Unintentional overestimation of an expected antihypertensive effect in drug and device trials: Mechanisms and solutions

James P. Howard a,⁎, Graham D. Cole a, Horst Sievert b, Deepak L. Bhatt c, Vasilios Papademetriou d, David E. Kandzari e, Justin E. Davies a, Darrel P. Francis a

a International Centre for Circulatory Health, National Heart and Lung Institute, Imperial College London, UK
b Cardiovascular Center Bethanien, Frankfurt, Germany
c Brigham and Women’s Hospital Heart and Vascular Center, Boston, MA 02115, United States
d Georgetown University Medical Center, Washington, DC, USA
e Piedmont Heart Institute, Atlanta, GA, USA

A R T I C L E   I N F O

Article history:
Received 28 November 2013
Accepted 28 December 2013
Available online 8 January 2014

Keywords:
Hypertension (C14.907.489)
Clinical trial (V03.200)
Bias (N05.715.350.150)
Research design (E05.581.500)

A B S T R A C T

In clinical practice we pay close attention to choosing an appropriate intervention for patients and performing it safely. We may put less thought into how to measure the effect without bias.

Clinical practice involving noisy values such as blood pressure requires intelligent processing to avoid confusing patients, but applying such discretion in clinical trials may inadvertently disrupt quantification of the benefit of the intervention, in unblinded trials.

In this article we explore two sources of bias, which label for convenience “big day bias” and “check once more bias”, which can lead to unintentional exaggeration of benefits from both drugs and device-based treatment for hypertension. We show in this article why, as denervation trials become increasingly bias resistant, the reported effect size may reduce. If inadvertent bias affects patients who are denervated in the same way as it affects those receiving antihypertensive drugs, the most meticulous denervation trials may show the effect size falling from around 30 mmHg to around 10 mmHg.

Some readers will doubt that these biases could be large enough to matter. We therefore include downloadable spreadsheets that any reader can use to explore how powerfully small biases affect the apparent effect sizes. The results may be surprising. A 10-mmHg reduction, without needing to adhere to an extra drug, would still substantially reduce events in the long term, but crucial to such reasoning is the reliable quantification of the blood pressure effect, free of bias.

© 2014 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Clinical medicine is a mixture of science and craft. Sadly a corner often cut in the scientific aspect of our education is inadequate consideration of measurement uncertainty and the consequences of its mishandling. Consequently doctors sometimes think this an esoteric and impractical backwater of statistics. However, for blood pressure measurement, we clinicians can instantly recognize our everyday clinical behaviours, indicating that for clinical cardiology this field is a simple and practical bedrock. Our international group presents here an analysis of how measurement uncertainty can cause unintentional exaggeration of the effect size of an intervention, wherever there is: 1). strong prior belief of a large effect; 2). biological noise in the measured variable; and 3). established clinical practice of selecting between multiple values based on clinical context. We use the example of blood pressure because it may next year be topical, but the concepts affect all clinical measurements where there is strong prior belief and a clinical habit of selection, even if innocent.

2. Two distinct sources of error

As clinicians we usually describe the performance of diagnostic tests against a “gold standard” test. We rarely consider that, when the same test is used serially in the same patient to quantify the size of effect from an intervention, there are two distinct and non-interchangeable reasons why a single measurement may not be representative. These reasons are called “error” in statistical terms, but this does not imply any mistake in the measurement. The errors can arise mainly from real fluctuation in pressure.

⁎ Corresponding author at: International Centre for Circulatory Health, National Heart and Lung Institute, Imperial College London, 59–61 North Wharf Road, London W2 1LA, UK. Tel.: +44 207 594 1093; fax: +44 208 012 6109.
E-mail address: jphoward@doctors.org.uk (J.P. Howard).

Available online 8 January 2014
Accepted 28 December 2013
Article history: Received 28 November 2013

© 2014 Elsevier Ireland Ltd. All rights reserved.
Random error, also called noise or scatter, cannot be predicted. It arises due to the measurement technique or from genuine biological variability. It is inherent in most biological variables and can be quantified in various ways such as by the standard deviation of replicate measurements. In the case of blood pressure, there is substantial beat-to-beat biological variability in pressure that is obvious in the catheter lab on invasive traces.

The size of this error generally expands with time between repeat measurements, since more processes can become interposed. Random error in blood pressure over a period of days is wider than that over a few seconds or minutes, since (for example) changes in volume status and medication concordance might become involved.

Consistent or systematic error, is the other type of error. This can arise from many sources, but if it tends to be in one direction preferentially, then its expected mean across all patients is not zero. This is bias. For a study quantifying the effect of an intervention, a great danger is allowing our routine clinical practice to influence our measurement technique. For many clinical variables, the degree of random error is large, and standard clinical practice has evolved to make several measurements until one that is satisfactory is obtained. This well-worn clinical habit, in combination with a strong prior expectation of what the measurement should be, can lead to innocent bias of the value in a particular direction. [1]

A particular form of systematic error is confounding. This will likely be present in both blinded and unblinded trials. Patients enrolled in trials of renal denervation, for example, may lose motivation to take their antihypertensive drugs through belief they have undergone a curative procedure. Conversely they may gain motivation, because the invasive procedure focuses their mind on the seriousness of hypertension. In unblinded trials this confounding may be unequal between trial arms, either masking or amplifying the true benefit of the intervention. In blinded trials both treated and placebo groups receive these effects equally, allowing us to quantify the true incremental benefit actually caused by the intervention, as it will be beyond the change seen in the control arm.

3. The statistical impact of increasing sample size is different between the two types of error

The two categories of error are worth distinguishing because they exhibit different and potentially opposite behaviors when we take the traditional pathway to making research more reliable: increasing the sample size. The increase can be achieved through either more subjects or more measurements per subject.

Random error is less able to disturb the study result (in statistical terms, the central estimate) if the sample size is increased, because it is likely that some random displacements are upwards and some downwards. The central limit theorem means the contribution of random error on the mean effect size shrinks in proportion to the square root of the sample size.

Random error is relevant to daily clinical practice: variation in office blood pressures is so large that some have argued that up to 40 blood pressure readings are necessary both before and after prescription of antihypertensive drugs, to reliably demonstrate a 5 mmHg reduction in blood pressure [2].

Bias, however, cannot be suppressed by enlarging a study. As sample size increases, just as a trial is more able to detect genuine biological effects, unfortunately it is also more able to falsely detect small biases as positive effects (Fig. 1 and interactive spreadsheet in Online Appendix 1). Statistical testing often focuses on whether effects are consistent with simple random error or not. If the deviation from zero is farther than commonly occurs with random error alone, e.g. when p < 0.05, this suggests that there is a non-random cause. But the test cannot separate a genuine effect from bias or confounding. The p-value is agnostic to the source of the effect.

As the study size is enlarged, the contribution of bias to the observed effect size does not become any closer to zero, and therefore the p value progressively becomes, misleadingly, more significant (Fig. 1).

4. Reasonable clinical practice cannot always prevent bias influencing research results

Physicians are trained to use all sources of information to ensure that the most clinically meaningful information is documented. This includes
rechecking values of a noisy variable if they are out of the expected range. [1]

A patient attending for review after commencing a full dose of a powerful antihypertensive drug is typically expected to show a lower blood pressure. If the measured pressure is higher, many physicians would be surprised and repeat the measurement. They do this because they believe that the first value is unrepresentative, even if the research protocol recommends against this practice.

In unblinded trials, the inequality, between arms, of the temptation to repeat measurements, occurs because a researcher has a priori knowledge of the expected result. Without this information an initial measurement might have been accepted. An ineffective intervention is performed, such as blowing on the dice (dotted line). On re-rolling, the dice revert to their usual average values which are significantly lower. Experiment B shows a more sophisticated design. The initially selected dice are re-rolled (blood pressures re-measured), hoping a true baseline will be obtained. It is, but unfortunately a second selection step takes place, again selecting a subset with an artificially high baseline, so an artificial drop to the final value still emerges. To eliminate big day bias requires an extra step shown in Experiment C. After enrolment, a separate baseline measurement must be made and the patient retained regardless of the value, i.e. with no selection at that stage. Most practicing clinicians consider this step an unnecessary waste of resources, but omission, with recycling of an entrance-criterion pressure as baseline, can cause effect sizes to be dramatically exaggerated.

**Fig. 2.** Big day bias, which statisticians call regression to the mean, can create a false conclusion that an ineffective intervention has an effect, which can be very large (depending on the degree of test–retest variability). The mechanism can easily be appreciated in a imaginary experiment with ordinary dice. In experimental design A, a trayful of dice are rolled. Only those dice scoring higher values are enrolled into the study; the others are rejected. The fatal flaw is to accept these chance high values as the baseline. An ineffective intervention is performed, such as blowing on the dice (dotted line). On re-rolling, the dice revert to their usual average values which are significantly lower. Experiment B shows a more sophisticated design. The initially selected dice are re-rolled (blood pressures re-measured), hoping a true baseline will be obtained. It is, but unfortunately a second selection step takes place, again selecting a subset with an artificially high baseline, so an artificial drop to the final value still emerges. To eliminate big day bias requires an extra step shown in Experiment C. After enrolment, a separate baseline measurement must be made and the patient retained regardless of the value, i.e. with no selection at that stage. Most practicing clinicians consider this step an unnecessary waste of resources, but omission, with recycling of an entrance-criterion pressure as baseline, can cause effect sizes to be dramatically exaggerated.
Blood pressure trials understandably only enroll patients with high blood pressure, typically using a threshold. Each patient’s blood pressure varies between visits (Fig. 3). If a measurement is chosen at random, the chances of the next value being larger or smaller are balanced. However if the trial enrolment process inadvertently preferentially uses as baseline a value tending to be above the patient’s true mean, then subsequent (unselected) measurements are more likely to be smaller than larger. This artifact is not an effect on the final measurement itself, only on the change (delta) between the baseline and the final measurement. Readers can explore the size of this effect in the downloadable spreadsheet online Appendix 2.

If a trial’s eligibility criteria require blood pressure to exceed a threshold, patients whose pressure is above their individual long-term mean are more likely to be selected than those whose pressure is below their true mean. If this enrolment value is used as the baseline for measuring changes, their later (unselected) values are more likely to show a negative rather than a positive change. Obtaining a larger sample size, only on the change (delta) between the baseline and the final measurement. Readers can explore the size of this effect in the downloadable spreadsheet online Appendix 2.

If a trial’s eligibility criteria require blood pressure to exceed a threshold, patients whose pressure is above their individual long-term mean are more likely to be selected than those whose pressure is below their true mean. If this enrolment value is used as the baseline for measuring changes, their later (unselected) values are more likely to show a negative rather than a positive change. Obtaining a larger sample size, only on the change (delta) between the baseline and the final measurement. Readers can explore the size of this effect in the downloadable spreadsheet online Appendix 2.

If a trial’s eligibility criteria require blood pressure to exceed a threshold, patients whose pressure is above their individual long-term mean are more likely to be selected than those whose pressure is below their true mean. If this enrolment value is used as the baseline for measuring changes, their later (unselected) values are more likely to show a negative rather than a positive change. Obtaining a larger sample size, only on the change (delta) between the baseline and the final measurement. Readers can explore the size of this effect in the downloadable spreadsheet online Appendix 2.

### 5. ‘Big day bias’: the consequence of using the enrolment BP as the baseline

Blood pressure trials understandably only enroll patients with high blood pressure, typically using a threshold. Each patient’s blood pressure varies between visits (Fig. 3). If a measurement is chosen at random, the chances of the next value being larger or smaller are balanced. However if the trial enrolment process inadvertently preferentially uses as baseline a value tending to be above the patient’s true mean, then subsequent (unselected) measurements are more likely to be smaller than larger. This artifact is not an effect on the final measurement itself, only on the change (delta) between the baseline and the final measurement. Readers can explore the size of this effect in the downloadable spreadsheet online Appendix 2.

If a trial’s eligibility criteria require blood pressure to exceed a threshold, patients whose pressure is above their individual long-term mean are more likely to be selected than those whose pressure is below their true mean. If this enrolment value is used as the baseline for measuring changes, their later (unselected) values are more likely to show a negative rather than a positive change. Obtaining a larger sample size, only on the change (delta) between the baseline and the final measurement. Readers can explore the size of this effect in the downloadable spreadsheet online Appendix 2.

If a trial’s eligibility criteria require blood pressure to exceed a threshold, patients whose pressure is above their individual long-term mean are more likely to be selected than those whose pressure is below their true mean. If this enrolment value is used as the baseline for measuring changes, their later (unselected) values are more likely to show a negative rather than a positive change. Obtaining a larger sample size, only on the change (delta) between the baseline and the final measurement. Readers can explore the size of this effect in the downloadable spreadsheet online Appendix 2.

If a trial’s eligibility criteria require blood pressure to exceed a threshold, patients whose pressure is above their individual long-term mean are more likely to be selected than those whose pressure is below their true mean. If this enrolment value is used as the baseline for measuring changes, their later (unselected) values are more likely to show a negative rather than a positive change. Obtaining a larger sample size, only on the change (delta) between the baseline and the final measurement. Readers can explore the size of this effect in the downloadable spreadsheet online Appendix 2.

### 6. The “check once more” clinical habit

Experienced clinicians realize that measurements in patients involve an element of chance. Some measurements are unrepresentative and should not be used, because of chance biological variation or equipment error (such as a clinically unrealistic oxygen saturation value). Increasing expertise brings increasing confidence in identifying and excluding these, even for reproducibility tests with clear thresholds. [5](Fig. 4) Guidelines for diagnostic modalities rarely forbid taking the whole clinical picture into account and indeed sometimes recommend it? [6]

Selecting the most believable value from a series of differing measurements while unblinded to clinical status, although clinically routine, exposes us to a powerful tendency to find what we expect to find. This can have a large effect, [1] Strict protocols for how many measurements are performed can be helpful, but only if they are followed. The temptation, especially for clinical medical staff, is to use common sense as they do in daily clinical practice, and may be irresistible.

The downloadable spreadsheet (Online Supplement 3) shows the effect of a clinician who is measuring blood pressure before and after an intervention that they believe to be effective. If the post-intervention measurement is higher than pre-intervention, the clinician rejects the measurement and makes another. In reality, however, the intervention is entirely ineffective, with any changes from pre- to post-intervention arising from chance alone. Readers are encouraged to adjust the spreadsheet to observe the size of the artifactual drop in blood pressure that can arise from this “check it once more” habit.

### 7. The evidence so far

The landmark randomized trials in renal denervation so far show office blood pressure drops far exceeding the corresponding ambulatory blood pressure drops, with a ratio approaching 3:1, which has been argued to be a genuine biological pattern. [7] However, amongst the plethora of drug trials for hypertension, it is only the nonrandomized, unblinded studies which show this; randomized blinded drug trials
show identical ambulatory and office blood pressure drops. [3] Moreover, renal denervation trials whose office pressure enrolment criteria are not so markedly high show office blood pressure drops similar in size to ambulatory blood pressure drops. [8]

Renal denervation could only create a novel physiological divergence between office and ambulatory pressure if it dramatically alters or obliterates the alerting response [9] which in turn could only happen if renal nerves are unexpectedly an isthmus through which all alerting response traffic passes. Such an explanation might be considered startling.

8. Clinical implications

Trial design is seen here to be a crucial determinant of observed effect size. What matters clinically is the true incremental effect of denervation, stripped of any systematic overstatement. Strict protocols can be attempted but it may be difficult for clinicians to resist the temptation to discard values that appear unrepresentative. In principle, an automated data capture system would free the investigators from the unpleasant necessity to document values they believe unrepresentative.

One widely-available automated capture system for blood pressure is an ambulatory blood pressure monitor, which may therefore be used as the primary endpoint, not because of its ability to capture diurnal variation, but because it frees the investigator from the burden of following a protocol which appears inappropriately rigid. However, ambulatory blood pressure monitoring is not routinely used in all parts of the world, and some clinical readers may distrust ambulatory data and prefer office data with which they are more familiar. Such readers may incorrectly focus their attention on the office pressure results, realizing neither that they are susceptible to bias nor that this is the reason that office pressure is not the primary endpoint.

An alternative solution which allows both office and ambulatory data to be reliable is to conduct a randomized sham-controlled blinded trial. However, such trials are difficult to establish and it is difficult to ensure complete blinding, especially in the early development of new technologies and therefore form only a small minority of studies embarked upon. None have been reported yet. The ongoing Symplicity HTN-3 trial is using both randomization and blinding to minimize the above limitations, and is also evaluating effects of renal denervation on ambulatory blood pressure. [10]

Unless the alerting response, or white coat hypertension, is mediated solely by the renal sympathetic nerves, it is likely that in future trials designed to resist bias (by randomization and double-blinding) ambulatory and office blood pressure effect sizes will converge. [3] Until such trial data emerge, the best available information on true blood pressure reduction comes from the careful examination of randomized controlled trial data using ambulatory blood pressure monitoring (Fig. 5).

9. Strategies to avoid unintentional exaggeration of effects of renal denervation

A large sample size does not help avoid bias, but rather increases the chance that the bias will be detected as a statistically significant effect and mistaken as a benefit.

Big-day bias can be eliminated if patients are randomized between active and control arms and the effect size is defined as the difference between active and control arms in the final measurements (rather than as the pre–post difference in the active arm alone).

Check-once-more bias can be neutralized by ensuring that the physicians treat both trial arms identically. This can be achieved in two ways.

One way would be to blind the physician assessing blood pressure to the study arm, which is difficult to achieve if the patient is unblinded, since they will inevitably communicate. This way would require the patient to undergo a sham intervention.

The alternative way, which does not need blinding and therefore does not need a sham procedure, is to make it impossible to “check once more”. Stern protocols alone are insufficient to overturn decades of clinical experience that obviously incorrect values should be discarded. Trialists are documented to subvert ingeniously even the sternest protocols. [11] Instead, a practical barrier should be raised against temptation to “check once more”, such as making the measurement a 24-hour recording to which the clinician has no access once initiated. A secondary advantage of this approach is that erroneous readings on an ambulatory monitor are no fault of the physician, and
so the physician would experience no psychological pressure to re-measure and conform to expectations.

8.1. Further advantages of double blinding

Beyond resolving the problems with measurements, double-blinding gives two further benefits. Firstly, clinicians will have no cue from the therapeutic arm to prompt them on whether or not to intensify drug regimens. Second, patient concordance with medication and other medical advice cannot be affected by knowledge of their assigned arm. Whether patient concordance is increased (the procedure serving as a reminder of the seriousness of hypertension) or decreased (through a false belief they have been cured) is unknown. Regardless, with double-blinding, the effect will be the same in both arms.

8.2. Fresh baseline measurement to minimize big-day-bias

Another way to deal with big-day-bias is for the patient to undergo a fresh baseline measurement of blood pressure some time after enrolment. It must be acquired and documented automatically with no opportunity to re-check and no impact on retention of the patient in the trial, regardless of its value. If implemented well, perhaps using an independent member of staff, this fresh post-enrolment baseline may be relatively free of big-day bias. One study implemented this approach inadvertently by using medical students to measure pre and post pressures without knowledge of prior information and without the possibility of the patient being excluded based on the measurement. Of office and ambulatory blood pressure drops were almost identical, at 13.1/5.0 mmHg and 11.3/4.5 mmHg respectively.

Embedded within existing trials there may already be examples of this. Many trials have used office blood pressure for enrolment qualification, and report effects on ambulatory blood pressure, from an unbiased baseline. These ambulatory drops may be the most reliable indication of the denervation effect on blood pressure available so far [3].

9. Implications for research design and clinician education

Blood pressure measurements in real-world clinical practice involve eliminating rogue values to obtain a simple meaningful result consistent with the clinical picture. When this happens in research it causes serious error.

Substantial effort is required to conduct a randomized trial with bias-resistant measurements. This includes planning, arranging ethical clearance, clinical trial registration, individual patient consent and a substantial cost commitment (Fig. 3). In exciting nascent fields, the burden may be too great for researchers and discrepancies can arise, which can have an impact on the observed effect size (Fig. 3). A systematic approach that prevents values being omitted by persons unblinded to study arm is needed. In the case of renal denervation studies, one approach is ambulatory blood pressure monitoring. An unblinded investigator obtaining a rogue high office value in a denervated patient may honestly feel clinical responsibility for checking-once-more (i.e. that the value was erroneous and editing it out), believing that they must have used poor clinical technique. However, they would not feel this pressure to edit out individual values from a 24-hour stream of data, since they would realize (a) that this was not a clinical failure in measurement technique and (b) that editing it would be misconduct. Thus honest desire to do a good job may encourage clinicians to edit office measurements, but to leave ambulatory measurements alone.

Fig. 5. Dose–response effect of bias-resistance features on effect size reported in renal denervation trials. A meta-analysis of 23 studies of renal denervation totaling 720 patients ([3], Howard et al.) shows a dose-response pattern. Studies with progressively more elaborate bias-resistance features show progressively smaller effects of renal denervation on blood pressure. This pattern has been previously recognized in response prediction in biventricular pacing ([13], Nijjer et al.). The raw data for this plot is in Online Appendix 4.
Nevertheless ambulatory measurements in open trials will not completely resolve the question, since the greater office pressure drops may create a belief that therapy has a specific effect on the alerting response. In drug trials, unblinded uncontrolled trials suggest this, but randomized placebo-controlled trials resoundingly dismiss it. [3] In the case of renal denervation, with of office drop at ~10 mmHg and ambulatory drop at ~30 mmHg it is all the more important to determine whether the gap between them is real biology or merely a mirage arising from unintended measurement bias. One way to resolve this is to conduct placebo-controlled trials, i.e. with one arm receiving sham procedures. A small number of such trials are now in progress and may report in the next year or two. [10,12] Without these efforts to eliminate big-day-bias (by having a control arm) and check-once-more bias (by using ambulatory pressure or a placebo procedure), the data acquired is vulnerable to biases large enough to dwarf the actual effect size.

This situation has arisen because clinical practice accommodates irreproducible methods largely uncritically. With limited time and resources, it seems impractical to pursue methods of reducing measurement scatter and bias. Instead, a pragmatic approach of selecting quantitative aspects of measurement variability including how to measure it, its impact on individual patient data, and on group data (under various measurement protocols). Our community would thereby be better placed to recognize trial designs that risk unintentional exaggeration.

10. Disclosures

Data used are in the public domain. This paper was written with no knowledge of the data of the Symplicity-HTN-3 trial.

Dr. Justin E Davies is a consultant for Medtronic Inc.

11. Contribution statement

JPH: Designed the study; retrieved and processed the data; designed the images; drafted and revised the manuscript;
GDC: Co-designed the study; retrieved and processed the data; designed the images; drafted and revised the manuscript;
HS: Co-designed the study; drafted, reviewed and revised the manuscript;
DLB: Co-designed the study; drafted, reviewed and revised the manuscript;
VP: Co-designed the study; drafted, reviewed and revised the manuscript;
DEK: Co-designed the study; drafted, reviewed and revised the manuscript;
JED: Co-designed the study; co-designed the study; designed the images; drafted and revised the manuscript;
DPF: Initiated the study; co-designed the study; designed the images; drafted and revised the manuscript.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ijcard.2013.12.183.

References