Discussion of Keiding

ROGER D. PENG

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
rpeng@jhsph.edu

Reproducible research is an increasingly important topic in many areas of inquiry and Professor Keiding’s commentary has raised a number of important issues regarding the role of reproducibility in science, biostatistics, and this journal. In general, I find it difficult to disagree with many of his major points. However, as the author of the editorial on which Professor Keiding has commented, I would like to provide some context to the editorial itself and perhaps shed some light on the role of a reproducibility policy at a journal like Biostatistics.

An important benefit to using reproducibility as a benchmark is that it is a clear and concrete standard that points to specific required materials. One can usually identify (at least in theory) the elements needed to reconstruct a specific numeric or graphical result. In an ideal world, we might ask for “everything” involved in a scientific investigation (including analyses attempted but not completed or published) but such a vague requirement would be impractical. The push for reproducibility is in part a reflection of the fact that various constraints on publishing simply prevent important details from being disseminated. It makes sense to publish those details in the lingua franca of the profession, which for biostatistics is increasingly the language of computer code.

It should be emphasized that reproducibility is a minimum standard in the context of validating scientific findings. Clearly, lack of reproducibility using the same data and methods would call into question, the validity of the conclusions drawn from those data. Professor Keiding seems to think that meeting such a minimum standard is trivial and therefore useless as a serious check on scientific validity. In many cases, this is no doubt true. However, I think it is important to point out that reproduction of a particular result is merely the end point; the data and computer code that leads to the result is of primary interest and making this available to others is often far from trivial (if even possible). Put a different way, it is perhaps uninteresting for me to claim that I have made a great pie, but it is a different story if I provide the precise ingredients and instructions for making that same pie. We might disagree about which ingredients or techniques should have been used, but the availability of the recipe provides a meaningful basis for discussion and analysis. Ultimately, demanding reproducibility is an example of the means justifying the ends.

I agree with Professor Keiding that our profession has a problematic tradition of analyzing certain data sets “to death” in order to demonstrate new methodology. Furthermore, the danger that a particular individual might obtain a data set and use it without making the proper effort to understand the scientific context is real. However, I do not think journals should try to police such behavior by not encouraging authors to publish data and code online with their articles. Indeed, what is to stop an unscrupulous individual from completely misinterpreting the article itself? Should a journal also refrain from publishing articles to prevent such behavior? Obviously not because the benefit to science of publishing the article far outweighs the cost of not making that information available. I believe a similar cost-benefit relationship exists with the publishing of data and software code. While there will always be the potential for individuals to use the data and code improperly, the benefit to those who do understand the scientific
context (and through them, the broader scientific community) can be tremendous. This leads me to my final point.

Professor Keiding seems to assume that statisticians will generally be ignorant of the scientific context surrounding a particular data analysis and that there is a critical need to collaborate with substantive scientists to obtain the full picture. I do not deny that close collaboration with substantive scientists is critical for all applied statisticians. However, I would argue that many statisticians today, including many that have published in the pages of Biostatistics, would be considered leaders in their substantive application areas as well as their statistical areas. The movement of statisticians between the statistical and substantive domains is increasingly common and one can observe this by noting the number of publications lead by statisticians in substantive journals. I am optimistic that the interdisciplinary nature of statistical work and the need for statisticians to be intimately familiar with scientific details results in an increasing number of statisticians who are in fact quite knowledgeable of the substantive context of the data they encounter. If Biostatistics can promote such interdisciplinary work and understanding through reproducible analyses, then I consider that an important contribution.