American Journal of Transplantation 2014; 14: 1271–1276
Wiley Periodicals Inc.

doi: 10.1111/ajt.12707

Personal Viewpoint

Bayesian Methods for Assessing Transplant Program Performance

N. Salkowski[1], J. J. Snyder[1,2], D. A. Zaun[1], T. Leighton[1], A. K. Israni[1,2,3] and B. L. Kasiske[1,3,*]

[1]*Scientific Registry of Transplant Recipients, Minneapolis Medical Research Foundation, Minneapolis, MN*
[2]*Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN*
[3]*Department of Medicine, Hennepin County Medical Center, Minneapolis, MN*
[*]*Corresponding author: Bertram L. Kasiske, kasis001@umn.edu*

Based on recommendations from a recent consensus conference and a report commissioned by the Centers for Medicare & Medicaid Services to the Committee of Presidents of Statistical Societies, the Scientific Registry of Transplant Recipients (SRTR) plans to adopt Bayesian methods for assessing transplant program performance. Current methods for calculating program-specific reports (PSRs) often generate implausible point estimates of program performance, wide confidence intervals and underpowered conventional statistical tests. Although technically correct, these methods produce statistical summaries that are prone to misinterpretation. The Bayesian approach assumes that performance of most programs is about average and few programs perform much better or much worse than average; thus, strong evidence is required to conclude that performance is extremely good or poor. In Bayesian statistics, inference is performed via a posterior probability distribution, which reflects both the available data and prior beliefs about what model parameter values are most likely. In the PSRs, the posterior distribution of a program-specific hazard ratio will show whether a program is likely to be performing better or worse than average. Bayesian-derived PSRs will be available for preview by programs on the private SRTR website in mid-2014 and will likely replace current methods for public reporting in early 2015.

Keywords: Graft survival, mortality, quality assurance, solid organ transplantation

Abbreviations: CMS, Centers for Medicare & Medicaid Services; E, expected event count; MPSC, Membership and Professional Standards Committee; O, observed event count; OPTN, Organ Procurement and Transplantation Network; PSRs, program-specific reports; SRTR, Scientific Registry of Transplant Recipients

Received 07 August 2013, revised 09 January 2014 and accepted for publication 19 January 2014

Introduction

We have entered the era of performance assessment in health care. Payers and other stakeholders have begun examining the outcomes of hospitals and other providers to assess quality performance. However, in arguably no area of health care in the United States has outcomes assessment received more attention than in solid organ transplantation. The National Organ Transplantation Act (1984 Pub.L. 98-507) mandates that the Scientific Registry of Transplant Recipients (SRTR) produce semiannual reports of transplant program performance (42 USC §121.11(b)) (1). These reports include information on risk-adjusted graft and patient survival after transplant.

The spirit of these reporting requirements is to inform both the general public and the agencies charged with regulatory oversight about the performance of individual transplant programs. Members of the general public may be interested in comparing program performance to make informed decisions about where to seek care. Regulatory bodies, for example, the Membership and Professional Standards Committee (MPSC) of the Organ Procurement and Transplantation Network (OPTN) and the Survey and Certification body of the Centers for Medicare & Medicaid Services (CMS), can use the reports to track whether programs are improving. The design of screening processes used by regulatory agencies to identify programs for review is an important issue, which is related to, but separate from, the design of program assessments. This article addresses development of new program assessment methodologies, rather than the use of those methodologies as a screening tool, which is addressed in a companion article (2).

SRTR and OPTN hosted a consensus conference on transplant program quality and surveillance in February 13–15, 2012 (3). A key recommendation of this conference was to explore use of Bayesian hierarchical, mixed-effects statistical methods to assess program performance.

Coincidentally, a report commissioned by CMS to the Committee of Presidents of Statistical Societies, published in January 2012, also recommended using Bayesian hierarchical, mixed-effects models in assessing hospital performance (4). SRTR has developed Bayesian methods for assessing transplant programs. In this overview for non-statisticians, we explain what these new methods are and how they will alter the SRTR program-specific reports (PSRs).

Rationale for a New Method to Assess Transplant Program Performance

Program assessments based on data analysis can never produce perfect certainty. Risk-adjustment models only approximate reality. Each program's outcomes result from the program's actions and from other seemingly random events. Despite these limitations, the program assessments included in the PSRs should provide a reasonable estimate of program performance and explain the precision of that estimate.

To date, SRTR has employed ''frequentist'' statistical methods (5) to assess transplant program performance (6). These methods are designed to answer a yes-or-no question, ''Is the performance of a program better or worse than expected?'' SRTR uses risk-adjustment models to calculate an expected event count (E), which is then compared to the observed event count (O). The statistical test assumes that E is the program's true event rate, then calculates the probability of observing O or more events and the probability of observing O or fewer events. If the probability of observing O or more events is less than 1 in 40, SRTR concludes that the program's performance is ''lower than expected.'' If the probability of observing O or fewer events is less than 1 in 40, SRTR concludes that the program's performance is ''higher than expected.'' Otherwise, SRTR concludes that the program's survival is ''as expected.'' That is, SRTR rejects the hypothesis that the program's performance is ''as expected'' only if the observed event count is unusually high or unusually low when the program's performance is assumed to be precisely ''as expected.'' The observed-to-expected ratio (O/E) is used to estimate the relative rate at which a program's transplant recipients experience graft failures or death compared with the expected event rate based on data from all transplant programs. A 95% confidence interval for the O/E ratio shows the range of possible O/E ratios that is reasonably consistent with the observed data.

The current frequentist statistical framework can lead to questionable summaries of program performance. Consider one extreme example from the July 2012 PSRs. One heart transplant program performed only a single adult transplant during the 30 months covered by the 1-year patient survival cohort. The recipient died, producing an O/E ratio of 42.16 for adult patient survival with a confidence interval of 1.07–234.91. Although this program did experience more adult patient deaths (1) than expected (0.02), the available data are limited to a single transplant. The frequentist analysis produces several bold claims that may not be supported by strong evidence: adult patient survival at the program in question is ''lower than expected'' ($p = 0.047$), the best estimate of the program's risk is over 40 times the risk of an average program, and the risk could be over 200 times the risk of an average program.

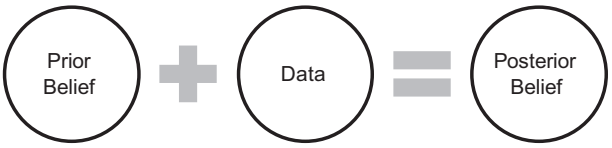

Figure 1: Diagram showing how Bayesian methods tell an observer how to update a prior belief about a program's performance after observing a set of new data, yielding a posterior belief.

Bayesian inference is a statistical framework based on a theorem developed by the Reverend Thomas Bayes (1701–1761) (7). Bayesian methods tell an observer how to update a prior belief about a program's performance after observing a set of new data. Thus, we start with a prior belief about the expected distribution of performance at transplant programs in the United States, then weigh that belief against the data we observe to yield a posterior belief (Figure 1). Both frequentist and Bayesian analyses depend upon conditional probability. As noted, the frequentist hypothesis test is based on the probability of observing the data conditioned on one particular belief, the hypothesis that a program's performance is exactly ''as expected.'' Bayesian analyses reverse the conditioning by employing a prior that describes the belief held before observing the data. A Bayesian analysis would thus produce a belief about a program's performance conditioned on the observed data and a prior belief about the program's performance. This new belief is called a posterior because it describes the belief held after observing the data.

The posterior depends on both the available data and the prior. When data are abundant, the posterior strongly reflects the data. When data are limited, for example, at smaller programs, the posterior tends to resemble the prior (Figure 2).

More precisely, the Bayesian method yields a ''posterior distribution,'' a probability distribution for a program's performance given the observed data and our prior beliefs. In the context of evaluating transplant program performance, the Bayesian method yields a probability distribution for the program's hazard ratio, telling the observer how probable various levels of performance are. In contrast to the frequentist approach, which answers a yes-or-no question, the Bayesian approach answers the question, ''What is the probability that the performance of Program X is worse than a certain threshold?''

Choosing the prior can sometimes be controversial. In practice, there is no way to choose a prior that pleases

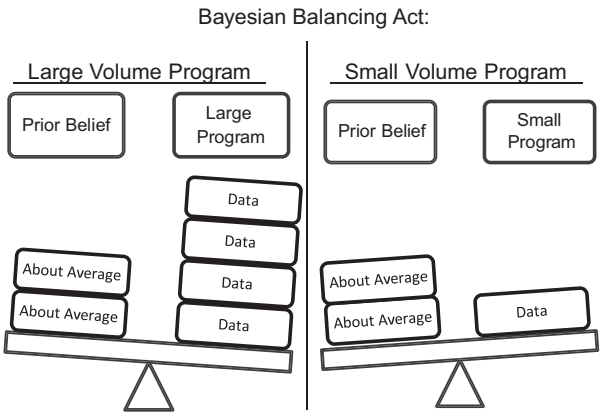

Figure 2: Diagram showing effect of data availability. When data are abundant, the posterior strongly reflects the data. When data are limited, for example, at smaller programs, the posterior tends to resemble the prior.

everyone. After much deliberation by the SRTR Technical Advisory Committee, use of the prior shown in Figure 3 was recommended for the following reasons:

1. It is conservative in the sense that it overestimates the true spread of program variability in the United States. Looking at the spread of this curve, we see that we believe (before looking at each program's data) that most programs are performing generally between hazard ratios of about 0.25 and 2.5, and few programs perform worse than 2.5. SRTR analyzed historical PSR data to estimate program variability, and found that the chosen prior has approximately double the standard deviation suggested by historical program variability in graft and patient survival for deceased donor heart, kidney, liver and lung transplants in adult recipients. This is attractive in the context of PSR evaluations because less data will be required for us to change our prior belief, in essence allowing the data from smaller programs to continue to influence our evaluation. If we used a more ''informative'' prior, one more in line with true variability in the United States, only larger programs would produce enough data to noticeably influence our assessment.
2. This prior has convenient mathematical properties that allow us to apply the Bayesian approach while using the existing risk-adjustment models to arrive at the expected event count. Using the existing risk-adjustment model structure will allow SRTR to continue to provide tools to transplant programs to run their own subgroup analyses of their data.

Hypothetical Examples

Consider a large transplant program, Program A, which performed 299 transplants in the most recent 2.5-year period of evaluation. During the first year of follow-up, Program A experienced 13 patient deaths. Is this a problem or a random happenstance? We estimate that Program A should have experienced 6.97 deaths given its case mix. Using the current frequentist methods, this yields an observed-to-expected ratio (O/E) of 1.87 (95% confidence interval 0.99–3.19) and a one-sided p-value of 0.026.

The Bayesian approach yields a probability distribution for Program A's hazard ratio (Figure 4, left panel). The bell-shaped curve indicates where Program A's hazard ratio (analogous to O/E) is most likely to be, given our prior belief and the observed outcomes. The average of this distribution is indicated by the vertical line on the figure (hazard ratio = 1.67). Our best guess is that Program A's death rate is 67% higher than expected based on national experience, and the general shape and location of the curve give a broader picture of how this program is performing. Of note, this estimate (1.67) is closer to 1.0 than the frequentist

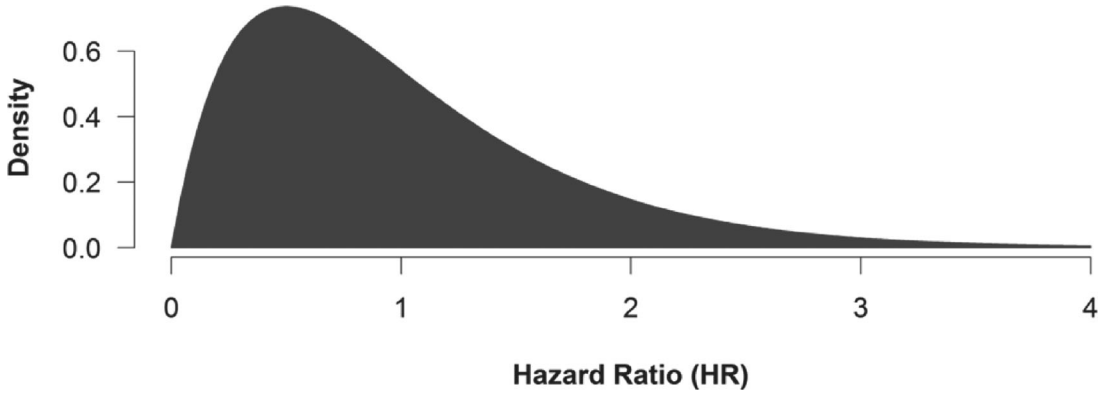

Figure 3: Gamma prior with mean of 1.0 and variance of 0.5 (SD = 0.71). The hazard ratio for each program is on the x-axis. A hazard ratio of 1 indicates a program that is performing exactly as expected and a hazard ratio of 2 a program with twice as many events as expected. The y-axis (labeled ''Density'') shows how frequent we believe this hazard ratio to be across all programs.

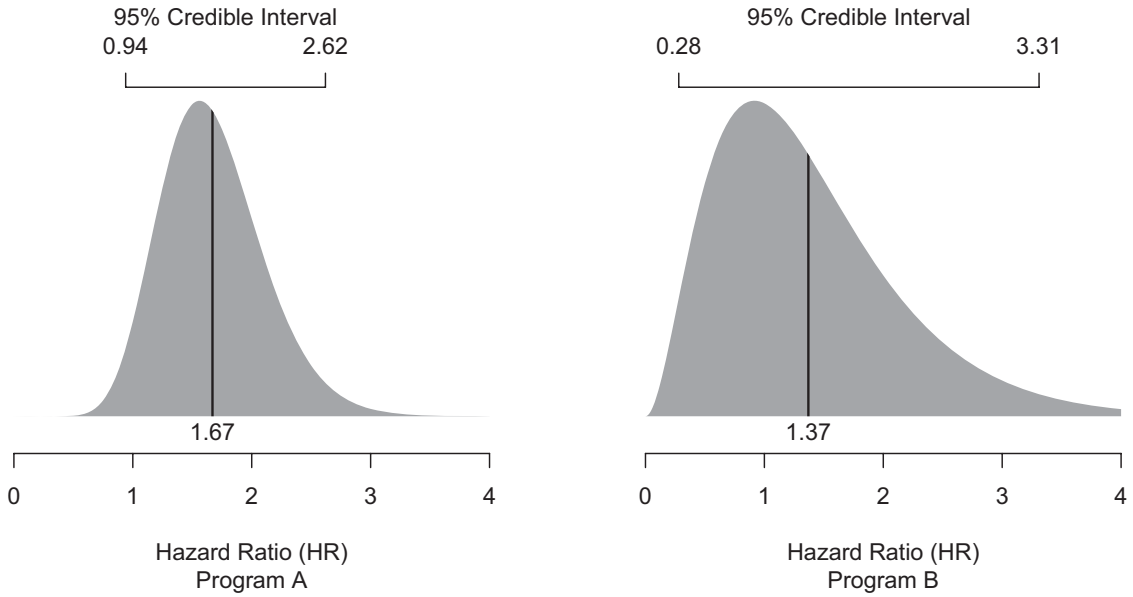

Figure 4: Probability distributions of the hazard ratios for graft survival at two hypothetical transplant programs. For large Program A (left panel), the average is indicated by the vertical line (hazard ratio = 1.67); the best guess is that this program's death rate is 67% higher than expected based on national experience. The program's true hazard ratio is within the 95% credible interval range (0.94–2.62). For small Program B (right panel), the average is indicated by the vertical line (hazard ratio = 1.37); the best guess is that this program's death rate is 37% higher than expected based on national experience. The program's true hazard ratio is within the 95% credible interval range (0.28–3.31). See text for details.

estimate of O/E, which was 1.87. This is because we incorporated our prior belief into the system. In this case, our prior belief was that most programs in the nation perform about average. Since this is a large program with plenty of data, we had plenty of evidence that it was underperforming, moving the hazard ratio substantially away from 1.0 but not as far as 1.87. In addition to the best guess estimate of the program's hazard ratio (1.67), we can calculate a 95% credible interval; the program's true hazard ratio is within this range with 95% certainty, given our prior belief. In this case, the 95% credible interval is 0.94–2.62. While the 95% credible interval is similar in nature to the traditional 95% confidence interval in the frequentist statistical framework, a subtle distinction is important (see Glossary).

Now consider Program B, a small program that performed only six transplants during a 2.5-year period and experienced one patient death. Is Program B underperforming? Is one death in six transplants extreme? Or could this death have occurred by chance even if the program's performance is not problematic? Using the frequentist methods, Program B has an O/E of 5.42 (95% confidence interval 0.14–30.20, one-sided p-value = 0.17). The Bayesian method yields a posterior distribution with a mean of 1.37 (95% credible interval 0.28–3.31; Figure 4, right panel).

Differences between the frequentist approach and the Bayesian approach are more pronounced in this example. The frequentist approach yielded an O/E estimate of 5.42 while the Bayesian approach yielded an estimate of 1.37. Our prior belief was that Program B's performance was likely to be about average. Six transplants are not enough to provide strong evidence for poor performance or for good performance. After observing the data, we are more likely to conclude that Program B's performance is worse than expected, but 1.37 is much less extreme than 5.42. The shape of the curve, however, conveys how uncertain we are of this conclusion; a substantial portion of the curve is still less than 1.0, indicating possible good performance. This approach is arguably more informative than simply providing one O/E ratio of 5.42 with a wide confidence interval. Do we really believe that mortality rates for some programs are more than five times the national average? Conversely, if Program B had experienced zero deaths among its six transplants, the O/E would be 0. Concluding that the risk of death is 0% after transplant for any program is nonsensical, especially given so little data. The Bayesian estimate in this instance would be a hazard ratio of 0.92, better than expected given the evidence observed, but not 0. The Bayesian approach will yield more reasonable estimates of performance given our prior belief that variation in program performance across the nation is not extreme.

Advantages of the Bayesian Approach

The Bayesian approach has several important advantages over current methods SRTR uses. Due to the plausible

assumption that a program's performance is unlikely to be extremely different from the performance of an average program, estimates of program performance will be less extreme (due to ''statistical shrinkage''), and evaluation cycles will show less fluctuation. The probability curve (''posterior distribution'') provides a visual display of the likely location of a program's hazard ratio. This visual display can quickly and easily convey a sense both of how far away from 1.0 the program's hazard ratio is and of how certain that assessment is (i.e., narrower bell-shaped curves convey more certainty and wider bell-shaped curves convey less certainty).

Bayesian posterior distributions will also allow SRTR to summarize each program's performance to more understandably convey difficult statistical concepts to stakeholders without statistical backgrounds. Most people are familiar with statements of probability, for example, ''There is a 75% chance of rain tomorrow,'' or ''This flight is on time only 33% of the time.'' Bayesian methods allow us to make statements like, ''The probability that a program is underperforming is 70%,'' or ''We are 80% certain that this program's graft failure rate is at least 33% higher than we would expect if the program were performing at the level of the national average.''

Limitations to the Bayesian Approach

Bayesian analysis requires the statistician to assert a prior belief about the situation under study. In this case, we assert a prior belief about the variation in performance across all US transplant programs. We introduce this belief into the Bayesian analysis in the form of a prior probability distribution. The choice of the prior distribution continues to spark controversy in the field of statistics. The prior SRTR uses was recommended by the SRTR Technical Advisory Committee because of its mathematical properties (it is the conjugate prior of the Poisson distribution), because it has a mean of 1 (forcing statistical shrinkage toward the national average), and because it provides a nice balance between an uninformative prior and a prior based on the empirical data; that is, it has a reasonable variance.

Using a gamma prior is attractive because it is the conjugate prior to the Poisson distribution. This mathematical property will enable SRTR to continue providing transplant programs with tools they can use to perform recipient subgroup analyses to better target quality improvement efforts. Without using a conjugate prior, it would be more difficult for SRTR to provide simple tools for programs to use.

The variance of the prior is 0.5, yielding a standard deviation of 0.71, suggesting that most programs should be performing between a hazard ratio of about 0.25 and 2.0. This is wider variation than is observed in the national data, but it is desirable for a prior to be more vague than the national data would suggest so that the observed data can still factor into the conclusions. If a smaller variance prior was chosen, the data from small-to-mid-volume programs would rarely matter and most would appear to be about average.

Also of note, moving to a Bayesian framework for assessing transplant program performance does not change the underlying risk-adjustment models used to estimate expected outcomes. This is perhaps a positive, in that programs are familiar with the current risk-adjustment models SRTR uses to estimate expected outcomes. However, any and all current limitations of the risk-adjustment models will carry forward into the Bayesian framework. In response to a recommendation from the PSR consensus conference (3), SRTR has implemented a 3-year model rebuild cycle to try to improve the performance of the risk-adjustment models in collaboration with the OPTN organ-specific committees. As new and better models are developed, they will improve the Bayesian results, but this is not a limitation of the Bayesian methodology per se.

Finally, statisticians will long debate whether frequentist or Bayesian philosophies should be the prevailing method employed. While SRTR cannot resolve this debate, we can follow the prevailing opinions expressed by the presidents of the statistical societies in the Committee of Presidents of Statistical Societies report, by experts in the field of transplantation at the PSR consensus conference, and by experts on the SRTR Technical Advisory Committee.

Future Directions

SRTR, under the direction of the SRTR Technical Advisory Committee, has completed development of a Bayesian framework for the PSRs. With the approval of the US Health Resources and Services Administration, SRTR will begin posting Bayesian-derived PSRs on programs' private websites for their review; results using the current method for calculating PSRs will appear along with results using Bayesian methods. If all goes well, PSRs calculated with Bayesian methods will be made publicly available in the future and will likely replace PSRs calculated with current methods. OPTN's MPSC is considering how best to use PSRs calculated with Bayesian methods and specifically what clinically significant thresholds to use for flagging programs for further scrutiny (a topic discussed in the companion piece to this article (2)). SRTR will continue to work with the MPSC if and when the new method is implemented to assess the performance of the system. CMS will also consider whether to use the same methods adopted by the MPSC.

Acknowledgments

This work was conducted under the auspices of the Minneapolis Medical Research Foundation, contractor for the SRTR, as a deliverable under contract no. HHSH250201000018C (US Department of Health and Human Services, Health Resources and Services Administration, Healthcare

Systems Bureau, Division of Transplantation) and United Network for Organ Sharing, contractor for the Organ Procurement and Transplantation Network, under contract no. 234-2005-370011C. As a US Government–sponsored work, there are no restrictions on its use. The views expressed herein are those of the authors and not necessarily those of the US Government, United Network for Organ Sharing, or OPTN. The authors thank SRTR colleagues Delaney Berrini, BS, for manuscript preparation, and Nan Booth, MSW, MPH, ELS, for manuscript editing. The authors also thank the members of the SRTR Technical Advisory Committee for their guidance, and the anonymous editors and reviewers of earlier drafts of the manuscript for their many helpful comments.

Disclosure

The authors of this manuscript have no conflicts of interest to disclose as described by the *American Journal of Transplantation*.

Glossary

Bayesian
An approach to statistics that assumes that earlier experiments influence the design of subsequent experiments, by updating beliefs
Frequentist
An approach to statistics that draws conclusions from samples by emphasizing the frequency of an occurrence; it assumes that an experiment can be modeled as one of an infinite number of possible repetitions, each producing independent results
Hazard ratio
Rate at which events occur at a program divided by the rate at which events occur at an ''average'' program
Interpretation of the Bayesian 95% credible interval (incorporates prior probability)
The program's true hazard ratio is contained within this interval with 95% certainty, with a 2.5% probability that it is below this range and a 2.5% probability that it is above this range
Interpretation of the frequentist 95% confidence interval (based only on observed data)
If the underlying process that resulted in the observed data was repeated over and over, 95% of the confidence intervals calculated would contain the true underlying hazard ratio
Posterior probability
The probability calculated after the relevant evidence is taken into account
Prior probability
The assumption made before viewing the data of how performance tends to vary among programs
Shrinkage
In Bayesian analyses, we update a prior belief by looking at the program's data. If there are a lot of data, that is, a large program, the posterior belief will largely reflect the data observed; however, if there are not a lot of data, that is, a smaller program, the posterior belief will shift more toward the prior belief (Figure 2). This shifting is referred to as shrinkage

References

1. National Organ Transplantation Act of 1984. Pubic Law 98-507, Title 42-Section 273. 10-19-0984. 98 Stat. 2339–2348.
2. Salkowski N, Snyder JJ, Zaun DA, et al. A Scientific Registry of Transplant Recipients Bayesian method for identifying underperforming transplant programs. Am J Transplant 2014; 14: 1310–1317.
3. Kasiske BL, McBride MA, Cornell DL, et al. Report of a consensus conference on transplant program quality and surveillance. Am J Transplant 2012; 12: 1988–1996.
4. Ash AS, Fienberg SE, Louis TA, Normand S-LT, Stukel TA, Utts J. Statistical Issues in Assessing Hospital Performance. 2012. Available at: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf Accessed January 8, 2014.
5. Guillies D. *The frequency theory. Philosophical theories of probability*. London, England: Psychology Press; 2000, p. 88.
6. Dickinson DM, Arrington CJ, Fant G, et al. SRTR program-specific reports on outcomes: A guide for the new reader. Am J Transplant 2008; 8: 1012–1026.
7. Bayes T, Price MR. An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond 1763; 53: 370–418. Available at: http://www.socsci.uci.edu/~bskyrms/bio/readings/bayes_essay.pdf. Accessed January 8, 2014.