Progress in Cardiovascular Diseases 55 (2012) 14–24
www.onlinepcd.com

Why Are Some Studies of Cardiovascular Markers Unreliable? The Role of Measurement Variability and What an Aspiring Clinician Scientist Can Do Before It Is Too Late

Matthew Shun-Shin*, Darrel P. Francis

International Centre for Circulatory Health, National Heart and Lung Institute, Imperial College London, United Kingdom

Abstract Cardiology research suffers from the scourge of unreliable results, despite honest conduct. Investigators' prior belief, compromised blinding, and scope for measurement variability are a fatally synergistic combination.

Can we stop these threats ruining the results?

First, clinical researchers must realize that healthy clinical practice (including intelligently integrating all available information) may be catastrophic to research.

Second, experienced clinicians know that variability may necessitate remeasurement to obtain a clinically correct result but must learn that doing so in research can cause surprisingly severe distortions of correlations or differences between groups.

For example, a "best-of-four" approach in comparing two 50-patient groups that are in reality identical, with a variable whose intraclass correlation is 0.8, easily generates highly significant *P* values.

Clinicians may be habituated to poorly reproducible clinical measurements and falsely reassured by their effectiveness for group mean effects in blinded randomized controlled trials.

We need a more critical approach to clinical tests if we care about evaluating individual patients reliably or want our research to be reliable.

Simple steps shown here, addressed during study design, will increase the reliability of research—if considered by researchers or the juniors whom they nurture. (Prog Cardiovasc Dis 2012; 55:14-24)

Keywords: Measurement variability; Bias; Blinding; Observational studies; Test-retest

> "In general, the performance of biomarkers is seldom as good in a second sample as in the sample in which they were initially assessed."[1]

Studies other than formal randomized trials with blinded assessment are well known to overestimate effect sizes[2] and even occasionally point in the wrong direction.[3] As the methodological quality of the trial decreases, the effect size tends to increase.[4] The basic sciences are not immune from this bias.[5]

A recent example comes from a meta-analysis of studies investigating the correlation between new imaging biomarkers of the mechanical dyssynchrony of ventricular contraction and the response to cardiac resynchronization therapy.[6] The observational studies reported values of the coefficient of determination (R^2) up to 10- or 20-fold higher than the externally monitored randomized controlled trials (RCTs). Further analysis revealed a progressive decline in the range of R^2 values, from reaching 0.8 in studies reporting

Statement of Conflict of Interest: see page 23.

* Address reprint requests to Matthew Shun Shin, BM, BCh, MRCP, International Centre for Circulatory Health, National Heart and Lung Institute, Imperial College London, 59-61 North Wharf Road, W2 1LA London, United Kingdom.

E-mail address: m@shun-shin.com (M. Shun-Shin).

0033-0620/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.pcad.2012.05.006

Abbreviations and Acronyms

BNP = B-type natriuretic peptide

ICC = intraclass correlation coefficient

RCT = randomized controlled trial

no blinding nor formal enrollment, to the 0 to 0.1 range in large studies that reported full blinding and formal enrollment.

Ioannidis and Panagiotou[7] have demonstrated a similar phenomenon in blood biomarkers across the specialties and specifically within cardiology.[2] Taking the use of C-reactive protein and Lp(a) lipoprotein in cardiology as an example, they note "If one considered only data from randomised controlled trials, probably neither…would be considered good biomarkers, whereas data from observational studies suggest the opposite."

Why is this effect occurring? Publication bias only provides part of the explanation. Individual studies made susceptible by compromised blinding are systematically contaminated by bias.

Blinding is often compromised…

Even large RCTs are susceptible to failures of blinding and randomization, which can be subverted by clinicians with strong prior belief acting in what they consider to be the patients' best interests. For example, the Captopril Prevention Project trial of angiotensin-converting enzyme inhibition was rendered uninterpretable because some investigators probably ($P = 1 \times 10^{-8}$) "peeked" into the randomization envelopes to help patients with the highest blood pressure not to be randomized to the placebo arm.[8,9]

Blinding is essential but requires additional effort to generate a data set that is independent of data acquired in normal clinical practice, which might directly or indirectly reveal the other variable to the researcher. In some cases, complete blinding may be practically unobtainable. For example, it may not be possible to hide the presence of a pacemaker lead or electrocardiographic spikes from an echocardiographer making a measurement of ventricular response to pacing.

…And bias is everywhere

For many researchers, the reason to conduct a study is to obtain a "positive result," that is, to confirm a suspected association or effect. One should not trust oneself to be unbiased merely because one is generally honest.

Even Nobel prize winners are not immune from bias. Millikan while determining the charge on a single electron selectively reported results from oil droplets that were consistent, biasing the estimate of the error with his technique to be smaller than it truly was so that the confidence interval on the estimate failed to contain the true value.[10]

Bias is everywhere, Sackett[11] categorized 35 different ways in which bias can contaminate analytical research, illustrated with ample examples from the literature.

For example, it may arise from the method of finding patients. If a study of mortality of aortic stenosis identified patients only from postmortem, it would tend to overestimate the mortality rate in a general population of aortic stenosis.[12]

It may also arise from the time point in the course of disease at which patients are selected. If a study examined the effect of an intervention on a biomarker, which showed some variability over time, but enrolled only patients with a high initial value (and had no control group), there is a tendency for a subsequent values to be lower, even if the intervention was ineffective, as the original high value may represent an "outlier" result and further reading are statistically more likely to be closer to the (lower) true underlying value. This pervasive effect is known as regression toward the mean.[13,14] This may be why so many ineffective remedies are incorrectly believed to be effective by members of the general public who use them only when they have a symptom: they are not dishonest but have merely not considered the biasing effects of their pattern of use.

Preferential enrollment of enthusiastic patients, especially if the intervention is considered sufficiently risky or unpleasant that many refuse it, can also introduce bias into an uncontrolled study.[15,16]

Measurement variability and expectation bias

One type of bias, termed *expectation bias*, arises from a strong clinical belief in a relationship among staff that make measurements, in the context of more than 1 possible, legitimate value being obtainable.

Because of the natural variability in many of our biomarkers, clinicians often choose the "most appropriate" of several potential values to represent the patient. Sackett[11] illustrates the occurrence of this phenomenon in simulated clinical obstetric practice when physicians misreported high fetal heart rates as being closer to the norm than automated measures do.[17]

However, if conducting a research project into the existence of a difference between groups or a relationship between variables and they begin with a positive belief, then this habit can become a destructive self-fulfilling prophecy.

Even requirements to average a few readings, for example, 3 beats, will not eliminate this, as the clinical researcher will still have a choice of which run of 3 to average.

This phenomenon is insidious, as it is not only legitimate but also obligatory within clinical practice; one must select a single reading to represent the supposed