Studying the Prevention of Acute Kidney Injury: Lessons from an 18th-Century Mathematician

Glenn M. Chertow,* Paul M. Palevsky,† and Thomas Greene‡

*Division of Nephrology, Departments of Medicine, Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California; †Renal Section, VA Pittsburgh Healthcare System, Pittsburgh, Pennsylvania; and ‡Department of Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, Ohio


Acute kidney injury (AKI) complicates approximately 5 to 15% of hospitalizations, depending on its definition (1–3), and is independently associated with a five-fold or more increase in in-hospital mortality rates (3–5). AKI also extends length of stay, obligates excess hospital expenditures, and may exert long-term adverse effects, including an increased risk for ESRD.

Several known mechanisms contribute to the development of AKI, including ischemia, vasoconstriction, toxic injury related to selected endogenous substances (e.g., myoglobin), radiocontrast and drugs (e.g., amphotericin B), and microcirculatory changes, as observed with sepsis and other inflammatory states (6). Given the dire consequences that are associated with AKI, efforts at prevention seem desirable and worthy of intense investigation.

Unfortunately, most episodes of AKI cannot be predicted readily, either from clinical criteria or from the timing of events. The vast majority of prevention trials in kidney disease have been conducted in the setting of radiocontrast exposure. Several recently published studies in AKI prevention have attracted significant attention and have changed practice considerably (7–11). In this report, we scrutinize three of these studies, focusing on issues related to effect estimates and statistical power, and incorporate principles of Bayes’ Theorem in our interpretation of study results. More general, we provide a cautionary note regarding the interpretation of “positive” studies with an insufficient sample size and a “significant” (P < 0.05) result.

Feasibility of AKI Prevention Strategies

Several routinely used clinical strategies may attenuate AKI. These include avoidance of volume depletion, a trial of volume expansion when feasible, avoidance of nonessential nephrotoxic agents (e.g., nonsteroidal anti-inflammatory drugs), and the judicious use of radiocontrast-enhanced imaging studies. Any prevention trial would need to consider these clinical factors in study design.

Determination of a study’s sample size requires multiple inputs, including an accurate estimate of the incidence rate in the placebo- or “usual therapy”–treated group, a reasonable estimate of the intervention’s effect, and an appropriate α error rate (i.e., the rate of falsely concluding a treatment effect when in fact a treatment effect is absent). Most important, the sample size should lead to adequate power to detect between the treatment groups a difference whose magnitude is clinically relevant and biologically plausible. In the setting of AKI, the incidence rate depends greatly on both the AKI definition used and the specific characteristics of the population studied. More specific definitions (e.g., large nominal or percentage changes in serum creatinine or other biomarkers) would yield lower incidence rates. In most published studies of prevention of radiocontrast-associated AKI, only relatively small changes in serum creatinine concentrations were required to meet the AKI definition (e.g., 25 to 50% or ≥0.5 mg/dl increase in serum creatinine).

Most published studies of AKI prevention failed to document the assumptions that were used in sample size calculations or failed to document sample size estimates altogether. However, estimates of effect size in excess of 20 to 40% of the control group event rate are unrealistic on the basis not only of the complexity of AKI but also of the proven efficacy of some of the most successful interventions to change medical practice (e.g., antibiotics for serious infections, lipid lowering in secondary prevention of cardiovascular disease). Ideally, the power of intervention trials to detect a biologically plausible treatment effect should be 90% or more, particularly when the intervention carries significant risk, although power estimates as low as 80% often are chosen. When the power is <80% (i.e., the β error exceeds 20%), the likelihood that a truly effective intervention may be deemed ineffective (i.e., a false negative) generally is considered unacceptably high. Issues of “false positive” clinical trials rarely are raised, either by investigators or by journal editors.

Potential Models for AKI Prevention Trials

For AKI prevention trials to be conducted, several necessary (but not sufficient) factors must be in place. These include a...
uniform definition of AKI, a predictable setting for injury, a high incidence of AKI, and, if possible, uncomplicated nonrenal manifestations. Three settings in modern medical practice qualify on these factors: Radiocontrast exposure, preoperative cardiovascular surgery (e.g., coronary artery bypass grafting, open abdominal aortic aneurysm repair), and, possibly, intensive care unit (ICU) admission. Whereas the risk for AKI among ICU patients is high, AKI is relatively uncommon after radiocontrast or cardiovascular surgery, except among individuals with moderate to advanced chronic kidney disease (CKD; e.g., estimated GFR or creatinine clearance <60 ml/min, particularly when accompanied by diabetes and/or atherosclerotic vascular disease (12,13).

Reasonable Sample Size Estimates

If we consider an exceptionally high-risk group (patients with CKD as a result of diabetic nephropathy) and modest definition of AKI (25 to 50% increase in serum creatinine), then a generous AKI incidence rate for control subjects after radiocontrast exposure or cardiovascular surgery might be 20%. Then consider an extraordinarily effective agent that could reduce the event rate by 50%, yielding a 10% incidence rate in the experimental group. Assuming an α error of 0.05 and a power of 90%, the required sample size would be 572 patients; with 80% power, the required sample size would be 438 patients. A robust but slightly more realistic effect estimate (40 rather than 90%) would require 928 patients. These estimates should be considered as we review the published literature on AKI prevention trials.

Key Publications on AKI Prevention

In 2000, Tepel et al. (7) published one of the most influential papers on the prevention of radiocontrast nephropathy to date. Eighty-three patients with mild to moderate CKD were enrolled in a randomized, clinical trial that compared standard therapy (0.45% NaCl administered at 1 ml/kg body wt per h for 12 h before and after radiocontrast exposure) and standard therapy plus N-acetylcysteine 600 mg twice daily on the day before and the day of radiocontrast exposure. Nine (21%) of 42 patients in the standard therapy group developed AKI compared with one (2%) of 41, yielding P = 0.01. No power calculations were reported. The relative risk was calculated at 0.10, with a 95% confidence interval of 0.02 to 0.90.

Marenzi et al. (10) published another high-profile paper on the use of hemofiltration for prevention of radiocontrast nephropathy. In this study, 114 consecutive patients who had moderate to severe CKD (serum creatinine > 2.0 mg/dl) and underwent percutaneous coronary intervention were randomly assigned to hemofiltration in an ICU versus 0.9% NaCl at 1 ml/kg body wt per h in a stepdown unit for 4 to 8 h before and 18 to 24 h after their procedure. An increase in serum creatinine of 25% or more was observed in 5% of hemofiltration-treated and 50% of saline-treated patients (P < 0.001). In-hospital and 1-yr mortality rates also were significantly reduced in the hemofiltration-treated group. In this study, a sample size calculation was provided. The authors estimated that radiocontrast nephropathy would develop in 40% of control subjects and in 30% of patients who were on hemofiltration (i.e., a 25% relative and 10% absolute risk reduction). The α error was 0.05, and the desired power was 80%. Whereas the authors estimated a required sample size of 50 patients per group, our estimates on the basis of the published assumptions would have yielded a study of 752 patients, or 376 patients per group (14). As with the Tepel et al. (7) paper, Marenzi et al. (10) offered no discussion on the extraordinary treatment effects observed and no comment on the possibility that the findings may not have reflected the true effect(s) of the intervention (indeed, the hemofiltration itself would be expected to lower serum creatinine concentrations).

More recently, Merten et al. (11) published the results of a study that compared the relative efficacy of a sodium bicarbonate–based versus sodium chloride–based intravenous fluid strategy for prevention of radiocontrast nephropathy in patients who underwent cardiac catheterization, computed tomography, or other procedures that required radiocontrast administration. As designed, the investigators planned to enroll a total of 260 patients to detect a 10% absolute difference in the incidence of radiocontrast nephropathy (15% in the sodium chloride–treated versus 5% in the sodium bicarbonate–treated group) with an α error rate of 0.05 and power of 80%. The study was halted at approximately the midpoint by a safety monitor because of a lower rate of radiocontrast nephropathy in the sodium bicarbonate–treated group. Eight (13.6%) of 59 patients in the sodium chloride–treated group had met the definition of AKI as compared with only one (1.7%) of 60 patients in the sodium bicarbonate–treated group, for a reduction in the incidence of radiocontrast nephropathy of 11.9% (95% confidence interval 2.6 to 21.2%). Although the reported P value was 0.02, it should be noted that if one additional patient who was treated with sodium bicarbonate had sustained an increase in serum creatinine of ≥25%, then the study would not have reached conventional levels of statistical significance. Although it could be argued that the study was terminated prematurely, to the authors’ credit, an additional 191 consecutive patients who were treated with open-label sodium bicarbonate (deemed a “registry phase”) were evaluated. Among these 191 patients, there were three documented cases of radiocontrast nephropathy, yielding an incidence of 1.6%—a rate virtually identical to that observed in the randomized trial. Detailed characteristics of the “registry” population were not provided. These three studies highlight the medical community’s ongoing fascination with the P value; because all three studies were “positive,” no objections were brought forward on the grounds of insufficient power.

Importance of Previous Information

To illustrate the importance of previous information in the interpretation and the design of intervention trials, we briefly digress to baseball and snack foods. First, let us consider a solid outfielder with home run totals of 17, 20, 25, 31, 20, 20, 30, and 24 in 151 to 162 games during 1998 through 2005, yielding an average home run production of 23 per 162-game season. Imagine that during April 2006, the outfielder hits nine home runs. Without consideration of the outfielder’s past performance, if
one were to extrapolate his performance over the 6-mo 2006 season, then one would estimate a total of 54 home runs, assuming no significant change in pitching prowess, strength, injuries, and the like. In this context, it is clear that previous information (i.e., past evidence of performance over the long term) should not be ignored when predicting future events.

The Reverend Thomas Bayes put forward these principles in 1764, in “An Essay Toward Solving a Problem in the Doctrine of Chances” (15). Bayes’ theorem suggests that the appropriate estimate of the odds that a hypothesis is true, after having observed the results of a study, depends equally on the previous odds that the hypothesis is true and the “likelihood ratio,” which reflects the compatibility of the data with the hypothesis being evaluated. From this perspective, the problem with underpowered studies is that the effect sizes that are necessary to achieve adequate power are implausibly large, suggesting that the previous odds of the research hypotheses that actually are being tested are small. If the previous odds of the research hypotheses are small, then Bayes’ theorem tells us that the probability that the hypotheses are true can remain low even in the presence of trends in the data that seem to support them. Despite the widespread application of Bayes’ theorem in multiple settings in medicine, including the evaluation of diagnostic studies with imperfect performance characteristics (i.e., sensitivity and specificity), few published clinical trials have been interpreted explicitly in this context.

Now imagine a bag of 10,000 nuts. Some are almonds and some are cashews, but the exact proportion of nuts is unknown. It is unnecessary to count all of the nuts to make some statement about this proportion. For example, a randomly acquired sample of 1000 nuts may be sufficient to make an inference about the proportion of almonds and cashews in the entire population. If almonds compose 40% of the 1000-nut sample, then we may be able to infer that approximately 40% of the population of nuts also are almonds. To laypersons (or to the editorial boards of some medical journals), this process may seem straightforward. In fact, it might seem that there is no need even to acquire a sample of 1000 nuts. A sample of 100 or even 10 nuts might do. However, as the sample size becomes smaller, the potential for error grows. For this reason, inferential statistics has developed numerous techniques for stating the level of confidence that can be placed on these inferences. The confidence in the estimate of the proportion of almonds (or cashews) will be higher with sample of 1000 than 100.

It is troubling that the most prominent studies of AKI prevention, with absent or incorrect power calculations but $P < 0.05$, generally have been accepted, despite that each study essentially is equivalent to bags of nuts that are too small for clear conclusions. As a result, there is so much variability in the data that these studies have adequate power to detect only very large, biologically implausible effects (which have low prior probabilities of being true). It is highly unlikely that N-acetyl cysteine or hemofiltration is so effective in such a complex disease as to exert a 90% treatment benefit. If one were to consider reasonable estimates of treatment effects given previous information—for example, a 10, 20, or even 30% treatment effect—then a larger bag of nuts would be required to provide a low enough error rate to yield confident conclusions. In other words, despite the significant $P$ value, there is a relatively high likelihood that the study results cited above were in error (i.e., false positive results). For example, other clinical trials and meta-analyses on the effects of N-acetyl cysteine for prevention of radiocontrast nephropathy have yielded conflicting results; none as positive as those reported by Tepel et al. (7).

### True versus False Positive Results: A Simulation

To illustrate these issues, we performed a small simulation study to show how the implications of the results of a study depend on its sample size. Consider as an example a research setting in which the true relative risks of 95% of the treatments that are studied range between 0.70 (corresponding to a 30% benefit) and 1.10 (corresponding to a 10% adverse effect), with a mean true relative risk of 0.90. For simplicity, we also assume that these true relative risks are normally distributed. This is a setting in which a substantial proportion of the interventions that are tested have beneficial effects of 10 to 30% but relatively few have benefits larger than 30%. The implications in this setting of various observed relative risks for studies of various sample sizes are illustrated in Table 1. For example, suppose a relative risk of 0.20 (corresponding to an 80% treatment effect) is reported in a randomized trial with 50 patients randomly assigned to each group (for $N = 100$ patients). This relative risk would be reported to be statistically significant at the $P < 0.05$ level. However, with a total $N$ of 100, the average true relative risk given a reported relative risk of 0.20 is 0.82. In other words, even when the observed relative risk seems to indicate a risk reduction of 80%, the mean true risk reduction in this setting is only 18%. Moreover, only 11% of the interventions that are associated with observed relative risks of 0.20 would be expected to have true relative risks smaller than 0.70. This example illustrates the point that even if one obtained a nominally significant positive result at the $P < 0.05$ level, this result will have very limited implications if the study was not powered adequately to detect a plausible treatment effect.

### Conclusion

Although we may be able to prevent some cases of AKI with novel therapeutic agents, it will be very difficult to prove. Existing clinical trials should be scrutinized for sample size considerations, regardless of whether the intervention is deemed efficacious or of nil effect. To be successful, future prevention trials will require better identification of very-high-risk subgroups (with the aid of biomarkers), earlier identification of injury (with the aid of laboratory studies [e.g., cystatin C] or imaging techniques), and very effective interventions. For now, clinical research efforts should focus on developing highly effective and efficient treatment strategies for established AKI. Moreover, clinical trial design should incorporate plausible effect estimates, and these should be considered when interpreting the strength of any study’s findings, regardless of the $P$ value reported.
Table 1. Dependence of implications of observed treatment effects on sample size

<table>
<thead>
<tr>
<th>Sample Size per Group</th>
<th>Summary Measure</th>
<th>0.20</th>
<th>0.40</th>
<th>0.60</th>
<th>0.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>Mean true risk ratio (% of studies for which true risk ratio is &lt;0.70)</td>
<td>0.82b (11%)</td>
<td>0.85 (6%)</td>
<td>0.88 (4%)</td>
<td>0.89 (2%)</td>
</tr>
<tr>
<td>100</td>
<td>Mean true risk ratio (% of studies for which true risk ratio is &lt;0.70)</td>
<td>0.76b (30%)</td>
<td>0.81b (13%)</td>
<td>0.85 (5%)</td>
<td>0.89 (2%)</td>
</tr>
<tr>
<td>200</td>
<td>Mean true risk ratio (% of studies for which true risk ratio is &lt;0.70)</td>
<td>Not observedc</td>
<td>0.75b (30%)</td>
<td>0.82b (9%)</td>
<td>0.88 (2%)</td>
</tr>
<tr>
<td>400</td>
<td>Mean true risk ratio (% of studies for which true risk ratio is &lt;0.70)</td>
<td>Not observedc</td>
<td>0.66b (65%)</td>
<td>0.77b (18%)</td>
<td>0.86 (2%)</td>
</tr>
<tr>
<td>800</td>
<td>Mean true risk ratio (% of studies for which true risk ratio is &lt;0.70)</td>
<td>Not observedc</td>
<td>Not observedc</td>
<td>0.72b (41%)</td>
<td>0.85b (1%)</td>
</tr>
</tbody>
</table>

aThe table summarizes the mean true risk ratios and the percentage of true risk ratios <0.70 that are associated with observed risk ratios of 0.20, 0.40, 0.60, and 0.80 on the basis of the indicated sample sizes for a research setting in which the true risk ratios are normally distributed with a mean of 0.9 and an SD of 0.1. The absolute risks in the control group are assumed to be normally distributed with a mean of 0.20 and SD of 0.025.

bThe observed risk ratio is significantly smaller than 1 with \( P < 0.05 \).
cIt is essentially impossible (probability <10\(^{-6}\)) that an observed risk ratio this small would be observed on the basis of the indicated sample size.

References