Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias

Robert M. Kaplan, Ph.D.[1], David A. Chambers, D.Phil.[2], and Russell E. Glasgow, Ph.D.[3]

Abstract

A number of commentaries have suggested that large studies are more reliable than smaller studies and there is a growing interest in the analysis of "big data" that integrates information from many thousands of persons and/or different data sources. We consider a variety of biases that are likely in the era of big data, including sampling error, measurement error, multiple comparisons errors, aggregation error, and errors associated with the systematic exclusion of information. Using examples from epidemiology, health services research, studies on determinants of health, and clinical trials, we conclude that it is necessary to exercise greater caution to be sure that big sample size does not lead to big inferential errors. Despite the advantages of big studies, large sample size can magnify the bias associated with error resulting from sampling or study design. Clin Trans Sci 2014; Volume 7: 342–346

Keywords: big data, research methods, bias, sampling

In preparation for the 1936 presidential election, the biggest public opinion poll in history of the United States was conducted. More than 2.4 million respondents indicated whether they intended to vote for governor Alfred Landon of Kansas or the incumbent president Franklin Delano Roosevelt. The poll results were clear and unambiguous. Landon would win by a landslide. Presidential elections provide one of the few opportunities to validate public opinion polls against verifiable outcomes. In the 1936 election the large sample size poll was wrong. Roosevelt won 46 of the then 48 states. Landon carried only Vermont and Maine, and because these states were small, he received only 8 electoral votes.[1]

With a sample size so large, how could the study have gone so wrong? The survey was conducted by a magazine known as the *Literary Digest*. The group surveyed its own readers and *Literary Digest* readership was skewed toward those subgroups who supported Landon, including registered automobile owners and telephone subscribers. The same year, a poll by the American Institute of Public Opinion used a sample size only 2% of the size of the *Literary Digest* poll and predicted the outcome of the election within 1% of the actual result. In fact, the popular vote in American elections can be accurately predicted with sample sizes as small as 1,500 or about 0.0006 the size of the sample in the *Literary Digest* poll. Several polls with sample sizes less than 1,000 were very accurate in predicting the result of the 2012 presidential election (available at wikipedia.org/wiki/Nationwide_opinion_polling_for_the_United_States_presidential_election,_2012).

Although large sample sizes and "big data" have a number of strengths, studies can be of relatively little value if the large sample size is not representative of the population to which the results will be generalized or is missing a key information, especially on a nonrandom basis. The *Literary Digest* example shows how sample nonrepresentativeness does not assure that large samples produce better results. The enthusiasm for "big data" encourages the use of ever-larger datasets with massive numbers of measured variables. "Big data" generally refers to "datasets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time."[1] Big data are clearly an important new direction for most areas of science as exemplified by the new National Institutes of Health initiative to transform big data into new knowledge (BD2K).[2] The goal of the BD2K initiative is to "capture the opportunities and challenges facing all biomedical researchers in accessing, managing, analyzing, and integrating datasets of diverse data types (e.g., imaging, phenotypic, molecular [including various "–omics"], exposure, health, behavioral) and many other types of biological, and biomedical and behavioral data that are increasingly larger, more diverse, and more complex, and that exceed the abilities of currently used approaches to manage and analyze effectively." (See more at: http://bd2k.nih.gov/about_bd2k.html#bigdata).

Certainly large datasets bring many advantages. They can be used to study rare events, and the integration of research protocols might greatly reduce the cost of investigation and the time required to evaluate research questions.[3] Although we share the enthusiasm for big data, investigators will surely have less intimate knowledge about the texture and often the quality of the many elements in their datasets. The purpose of this paper is to offer a cautionary perspective on the potential for serious bias in studies using very large sample sizes and large numbers of outcome measures, and to provide suggestions for maximizing interpretability of these studies. Our comments focus on one assumption that underlies much of the enthusiasm for big data: the assumption that large sample sizes yield more meaningful results than small sample sizes.

Big Data, Electronic Health Records, and the Future of Epidemiologic Research

The current enthusiasm about electronic health records (EHRs) exemplifies why these issues of potential bias needs to be understood.[4] Many investigators believe that the future of epidemiologic research will use information obtained from patients in community settings and recorded in EHRs. The proposal to build an EHR based on National Patient-Centered Clinical Research Network (PCORnet)[4] has been greeted with great enthusiasm. However, there can be systematic biases in the sample of people in EHR systems and there are often biases in the way information is obtained and recorded.[3,5] If we ignore these biases and assume they will be resolved through sheer sample size,

[1]Office of Behavioral and Social Sciences Research and Department of Rehabilitation Medicine, National Institutes of Health, Bethesda, Maryland, USA; [2]Division of Services and Intervention Research, National Institute of Mental Health, Bethesda, Maryland, USA; [3]Colorado Health Outcomes Program, University of Colorado, Anschutz, Colorado, USA.
Correspondence: Robert M. Kaplan (robert.kaplan@ahrq.hhs.gov)

The views expressed in this paper are those of the authors and do not reflect the positions of the National Institutes of Health, the National Institute of Mental Health.

DOI: 10.1111/cts.12178

we compromise the utility of the findings from our research. Here are a few of the biases that can be expected if we overemphasize sample size in relation to other aspects of study design.

Sampling bias

A substantial portion of the US population remains uninsured and even a larger group uses healthcare rarely only. Although the trend is toward greater use of EHRs, only about 40% of patients currently have their information recorded in EHRs. Population scientists have offered countless examples of the consequences of sampling bias. For instance, the large Nurses Health Study followed 48,470 postmenopausal women, 30–63 years of age for 10 years (337,854 person-years). The study concluded that use of hormone replacement therapy cut the rate of serious coronary heart disease nearly in half.[6] Despite the enormous sample size, the study failed to recognize the atypical nature of the sample and the confounding of estrogen therapy use with other positive health habits. The clinical trial portion of the Women's Health Initiative (WHI) was able to control for self-selection and suggested that estrogen replacement did not lower the risk of heart disease and may actually be harmful.[7] Later, investigators were able to show that the results of the Nurses Health Study do map to the WHI if the focus is on new hormone replacement users.[8] Observational studies can provide valuable causal information, but only when the investigators have the right model. When the underlying sampling model is wrong, large sample size can magnify the bias.

Big Data and the Future of Health Services Research

Big data are likely to revolutionize health services research.[9] But, despite the advantages of EHRs, the actual rollout in the United States is creating its own set of biases. For example, extensive customization of EHRs now makes it very difficult to combine data across systems. There are also serious problems with intraoperability, or compartmentalization of data across systems that often do not communicate with each other.[10] Patients often change providers and may move from one EHR or insurance claims system to another, and as a result their trajectories are hard to follow. Changes in coding conventions and the frequent conversion of numerical data to text may result in information that cannot be easily tracked over time. Merging data can be challenging, particularly when the observations are stored in different proprietary systems that cannot be easily accessed.

Overhage and Overhage describe EHR data as a "cartoon" of the patient they are derived from. Claims forms are often incomplete and known to contain substantial errors. Some of these errors are intentional. For example upcoding or use of specific diagnostic codes is a common source of systematic error. Claims data often have missing observations, or a "Swiss cheese" quality.[5] In addition to inaccuracy, observations from claims are often transformed in ways that obscure their original meaning.[10] For example, a patient may be coded as having hypertension based on blood pressure readings or, alternatively, because she takes a medication commonly prescribed for high blood pressure.[5] These problems may be more likely in big studies in which the investigators are not involved in the original data collection.

Ascertainment bias

Even for those captured by EHRs, there are serious problems with the sampling of health system encounters because many people see a physician as a result of a health event. Event driven visitation creates a serious bias toward the overestimation of illness and disability. A bias in the other direction can occur when the most health conscious people visit their providers more often than less health conscious peers in the community. In planned prospective research, follow-up is typically achieved on a fixed interval. Sometimes these follow-ups occur on a random basis. The analysis strategies use either fixed effects designed for assessment at defined intervals, or random effects models that assume follow is on a random time schedule. However, event driven follow-ups violate the assumptions of both the fixed effects and the random effects models because follow-up schedules are neither on a planned nor a random schedule.

Big Data and the Study of Health Determinants

As with any type of data, the data in EHRs are only as good as the information captured by the systems. For example, several analyses have concluded that more than half of the variance in health outcomes is driven by behavioral, environmental, and social factors while medical care explains only about 10%.[11] Yet, social and behavioral factors are rarely recorded in the EHR. For example, regular consumption of fruits and vegetables is considered central to cancer prevention, but almost never recorded in the EHR.[12] When behavioral and environmental data are recorded they are rarely captured in a uniform way.[13] As a result, there is a tendency to estimate determinants of health outcome based on the variables that are included in the record. To use the tired analogy, investigators are "looking for the keys under the lamp post." Another problem is that missing data in EHRs are the rule rather than the exception. Especially problematic is that when data are missing, they are rarely missing at random. Even in very large studies, conclusions can be inaccurate because the most important variables are either not measured or are measured in different ways, with error, or are not included in the analysis model.

Measurement error

For any measured variable, the difference between the true score and the observed score results from measurement error. This error is common in all biological and behavioral sciences, but it can be controlled through systematic measure development. Small controlled studies often devote great attention to measurement precision. In contrast, users of big datasets may be unaware of how study variables were measured and low reliability of some measures can be expected. This can be a serious problem because it reduces the chances of finding significant relationships among measures. Thus, estimates of correlations and effect sizes are attenuated by measurement error. Large sample size may help reduce this bias, but if the measures are of very low reliability, the analysis will be focused on random variation. For example suppose two variables in a big dataset have reliability coefficients of 0.050 and 0.30. Even if these two variables were perfectly correlated in the real world ($r = 1.0$), the maximum expected observed correlation would be only 0.38. A modest correlation would be unlikely to be detected, even in a very large datasets.

Internal versus External Validity: To Whom and What Do We Wish to Generalize?

One of the most important challenges is obtaining data in circumstances that are representative of settings in which healthcare is practiced. Green and colleagues described one of the major problems in the generalization from randomized clinical

trials.[14] Considering a population of 1,000 people, about 800 experience symptoms over an interval of 1 month. Among these an estimated 327 consider seeking medical care. About 217 will get care in a physician's office. Even among those seeking physician care, only about 113 will get care from a primary care provider. About 65 of the 1,000 will go to an alternative care provider. Only 21 of the 1,000 will visit a hospital or an outpatient clinic and 13 will go to an emergency department. Among the entire population only eight will be hospitalized and less than one will be hospitalized in an academic medical center.

Even though only one in each thousand are hospitalized in an academic medical center, much of the research published in our major medical journals is based on referrals and data from patients who received care in academic settings. This raises serious questions about the generalizability of many of our large randomized trials. Even when recruiting through practice-based research networks, we only get access to less than 1/3 of the members of the general population. Most "Big Data" approaches that depend on EHRs only get data from the 217 patients who get to a physician's office. Even with big data, the focus is on the 21% who would be available for analysis. And if the 21% are a skewed sample, efforts to generalize will be compromised.

Big Data and the Randomized Clinical Trial

Randomized clinical trials do an excellent job of reducing biases internal to studies. However, randomization does little to assure that the results of the trial generalize to populations that may differ from the trial participants. It is common for therapeutic agents to look promising in phase 2 clinical trials, but to fail in larger phase 3 trials that are based on more representative study populations. Large samples in phase 2 trials may not address the problem, particularly if study participants are not selected at random from or representative of the population to which the results will be applied. Although cluster randomized trials help address this issue, they come with another set of limitations. Observations from a variety of disciplines suggest that the effects of interventions systematically decrease over time.[11] In addition, the effects of interventions are expected to decrease as populations become more heterogeneous and complex, a phenomenon referred to as "voltage drop."[15]

Multiple comparisons bias

Replication of study results has become a challenging problem for contemporary science.[16] Fraud is rarely the explanation for the nonreproducibility of results. One of the most important factors is that investigators have too many degrees of freedom when analyzing outcomes. For example, if 100 outcome variables are available for analysis, we would expect significant statistical effects for about five variables by chance alone. When big datasets include thousands of outcome variables and modern software is capable of examining all of these potential outcomes, the urge to attend to those that are statistically significant becomes irresistible.

The multiple comparisons problem in analysis of big datasets emerges in many different areas of biomedical science. The term "voodoo" has been used to describe the relationships between emotions and activity captured by functional magnetic resonance imaging (fMRI).[17] Activity in fMRI might take advantage of thousands of potential response patterns and more than half of the published papers fail to make a meaningful adjustment for this problem.[17] An independent review of 100 brain imaging papers in the top five journals similarly found that 40% of the papers did not correctly adjust for multiple comparisons.[18] In studies of tumor biology, the title of a recent paper tells the whole story: "Almost all articles on cancer prognostics markers report statistically significant results."[19] Incorrect conclusions from studies on tumor biomarkers result not only from multiple comparisons problems, but also because "predictions" actually come from retrospective rather than prospective observations.[20] Some high quality EHR studies do adjust for multiple comparisons.[21] However, unlike clinical trials, primary outcome variables are rarely declared prospectively in observational studies. Prospective declaration of primary outcome variables eliminates the possibility of selecting a significant outcome from among many variables. One recent analysis from the Amgen Pharmaceutical company suggested that promising preclinical trial results were replicated in only 11% of the tests.[22] One of the reasons preclinical studies are so difficult to repeat is that the replications focus on a single outcome. However, unreliability of information can be found in the most basic lines of research. For example, mammalian cell lines are often cross-contaminated. An investigator may purchase mouse cells but later learn that they are using guinea pig cells. Investigators studying ovarian cancer cells have later learned that they were really using breast cancer cells. Investigations of major cell repositories suggest that 14-30% of cell lines are misidentified.[23] Increasing sample size is not the remedy for this basic specification problem.

Big data studies can identify significant but inconsequential effects

Imagine that two randomized clinical trials compared new treatments for pneumonia. Both trials produce statistically significant results at exactly $p = 0.05$. One trial was based on 150 patients randomized in one of the two groups. The second trial was based on 15,000 patients randomized into two groups. Which treatment should be preferred? Many people would prefer the treatment based on the larger trial. However, the effect size for the smaller trial would need to be significantly larger in order to achieve the same $p = 0.05$ significance level. At a constant p level, effect size declines as a function of sample size.[24] In other words, the number needed to treat would be considerably smaller for the treatment evaluated in the small trial. Although large trials have many advantages, marginally significant effects observed in large trials typically mean that the effect of the treatment is quite modest. At a constant exact p level, an individual patient would be more likely to benefit from a treatment evaluated in a small trial. Treatment effects identified in big datasets, although statistically significant, might be almost trivial at the individual level as statistical significance testing is designed for use in small rather than enormous datasets. *Figure 1* shows the effect size in two hypothetical trials, each with a difference of $p = 0.05$. The effect size of the small trial is much more meaningful for the individual patient.

Current practice in trials versus big data

Contemporary clinical trials are required to register in Clinicaltrials.gov. When trials are prospectively registered, the investigators are required to declare their primary and secondary outcome variables well in advance of data collection. As a result, they have fewer opportunities to report on other variables that turned out to support the value of their treatments. Analysts of big data are usually not restricted to just one or two variables and, as a result, their conclusions are often more conservative. On the other hand, there may be other unexpected consequences

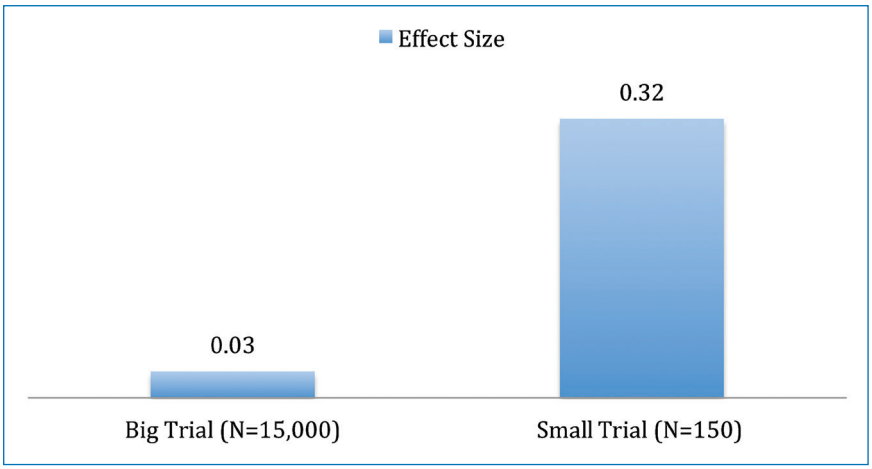

Figure 1. Effect sizes for two trials, both $p = 0.05$.

of using multiple outcome variables. For example, if studies are registered prospectively, statisticians are likely to request adjustments for multiple comparisons. If analysts do adjust for the hundreds of tests, it is likely that many potentially true effects will be overlooked.

Table 1 summarizes the potential biases, questions that should be asked when considering using large datasets, and some potential practices that might be used to minimize the bias. Despite the enthusiasm for basing a new epidemiologic science on "big data" repositories, we urge caution in uncritically accepting the interpretation of these analyses. Our statistical methodologies were developed for inferences about small datasets and are often irrelevant or inappropriate for the analysis of big data. Further, there are many occasions in which carefully done small studies can provide a more reliable answer at a lower cost. Newly emerging big datasets offer promise to answer questions not previously possible. However, we urge users, purchasers, and researchers to be aware of the potential biases outlined above, to ask questions and use procedures such as those in *Table 1*, and to work to enhance the quality of datasets, analyses, and conclusions.

The Potential of Big Data

We do not want to leave the reader with the impression that big data has no place in biomedical research. Investigators will come to master the biases and new methodologies are likely to revolutionize several different areas of research. For example identification of rare medication side effects or observation of outcomes among people with rare diseases will benefit greatly from big data methodologies. We only caution that information can mislead when not acquired systematically.

Conclusion

In biomedical research, as in other endeavors, we often assume that bigger is better. There are many circumstances in which very large studies include systematic biases or have large amounts of missing information, and even missing key variables. Large sample size does not overcome these problems: in fact, large sample studies can magnify biases resulting from other study design problems. Unlike the public opinion poll for the 1936 presidential election, it often is difficult to validate the results of many biomedical studies with clearly observed public events. We suggest that, in the near term, big data and large sample size alone is unlikely to improve the validity of most studies in epidemiology, health services research, determinants of health, or clinical trials.

We often assume that large sample size is the remedy to inferential bias.[25] For many inferences, representativeness is more important than sample size. Bigger is not necessarily better. The

Potential biases	Questions to ask	Potential practices to minimize problem
Sampling bias	Is the study sample representative of the population the results will be generalized to?	Whenever possible randomly sample from the population; if not, ensure diverse sample and include those at risk.
Ascertainment bias	Were individuals entered into the study because of a health event?	Review biases associated with case–control studies; see below.
Retrospective bias	Did the study begin with an established event and work backwards?	Use population denominator rather than clinical practice denominator when estimating prevalence.
Bias and lack of scope in the information recorded	Did the electronic health record include the full range of health determinants e.g., patient reported information, contextual and environmental factors?	Whenever possible, work to expand the comprehensiveness of information in the EHR.
Measurement error	Was there an assessment of the reliability of key outcome measures?	Estimate the reliability of key measures and determine how much attenuation would be expected given the imperfection of the measurement instruments.
Multiple comparisons bias	How many outcome measures were used? How many analyses were conducted?	The primary outcomes should be specified in advance of the study. If multiple outcome measures are used, statistical corrections should be applied.
Effect size	Was the effect size reported? If so, was the reported effect clinically meaningful?	Effect sizes should be calculated and reported whenever possible.
Study registration	Was the study registered in a service such as clinicaltrials.gov?	Studies that are registered with prespecification of outcome variables should be assumed to have a higher level of credibility.

Table 1. Potential strategies before addressing bias when analyzing big data.

examples used in this commentary are based on large versus small sample sizes in contemporary research investigation. These examples represent the small data era. The problems we describe are likely to be much more severe in the new era of big data when investigators are more likely to be separated from the data collection process. Clearly the analysis of large databases will be an important part of our future, especially if combined with other data sources, with purposeful attention to the issues above, and with appropriate sensitivity analyses to address potential biases.

References

1. Freedman D, Pisani R, Purves R. *Statistics.* 4th edn. New York: WW Norton; 2004.

2. Lauer MS. Time for a creative transformation of epidemiology in the United States. *JAMA.* 2012; 308: 1804–1805.

3. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol.* 2005; 58: 323–337.

4. Selby JV, Krumholz HM, Kuntz RE, Collins FS. Network news: powering clinical research. *Sci Transl Med* 2013; 5: 182fs13.

5. Overhage JM, Overhage LM. Sensible use of observational clinical data. *Stat Methods Med Res.* 2013; 22: 7–13.

6. Stampfer MJ, Colditz GA, Willett WC, Manson JE, Rosner B, Speizer FE, Hennekens CH. Postmenopausal estrogen therapy and cardiovascular disease. Ten-year follow-up from the nurses' health study. *N Engl J Med.* 1991; 325: 756–762.

7. Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, Barad D, Curb JD, Kotchen J, Kuller L et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *Am J Epidemiol*. 2005; 162: 404–414.

8. Grodstein F, Manson JE, Stampfer MJ. Postmenopausal hormone use and secondary prevention of coronary events in the nurses' health study. a prospective, observational study. *Ann Intern Med.* 2001; 135: 1–8.

9. Lagu T, Krumholz HM, Dharmarajan K, Partovian C, Kim N, Mody PS, Li SX, Strait KM, Lindenauer PK. Spending more, doing more, or both? An alternative method for quantifying utilization during hospitalizations. *J Hosp Med.* 2013; 8: 373–379.

10. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care.* 2013; 51: S30–S37.

11. Schroeder SA. Shattuck Lecture. We can do better–improving the health of the American people. *N Engl J Med.* 2007; 357: 1221–1228.

12. Council NR. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. Washington, DC: The National Academies Press; 2014.

13. Estabrooks PA, Boyle M, Emmons KM, Glasgow RE, Hesse BW, Kaplan RM, Krist AH, Moser RP, Taylor MV. Harmonized patient-reported data elements in the electronic health record: supporting meaningful use by primary care action on health behaviors and key psychosocial factors. *J Am Med Inform Assoc.* 2012; 19: 575–582.

14. Green LA, Fryer GE Jr., Yawn BP, Lanier D, Dovey SM. The ecology of medical care revisited. *N Engl J Med.* 2001; 344: 2021–2025.

15. Kilbourne AM, Neumann MS, Pincus HA, Bauer MS, Stall R. Implementing evidence-based interventions in health care: application of the replicating effective programs framework. *Implement Sci.* 2007; 2: 42.

16. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet.* 2001; 29: 306–309.

17. Vul E, Pashler H. Voodoo and circularity errors. *Neuroimage.* 2012; 62: 945–948.

18. Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci.* 2009; 12: 535–540.

19. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer.* 2007; 43: 2559–2579.

20. Vermorken J. Annals of Oncology: a statement of editorial intent. *Ann Oncol.* 2012; 23: 1931–1932.

21. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Crane PK, Pathak J, Chute CG, Bielinski SJ et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med.* 2011; 3: 79re1.

22. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature.* 2012; 483: 531–533.

23. Masters JR. Cell-line authentication: end the scandal of false cell lines. *Nature.* 2012; 492: 186.

24. Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. Effect sizes and *p* values: what should be reported and what should be replicated? *Psychophysiology.* 1996; 33: 175–183.

25. Lauer MS. From hot hands to declining effects: the risks of small numbers. *J Am Coll Cardiol.* 2012; 60: 72–74.