L. Hui. 2000. "Deaths Due to Medical Errors Are Exaggerated in Institute of Medicine Report." *Journal of the American Medical Association* 284: 93–5.

Oppenheimer, L., and U. Kher. 1999. "The Impact of Measurement Error on Comparison of Two Treatments Using A Responder Analysis." *Statistical Medicine* 18: 2177–88.

Schulzer, M., D. R. Anderson, and S. M. Drance. 1991. "Sensitivity and Specificity of a Diagnostic Test Determined by Repeated Observations in the Absence of an External Standard." *Journal of Clinical Epidemiology* 44: 1167–79.

Shen, W., and T. A. Louis. 2000. "Triple-Goal Estimates for Disease Mapping." *Statistical Medicine* 19: 2295–308.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: Chapman & Hall/CRC.

Sox, H. C. Jr, and S. Woloshin. 2000. "How Many Deaths Are Due to Medical Error? Getting the Number Right." *Effective Clinical Practice* 3: 277–83.

Thomas, E. J., S. R. Lipsitz, D. M. Studdert, and T. A. Brennan. 2002. "The Reliability of Medical Record Review for Estimating Adverse Event Rates." *Annals Internal Medicine* 136: 812–6.

Thomas, E. J., D. M. Studdert, H. R. Burstin, E. J. Orav, T. Zeena, E. J. Williams, K. M. Howard, P. C. Weiler, and T. A. Brennan. 2000. "Incidence and Types of Adverse Events and Negligent Care in Utah and Colorado." *Medical Care* 38: 261–71.

Thomas, E. J., D. M. Studdert, W. B. Runciman, R. K. Webb, E. J. Sexton, R. M. Wilson, R. W. Gibberd, B. T. Harrison, and T. A. Brennan. 2000. "A Comparison of Iatrogenic Injury Studies in Australia and the USA. I: Context, Methods, Casemix, Population, Patient and Hospital Characteristics." *International Journal of Quality Health Care* 12: 371–8.