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

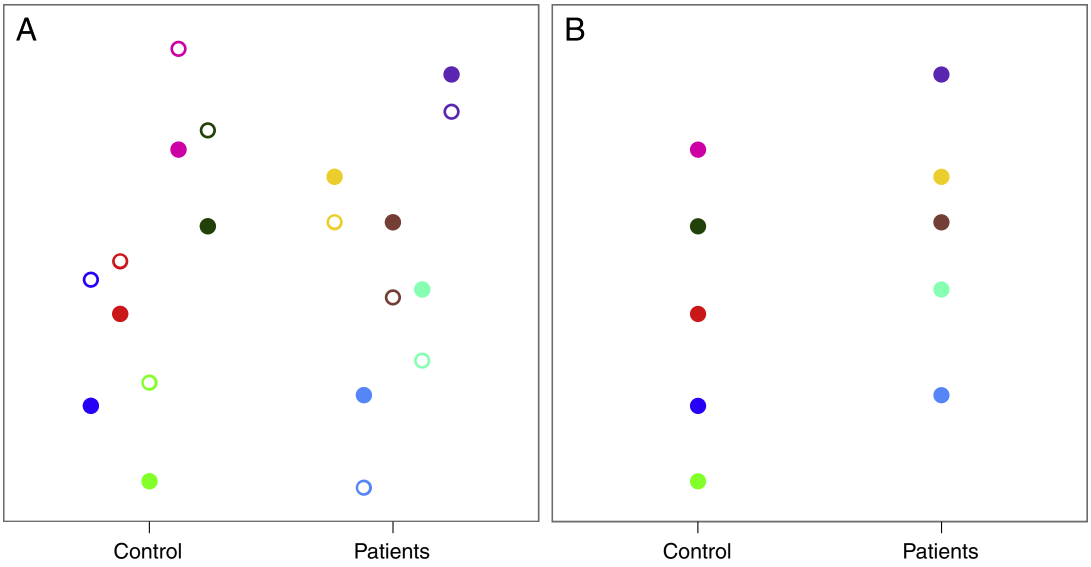

Fig 1. How choosing between more than 1 value for a variable causes a biased, unblinded observer to create a false difference between 2 groups. In reality, there is no difference between patients and the control group in the biomarker. However, if the experimental setup allows the researcher to make more than 1 measurement of the variable in each patient and select the one that most fits with his hypothesis, a difference between the groups can readily be generated. Panel A, The researcher has made 2 measurements of the biomarker in 5 different patients (colored circles) in each group. There is no apparent difference between the patients and the controls. Panel B, However, if the researcher selects the smallest reading in the controls and the largest in the patients (closed circles), an apparent difference between the groups begins to develop.

true value. It is especially dangerous within cardiology because some measurements can be made every heartbeat giving a vast choice of possible readings; moreover, some measurements may have a significant analytical variability, giving a wide selection of values.

Studies conducted largely by a sole researcher may also be particularly susceptible as the individual who designed the hypothesis, obtained the funding, carried out the intervention, and is benevolently mindful of the impact of disappointing findings on the field as a whole is often the one who makes the measurements.

Given the existence of this bias and the susceptibility of cardiology research to it, the remaining question is how powerful can this effect be?

## How measurement variability creates unreliable results

In Figs 1 to 4, we demonstrate that not only is the impact of this form of bias greatest when measurement variability is large and when there are numerous potential choices but also that it is surprisingly (and disturbingly) powerful even when the reproducibility of measurement is high (intraclass correlation coefficient [ICC] of 0.9) and when the researcher has only 2 readings to choose between.

In Fig 1, we show the mechanism by which an unblinded researcher who has an honest (but incorrect) belief that values of a measurement are different can create an artefactual difference, if there is some variability in the measurement and the researcher honestly documents the most plausible value for each subject. If the experimental setup permitted more than 1 measurement to be made on each patient and the most appropriate one selected, then the clinician who strongly believed in a difference between groups would tend to select the higher measurement for patients in the group which was expected to be higher and the lower measurement for the other patients.

Fig 2 shows the size of the effect on a simulated study comparing measurements in 2 groups of 50 patients across a range of ICCs and number of possible measurements. In reality, there is no difference between the populations from which these groups are drawn, and an unbiased study would, on average, find no difference. Yet, even for well-reproducible variable with an ICC of 0.8 (top row) and with a choice of only 2 values (left column), a false difference between the groups appears and easily meets criteria for statistical significance. When measurements with greater test-retest variability are used or the researcher has the ability to take more measurements, the effect becomes very much more statistically significant.

In Fig 3, we show the mechanism of how applying these habits can artefactually create a correlation between 2 inherently uncorrelated variables. The clinical habit of selecting plausible measurements, that is, those most consistent with the tested hypothesis, reduces the scatter on the plot and causes a false correlation to develop.

Fig 4 shows the effect of this mechanism in a group of 100 patients across a range of ICCs and number of

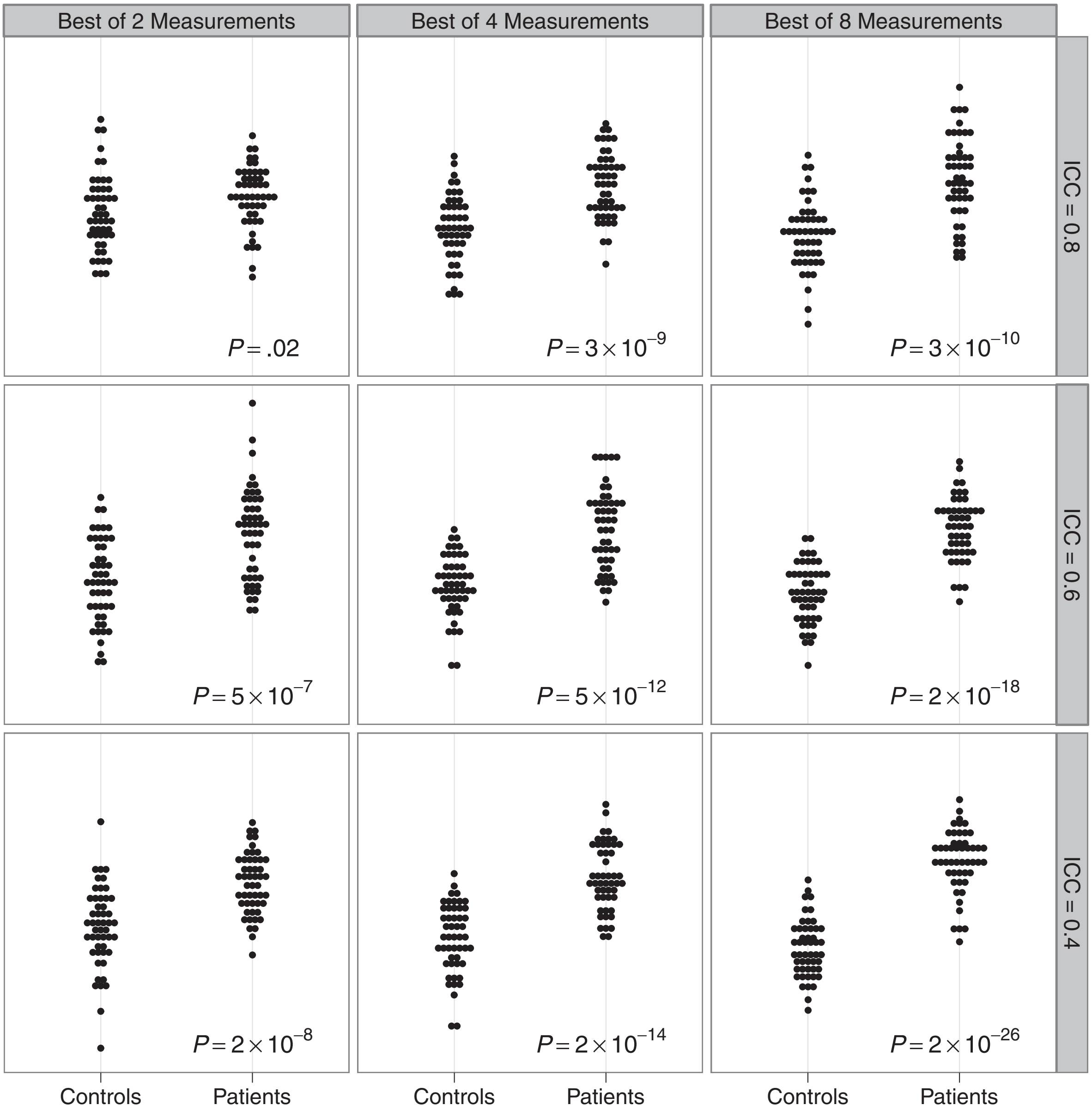

Fig 2. Effect the level of reproducibility and the number of choices of a measurement has on generating false and exaggerated differences between groups by an honest researcher who believes there is a difference between the two groups. If the mechanism seen in Fig 1 is continued to larger studies, statistically significant differences between 2 groups will be generated. Even if the measurement has good test-retest reproducibility with an ICC of 0.8 (top row) and only 2 measurements per patient are made (left column), the otherwise-negative study easily becomes falsely positive (top left panel) with $P < .05$ in a computer simulation with 50 patients per group. If the reproducibility of the variable is worse (middle and bottom row), the effect becomes stronger. Similarly, if there are more measurements to choose among (middle and right columns), again, the effect becomes stronger. For a poorly reproducible variable with many possible values, for a study of this size, one should expect an artefactual difference between groups that meets statistical significance criteria at $P < 2 \times 10^{-26}$.

possible measurements. In reality, there is no correlation, but with this bias, a persuasive graphical appearance of correlation readily develops. Tests of statistical significance rapidly become positive.

This extends our previous report[18] that correlations easily arise when researchers select among very irreproducible variables whose ICC is 0. In this article, we find that this effect can occur at any measurement with imperfect reproducibility (ICC <1). The worse the reproducibility and the more alternative values the researcher may choose between, the greater the opportunity for bias to distort the results. In principle, perfect test-retest reproducibility would protect against this.

No fabrication or falsification is required for these exaggerated or false results: they arise spontaneously when poor reproducibility combines with compromised blinding and a researcher's prior belief.

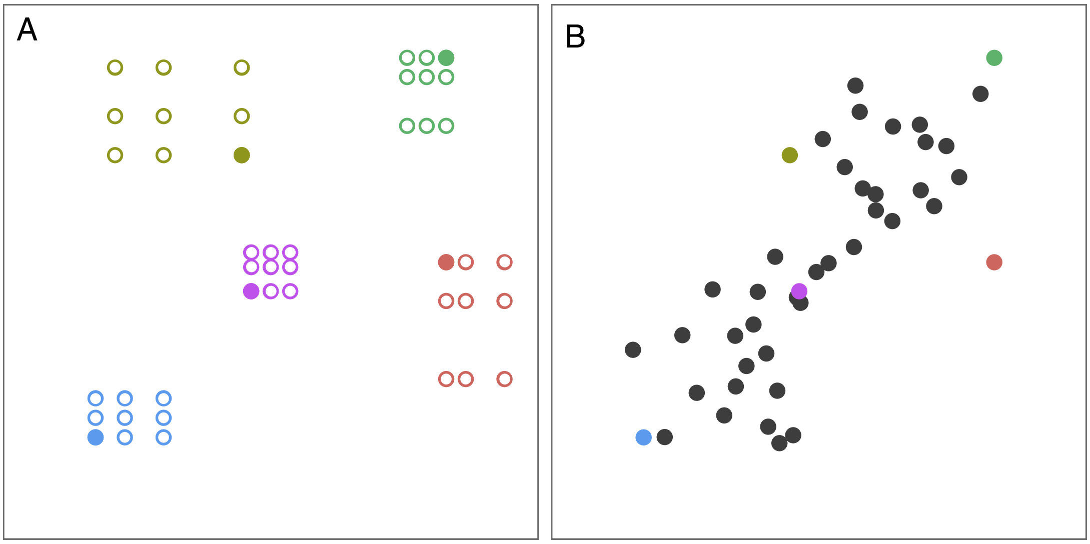

Fig 3. How choosing between more than 1 value for each variable causes an unblinded observer to create a false correlation. In reality, there no is correlation between the biomarker (x-axis) and the outcome measure (y-axis) in patients. If the experimental setup does not force the researcher to make only 1 measurement and the researcher is allowed to pick the combination of measurements that most supports a positive correlation, false correlations can readily be generated between uncorrelated variables. Panel A, The researcher has made 3 measurements of both variables (the biomarker and the outcome) in 5 different patients (colored circles). The 2 variables appear uncorrelated. However, if the researcher picks the measurements that are most consistent with a positive relationship between the 2 variables (closed circles), a correlation begins to develop. Panel B, The correlation becomes more apparent as more patients are recruited—the original 5 patients (colored circles), in the context of the following 40 in which the same practice has continued has generated a false correlation. The effect will become stronger as more measurements are allowed per patient and if less reproducible measures are used. It can be detected using the Enron bite test.[18]

## Measurement variability

If false correlations are so easily generated from techniques with poor reproducibility, researchers must take great care to establish reliably that techniques are reproducible. Perhaps clinicians might also demand such data. Without a quantitative indication of reliability, measurement results cannot be compared, either against reference values or between one time point and another, for example, in serial follow-up of disease progression or response to an intervention.[19]

Both clinicians and researchers will be disappointed by current practice and evidence. Rarely, if ever, is the individual patient uncertainty associated with a measurement quoted during clinical practice. Worryingly, for example, echocardiography guidelines do not habitually report neutral, blinded test-retest reproducibility of imaging biomarkers that they recommend, and clinicians seem unwilling to inquire why. Sometimes guidelines suggest steps to improve reproducibility but do not indicate what precision should be expected to be achieved.

## What do we need to know about measurement variability in cardiology?

The natural variability of blood pressure and the hence inability to reliably compare successive single readings are well appreciated. Even 3 readings within a visit are now recognized to be insufficient. Consequently, ambulatory blood pressure monitoring has become the gold standard for diagnosis of hypertension and monitoring the response to therapy.[20]

Do other measurements in cardiology warrant the same treatment? In both imaging and biomarker research, it is becoming increasingly clear that the answer is yes. For example, B-type natriuretic peptide (BNP) has a coefficient of variation of 40% over a period of 1 week[21] in stable outpatients. This is higher than the reported value for blood pressure (around 9%), and yet current clinical practice has not moved beyond taking a single sample. In current practice, if more than 1 is carried out, the principal driver for this is often to determine if there has been a change, rather than with the intention of averaging the 2 values. Yet with spontaneous variability having an SD of 40%, almost all numerical changes seen in an individual are not worth clinical time poring over because they cannot be reliably distinguished from no information at all.

This decision to attempt to interpret single measurements may be pragmatic in cost to the hospital and inconvenience to the patient but does limit the prognostic value of the marker and its ability to reliably detect changes of a moderate size in serial follow-up of a patient. In scientific practice, use of a single value for BNP will result in a study having to recruit many more patients to reliably detect the same-sized effect. If a study is hoping to detect a correlation between BNP and another marker, it will have to be very much larger if it uses a single BNP measurement rather than if it uses an

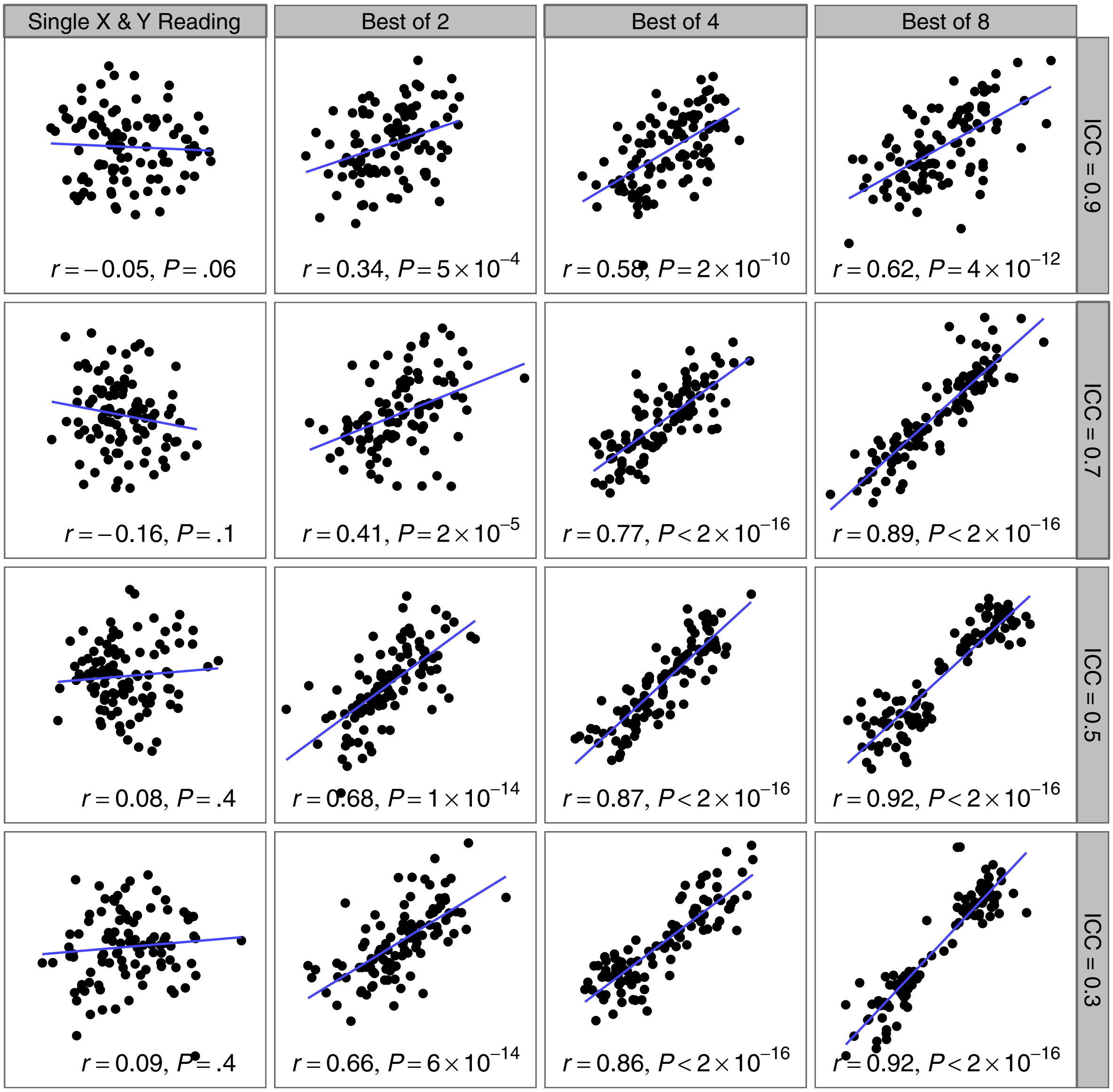

Fig 4. Effect the level of reproducibility and the number of choices of a measurement has on the generation on false and exaggerated correlations between 2 variables by honest researchers with biased beliefs. In reality, there is no underlying association between these 2 variables in the population ($r = 0$), but the simulated researcher measures each variable more than once and takes the "most appropriate" combination to represent that patient, that is, that which most concords with the researcher's belief of the underlying correlation. Even when only 2 measurements are taken of 2 highly reproducible variables (ICC, 0.9) in a 100-patient study, correlations of a significant magnitude are easily generated (top row, second column, $r = 0.03$). The effect becomes stronger as the number of possible measurements increases and the reproducibility of the variable falls.

appropriate average (such as a geometric mean) of a protocol-defined sequence of measurements.

### *Recognize the components—otherwise you may miss the most important one*

The variability of a measurement can be partitioned into its contributory components. We must not mistake interobserver or intraobserver remeasurement of an identical sample for test-retest reproducibility. In clinical measurements, variability over time is often much larger than that between observers reexamining one frozen moment in time repeatedly. If the study is of the chronic state of a patient, then it is essential that measurement of variability addresses this range of times.

For example, the variability of peak aortic velocity as assessed by echocardiography is well known to have a biologic component that varies over time, both from beat to beat reflecting variable filling and over longer periods reflecting different inotropic and volume states. It also has an analytical component that will consist of both intraoperator and interoperator (i.e., who does the scan, exactly where the probe is placed) and intrareader and interreader variability (i.e., who interprets the images, which may not be the same).

Biologic variability is just as present in blood biomarkers as with imaging biomarkers but is less thoroughly discussed.

### *Do not assume that well-established clinical protocols are automatically suitable for research*

Clinical guidelines may have a preference for limiting workload rather than delivering particular levels of reliability and may not be assumed to be a suitable basis for research design. For example, they may imply that it is possible and advisable to measure 1 best value for a variable, even when it has an inherently wide test-retest variability. In other cases, they recommend averaging across several beats, typically 3, but do not say where this number arises from nor on what basis they have decided that doing so produces a result with suitable reproducibility.

In general, averaging a few beats will scale down the between-beat variation by a factor of the square root of the number of beats; in the case of 3 beats, it would reduce it by approximately 40%. However, selecting beats from a single trace does not capture all the variability. The averaging process can only reduce the influence of the variability of the components captured within the sample. Those that are not captured (eg, probe position or operator) will not be reduced. Particularly counterproductive would be to wait for and capture a run of 3 beats that seem especially consistent.

It is advisable, but perhaps not immediately intuitive, that one should attempt to capture as much as possible of the variability that will occur between this visit and a follow-up visit. This means that it would be ideal to reposition the probe afresh and perhaps even the patient. It may seem strange to be doing this on purpose, but unless this is done, much of the benefit of averaging is lost. Going to such lengths has a cost. How much is worthwhile to do depends on what level of test-retest variability is desired and what level occurs without these special steps.

Even more surprising would be to make a series of measurements of a chemical biomarker on different days or weeks, with the intention to systematically use the average instead of just 1 value, for statistical analysis against other features. Nevertheless, doing so has twin advantages. First, it permits a study to be reliable with fewer patients. Second, it permits the detection of underlying relationships between variables that might be much stronger than they appear from single measurements that are heavily influenced by noise.

Knowing the test-retest variability of the biomarker is pivotal to this decision.

### *Assess test-retest variability in your own hands, rather than gambling on the literature being reliable*

True clinically relevant reproducibility requires conditions of “other days, other hands, and other eyes.” Unfortunately, such data are not often readily available. Often, in what is presented as reproducibility, only part of the variability has been captured.

One useful trick to obtain this information without having the bias of investigators understating variability is to look at data revealed as a by-product of randomized clinical trials. The SD of the change ($\delta$) from baseline to follow-up in the control-arm, in which no true change is expected (as can be demonstrated by comparing the means), will provide an estimate of test-retest reproducibility. Furthermore, the estimate is likely to be unbiased compared with direct reproducibility studies, as RCTs tend to have a higher methodological quality, and researchers will have had little motivation to deflate the variance of the change in the control arm.

### *Have realistic expectations*

Remember that the ceiling on the correlation coefficient that can be consistently observed, even between 2 variables, which are in principle perfectly related, is not 1 but lower. This is because the irreproducibility of each variable depresses the observed correlation coefficient. If the ICC of the variables is described as $ICC_x$ and $ICC_y$, respectively, then the relationship between the observed correlation coefficient ($r_{\text{observed}}$) and the theoretical underlying correlation coefficient ($r_{\text{underlying}}$) is as follows:

$$r_{\text{observed}} = r_{\text{underlying}} \times \sqrt{ICC_x \times ICC_y}$$

As ICCs are always less than 1, $r_{\text{observed}}$ is always less than $r_{\text{underlying}}$.

Consequently, readers of an article can calculate the implied underlying correlation coefficient from the correlation coefficient reported to have been observed by the authors, as follows:

$$r_{\text{underlying}} = \frac{r_{\text{observed}}}{\sqrt{ICC_x \times ICC_y}}$$

If $r_{\text{underlying}}$ is calculated to be greater than 1, either the $r_{\text{observed}}$ is exaggerated by chance or bias, or the study somehow achieved better test-retest variability than encapsulated in $ICC_x$ and $ICC_y$.[22]

## What researchers can do to prevent these problems

Trying to tackle this at the reporting stage (as, for example, was suggested in the STROBE guidelines[23]) is one approach, suitable for journals as they form an accessible pinch point at which to filter compromised studies. However, if studies have already been conducted with inbuilt poor design, authors may be defensive ("mistakes were made, but not by me"[24]), misunderstanding the suggestion of bias for a suggestion of impropriety, and redirect the article to a less rigorous journal. These issues must be tackled not only at the start of studies but also at the start of researchers' careers, preventing bad habits before they are formed.

## Advice for junior researchers

Do not allow the pressure to be "doing some research" to goad you into embarking a protocol without carefully thinking it through. Your work will not automatically escape inadvertent bias arising from a combination of prior belief, routine clinical habit of selectively reporting clinically consistent information, imperfect or absent blinding, and inherent measurement variability.

### *Recognize and reject these classic myths*

- "Minor bias can only have a minor effect" (see $P$ value of ~$10^{-26}$ in Fig 2 to be disabused of this).
- "A larger sample size protects against bias" (in fact, it makes the false-positive effect more statistically significant) (Fig 5).
- "Being honest protects me against bias" (see Figs 1 to 4 showing effect of honest belief).
- "Well-established clinical protocols are a safe bet" (in reality, they are rarely designed to deliver a specified level of precision, and they typically permit or encourage unblinded manipulation under the guise of taking the whole clinical context into account).
- "Good clinical practice is good research practice" (you are right to give clinical patients a coherent summary, but favoring coherence during research data collection produces false associations as shown in Figs 3 and 4)
- "My research is sound because the data was collected by others with nothing to gain from exaggeration" (but their beliefs may have colored what they recorded as the patients' values, Figs 1-4)
- "Randomisation protects my study against bias" (it protects only against biased allocation between therapeutic arms and not automatically against biased measurement).

### *Your study hinges on the measurements—scrutinize them mercilessly before you begin*

Believe nothing anyone says or writes about the measurement of the variable. They will not be sweating all night with you in a year or two when your study gives unreliable results. If you find your measurements have wider test-retest variability than the claimed "reproducibility," do not feel that you are a failure. Do not be cowed by assertions that you need "more training" unless evaluation from credible, independent sources shows that the training generates genuine narrow spread of blinded test-retest variability.

### *Say no to silly design*

You must have backbone to stand up to seniors who might exert pressure to do as you are told. Only you can prevent yourself wasting your research career on inherently doomed studies.

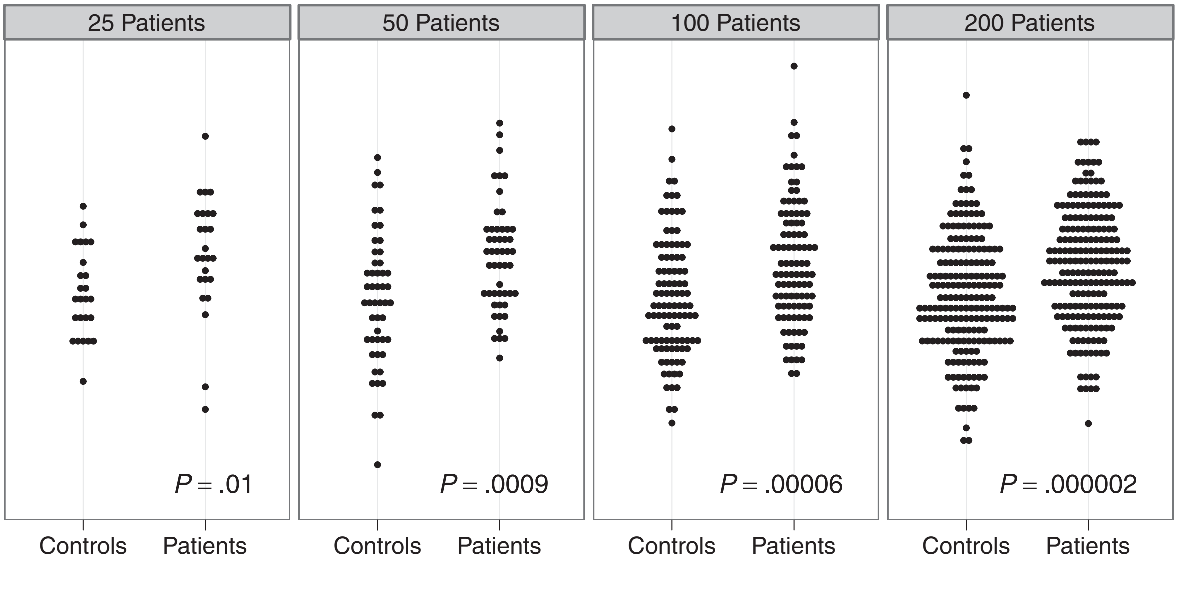

Fig 5. The effect of study size on the significance of a false difference between 2 groups generated by an honest researcher with prior beliefs who has the choice of more than 1 value for a variable. Study size is no protection against bias; indeed, it gives greater statistical significance for false differences. As the number of patients increases in an example run of the simulation (left to right), the statistical significance of the false difference generated between the 2 groups increases (ICC, 0.8; 2 possible measurements).

Ask the proposed boss what will happen if the study result is not as expected. If the answer is "Collect more patients" or "look at subgroups," it is a red flag that they have not actively planned the study to be reliable enough for the stated purpose.

You now already know that selecting between measurements to represent a patient, unblinded to their group and with your career on the line, produces worthless data (Fig 2). You have also seen that looking for confirmation of a correlation that your department strongly believes in can be equally meaningless (Fig 4).

Ask the awkward questions, and be ready to turn down the offer of your involvement if the study is not demonstrably well designed.

### *Challenge the gold standard*

Biomarkers for diagnosis typically rely implicitly on there being a gold standard. The possibility that this gold standard is incorrect, contentious, or just plain variable is rarely considered because even the thought might jeopardize the whole study whose funding is already in place. For example, if you are being encouraged to spend your life adding to the battery of markers for the severity of heart failure, ask about the extent of agreement between the local gold standard diagnostic test for heart failure and other potential gold standards, and if it is only partial, ask why the proposed biomarker is hoped to match this locally available gold standard rather than another one. Genuinely good gold standards withstand modest scepticism.

### *Enough vision, but no more*

If the proposed boss advises you that your role will be to "show that …," be careful. If the result of your scheduled years of research is already known to your boss, then either your work is redemonstrating what is already widely known or your boss has extraordinary insight.

### *Resist the isolated hunt for a correlation coefficient*

Almost every biomarker has some degree of dependency on every other, although the correlation coefficient for most pairs is modest. Setting out on years of work to confirm a universally expected association may not be a good use of your scarcest resources: time and concentration. Worse, if the relationship is so obvious that the only way to publish it is to report that the correlation is higher than expected, there is a risk of entering a game of competitive lying.[18]

### *Ask how your work will add usefully to useful knowledge*

Expect a well-developed answer. It need not change practice the day after you finish but should make a difference to what is done by someone, somewhere. The first requirement is that the result should be demonstrably unbiased. The second is that it should be designed to report sufficiently narrow confidence intervals to contribute usefully to understanding. The third is that it is a question that impacts on clinical practice or on the design of other experiments that are sensibly on a chain that affects clinical practice.

If you are studying a prognostic marker, beware that the market stalls are heaving with alternatives[25] and that customers are interested in cheapness, speed, and accessibility to the extent that less than 1% of all available biomarkers are used in practice. If your research culminates in proving that your marker predicts prognosis, it takes its place at the bottom of the pile. Even if it adds prognostic value over standard clinical information, beware that it might not be additive over the dozens of existing biomarkers.[26,27] Even if it does, with its predecessors not routinely used, it too might suffer the same fate. For many cardiovascular disease groups, there is a standard panel of therapies that are applied to almost all patients in the spectrum systematically, to improve prognosis. Only if it were plausibly believed that a drug may be helpful in one subgroup and yet harmful in another might biomarkers be crucial to that therapeutic decision.[28]

If your marker is diagnostic, beware that it is very unlikely that clinicians will agree to rely upon it for decision making. At best, then, it will be part of a panel; far more likely, it will fade away from routine practice. For example, detailed evaluation of subfractions of lipids held great interest in decades past, but this has faded because we have 1 dominant class of prognostically effective therapeutic intervention and so currently little appetite to expend effort in identifying and planning individualized attacks on these subfractions.[29]

### *Improve reproducibility by systematically averaging enough measurements per subject according to a prespecified protocol*

If the test-test reproducibility of your biomarker is poor, the study will be extremely vulnerable to bias (Figs 2 and 4). Even a traditionally good ICC of 0.8, which means that only 20% of the variance between single measurements in different patients is noise, provides enough flexibility to for the study to give false-positive results (Fig 2, top row).

Applying a protocol that insists on exactly $n$ measurements (with none discarded), which are then averaged, causes test-retest variance to fall by a factor of $n$, that is, SD to fall by $\sqrt{n}$.

This process brings the ICC approximately $n$ times closer to 1.

### *Be wary of studies involving one man and his biomarker*

When a junior fellow embarks upon a project, they often undertake all roles of the study. They may recruit the patients, perform the measurements, collect and digitize the data, and analyze the data. It is almost impossible for the sole researcher to remain blinded. Consequently, they are particularly susceptible to measurement variability, allowing their study to become biased.

### *Data collected for routine clinical purposes may not be unbiased*

Even using data already collected from routine clinical purposes, from clinicians who had no knowledge that a study may later be undertaken using the data, does not guarantee freedom from bias. The clinicians have no obligation to be unbiased in a scientific sense but do have an obligation to the best interests of the individual patient. For example, it is right and proper that a patient having undergone a hazardous intervention is given information in appropriate context of its psychologic effects on their well-being. A good physician may rightly report to a patient and then record in the notes that a marker that has numerically deteriorated within its range of variability is "essentially unchanged" and to the next patient whose marker has improved to the same extent, that it has "improved, and you should soon start to feel better!" Encouraging patients and avoiding discouraging them are key roles as clinicians. However, to start with information that was filtered in that way and hope to derive reliable research findings is unrealistic.

### *Implement aggressive debiasing into the protocol*

Imagine this important project will in fact have to be carried out not by you but by an associate whom you know to be a manipulative career climber. Moreover, it so happens that you will be judged not by the extremeness of your findings (as is usual in clinical research) but by whether your findings become reconfirmed when they are used as a basis for future work. If this is difficult to imagine, imagine instead that a dear relative has an early form of this condition and you are therefore even more determined that only correct answers come from this research—regardless of its "publishability."

Think how you would design the study to protect from the results from this coworker of dubious integrity. Perhaps you would insist on a neutral assistant to ensure blinding. You might ensure that there was feasible protocol for averaging multiple samples systematically acquired and measured. You might insist that predictor and outcome variables are separately measured. You might ask for data to be archived as collected, with no scope for subsequent removal.

The steps you devise, for this imaginary case of a researcher with a nefarious Midas touch, might be worth applying even if the study will now in fact be carried out by a paragon of honesty such as yourself.

## Conclusion

Research studies are easily wrecked by normal clinical practices. Researchers aware of the study hypothesis may have multiple opportunities to inadvertently make it self-fulfilling. The dangers are great when there is scope for remeasurement of variables that have any test-retest variability.

For study results to be reliable, obsessive bias resistance must be incorporated into all measurement processes. Researchers embarking on studies should not assume that honesty protects them from unreliable results, that a larger sample size protects against bias, or that minor bias has only minor effects. We are all only human.

## Acknowledgments

DPF is supported by the British Heart Foundation (FS/10/038) which kindly funds this article to be Open Access.

## Statement of Conflict of Interest

All authors declare that there are no conflicts of interest.

## References

1. Vasan RS: Biomarkers of cardiovascular disease: molecular basis and practical considerations. Circulation 2006;113:2335-2362.
2. Tzoulaki I, Siontis KCM, Ioannidis JPA: Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. BMJ 2011; 343:d6829.
3. Ioannidis JPA: Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218-228.
4. Rutjes AWS, Reitsma JB, Di Nisio M, et al: Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469-476.
5. Ransohoff DF: Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 2005;5:142-149.
6. Nijjer SS, Pabari PA, Stegemann B, et al: Limits of plausibility for predictors of response to cardiac resynchronisation therapy: systematic review and design steps for reliable research. JACC Cardiovasc Imaging 2012 [in press].
7. Ioannidis JPA, Panagiotou OA: Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. JAMA 2011;305: 2200-2210.
8. Hansson L, Lindholm LH, Niskanen L, et al: Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial. Lancet 1999;353:611-616.
9. Peto R: Failure of randomisation by "sealed" envelope. Lancet 1999;354:73.
10. Broad W, Broad WJ, Wade N: Betrayers of the truth. Simon & Schuster; 1983.
11. Sackett DL: Bias in analytic research. J Chronic Dis 1979;32:51-63.
12. Ross J, Braunwald E: Aortic stenosis. Circulation 1968;38(1 Suppl): 61-67.
13. Bland JM, Altman DG: Regression towards the mean. BMJ 1994;308:1499.
14. Herpin D, Demange J: Effect of regression to the mean in serial echocardiographic measurements of left ventricular mass.

Quantification and clinical implications. Am J Hypertens 1994;7(9 Pt 1):824-828.

15. Strauer B-E, Yousef M, Schannwell CM: The acute and long-term effects of intracoronary Stem cell Transplantation in 191 patients with chronic heARt failure: the STAR-heart study. Eur J Heart Fail 2010;12:721-729.
16. Cole G, Francis DP: Comparable or STAR-heartlingly different left ventricular ejection fraction at baseline? Eur J Heart Fail 2011;13:234.
17. Day E, Maddern L, Wood C: Auscultation of foetal heart rate: an assessment of its error and significance. Br Med J 1968;4:422-424.
18. Francis DP: How easily can omission of patients, or selection amongst poorly-reproducible measurements, create artificial correlations? Methods for detection and implications for observational research design in cardiology. Int J Cardiol 2012:1-12.
19. Working Group 1 of the Joint Committee for Guides in Metrology: Evaluation of measurement data—guide to the expression of uncertainty in measurement. Bureau International des Poids et Mesures; 2008.
20. National Institute for Health and Clinical Excellence 2011 Hypertension: CG127. London: National institue for Health and Clinical Excellence; 2011.
21. Bruins S, Fokkema MR, Römer JWP, et al: High intraindividual variation of B-type natriuretic peptide (BNP) and amino-terminal proBNP in patients with stable chronic heart failure. Clin Chem 2004;50:2052-2058.
22. Francis DP, Coats AJ, Gibson DG: How high can a correlation coefficient be? Effects of limited reproducibility of common cardiological measures. Int J Cardiol 1999;69:185-189.
23. Elm von E, Altman DG, Egger M, et al: Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335: 806-808.
24. Tavris C, Aronson E: Mistakes Were Made (But Not by Me). Orlando, FL: Houghton Mifflin Harcourt (HMH); 2008.
25. Ahmad T, Fiuzat M, Felker GM, et al: Novel biomarkers in chronic heart failure. Nat Rev Cardiol 2012:347-359.
26. Kattan MW: Evaluating a new marker’s predictive contribution. Clin Cancer Res 2004;10:822-824.
27. Greenland P, O’Malley PG: When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. Arch Intern Med 2005;165: 2454-2456.
28. Lip GYH, Nieuwlaat R, Pisters R, et al: Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010;137:263-272.
29. Ip S, Lichtenstein AH, Chung M, et al: Systematic review: association of low-density lipoprotein subfractions with cardiovascular outcomes. Ann Intern Med 2009;150:474-484.