JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2015 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER INC.

VOL. 66, NO. 22, 2015
ISSN 0735-1097/$36.00
http://dx.doi.org/10.1016/j.jacc.2015.10.014

THE PRESENT AND FUTURE

STATE-OF-THE-ART REVIEW

Making Sense of Statistics in Clinical Trial Reports

Part 1 of a 4-Part Series on Statistics for Clinical Trials

Stuart J. Pocock, PhD,* John J.V. McMurray, MD,† Tim J. Collier, MSc*

ABSTRACT

This paper is a practical guide to the essentials of statistical analysis and reporting of randomized clinical trials (RCTs). It is the first in a series of 4 educational papers on statistical issues for RCTs, which will also include statistical controversies in RCT reporting and interpretation, the fundamentals of design for RCTs, and statistical challenges in the design and monitoring of RCTs. Here, we concentrate on displaying results in tables and figures, estimating treatment effects, expressing uncertainty using confidence intervals, and using p values wisely to assess the strength of evidence for a treatment difference. The various methods and their interpretation are illustrated by recent, topical cardiology trial results. (J Am Coll Cardiol 2015;66:2536–49) © 2015 by the American College of Cardiology Foundation.

Statistical methods are an essential part of virtually all published medical research. Yet, a sound understanding of statistical principles is often lacking amongst researchers and journal readers, and cardiologists are no exception to this limitation. In this series of 4 papers in consecutive issues of the *Journal*, our aim is to illuminate readers on statistical matters, our focus being on the design and reporting of randomized controlled trials (RCTs).

After these first 2 papers on *statistical analysis and reporting of clinical trials*, 2 subsequent papers will focus on statistical *design of randomized trials* and also *data monitoring*. The principles are brought to life by real topical examples, and besides laying out the fundamentals, we also tackle some common misperceptions and some ongoing controversies that affect the quality of research and its valid interpretation.

Constructive critical appraisal is an art continually exercised by journal editors, reviewers, and readers, and is also an integral part of good statistical science that we hope to encourage via our choice of examples. Throughout this series, we concentrate on concepts rather than providing formulae or calculation techniques, therefore ensuring that readers without a mathematical or technical background can grasp the essential messages we wish to convey.

THE ESSENTIALS OF STATISTICAL ANALYSIS

The 4 main steps in data analysis are:

1. Displaying results in tables and figures
2. Quantifying any associations (e.g., estimates of treatment differences in patient outcomes)
3. Expressing the uncertainty in those associations by use of confidence intervals (CIs)
4. Assessing the strength of evidence that the association is "real" (i.e., more than could be expected by chance) by using p values (statistical tests of significance)

The next few sections take us through these essentials, illustrated by examples from randomized

Listen to this manuscript's audio summary by *JACC* Editor-in-Chief Dr. Valentin Fuster.

From the *Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, United Kingdom; and the †Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, United Kingdom. The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

Manuscript received September 8, 2015; revised manuscript received October 12, 2015, accepted October 18, 2015.

trials. The same principles broadly apply to observational studies, with 1 major proviso: in nonrandomized studies, one cannot readily infer that any association not due to chance indicates a causal relationship.

Also, next week we discuss some of the more challenging issues when reporting clinical trials.

DISPLAYING RESULTS IN TABLES AND FIGURES

TABLE OF BASELINE DATA. The first table in any clinical trial report shows patients' baseline characteristics by treatment group. Which characteristics to present will vary by trial, but will almost always include key demographic variables, related medical history, and other variables that might be strongly related to the trial endpoints. See **Table 1** as an example from the PARADIGM-HF trial (Prospective Comparison of Angiotensin Receptor–Neprilysin Inhibitor with Angiotensin-Converting-Enzyme Inhibitor to Determine Impact on Global Mortality and Morbidity in Heart Failure) (1). Note that categorical variables are shown as number (%) by group. For quantitative variables, there are 2 common options: means (and SDs) or median (and interquartile range). For variables with a skew distribution, the latter is often preferable, geometric means being another option. In addition, some such variables may be formed into categories, for example, age groups or specific (abnormal) cut-offs for biochemical variables. This (and indeed any other table) should include the total number of patients per group at the top. To limit the size of **Table 1**, a third column showing results for all groups combined may be unnecessary. Also, for some binary variables (e.g., sex or disease history) only 1 category (e.g., male or diabetic) need be shown. Unnecessary precision in reporting means or percentages should be avoided, with 1 decimal place usually being sufficient. The use of p values in baseline tables should also be avoided, because in the setting of a well-conducted RCT, any differences at baseline must have arisen by chance.

TABLE OF MAIN OUTCOME EVENTS. The key table for any clinical trial displays the main outcomes by treatment group. For trials concentrating on clinical events during follow-up, the numbers (%) by group experiencing each type of event should be shown. See **Table 2** as an example from the SAVOR-TIMI 53 trial (Saxagliptin Assessment of Vascular Outcomes Recorded in Patients with Diabetes Mellitus-Thrombolysis In Myocardial Infarction 53) (2).

For any composite event (e.g., death, myocardial infarction, and stroke), the number of patients experiencing any of them (i.e., the composite) plus the numbers in each component should all be shown. Because some patients can have more than 1 type of event (e.g., nonfatal myocardial infarction followed by death), the numbers in each component usually add up to slightly more than the numbers with composite events.

The focus is often on time to first event, so any subsequent (repeat) events (e.g., a second or third myocardial infarction) do not get included in the main analyses. This is not a problem when the frequency of repeat events is low. But for certain chronic disease outcomes, such as hospitalization for heart failure, repeat events are more common. For instance, in the CORONA (Controlled Rosuvastatin Multinational Trial in Heart Failure) trial (3) of rosuvastatin versus placebo in chronic heart failure, there were a total of 2,408 heart failure hospitalizations in 1,291 of 5,011 randomized patients. Conventional analyses of time to first hospitalization was inconclusive, but analyses using all hospitalizations (including repeats) gave strong evidence of a treatment benefit in that secondary outcome (4).

In trials of chronic diseases (e.g., chronic heart failure), in which the incidence rates over time are fairly steady, it may be useful to replace % by the incidence rate per 100 patient-years, for example, of follow-up in each group: to calculate the incidence rate one divides the number of patients with the relevant event by the total follow-up time in years of all patients (excluding any follow-up after an event occurs). Such a table will usually add in estimates of treatment effect, CIs, and p values, as dealt with in the next 3 sections, and already shown in **Table 2**. Another important table concerns adverse events by treatment group.

KAPLAN-MEIER PLOT. The most common type of Figure in major trial reports is a Kaplan-Meier plot of time-to-event outcomes. **Figure 1** shows this for the primary outcome (death, myocardial infarction, or stroke) of PLATO (Study of Platelet Inhibition and Patient Outcomes) (5). The figure clearly displays the steadily accumulating difference in incidence rates between ticagrelor and clopidogrel. There are several features that make for a good quality Kaplan-Meier plot (6). The numbers at risk in each group should be shown at regular time intervals of follow-up. In this case, we see that nearly all patients had 6 months of follow-up, but only around one-half of patients were followed for 1 year. In connection with this, we recommend that the time axis should not be extended too far, perhaps not beyond the time when <10% of patients are still under follow-up.

ABBREVIATIONS AND ACRONYMS

ANCOVA = analysis of covariance

CABG = coronary artery bypass grafting

CI = confidence interval

PCI = percutaneous coronary intervention

RCT = randomized clinical trial

SBP = systolic blood pressure

SD = standard deviation

SE = standard error

TABLE 1 Characteristics of the Patients at Baseline in the PARADIGM-HF Trial

	LCZ696 (N = 4187)	Enalapril (N = 4212)
Age, yrs	63.8 ± 11.5	63.8 ± 11.3
Female	879 (21.0)	953 (22.6)
Race or ethnic group		
White	2,763 (66.0)	2,781 (66.0)
Black	213 (5.1)	215 (5.1)
Asian	759 (18.1)	750 (17.8)
Other	452 (10.8)	466 (11.1)
Region		
North America	310 (7.4)	292 (6.9)
Latin America	713 (17.0)	720 (17.1)
Western Europe and other	1,026 (24.5)	1,025 (24.3)
Central Europe	1,393 (33.3)	1,433 (34.0)
Asia-Pacific	745 (17.8)	742 (17.6)
Systolic blood pressure, mm Hg	122 ± 15	121 ± 15
Heart rate, beats/min	72 ± 12	73 ± 12
Body mass index	28.1 ± 5.5	28.2 ± 5.5
Serum creatinine, mg/dl	1.13 ± 0.3	1.12 ± 0.3
Clinical features of heart failure		
Ischemic cardiomyopathy	2,506 (59.9)	2,530 (60.1)
Left ventricular ejection fraction, %	29.6 ± 6.1	29.4 ± 6.3
Median B-type natriuretic peptide, pg/ml	255 (155–474)	251 (153–465)
Median N-terminal pro-B-type natriuretic peptide, pg/ml	1,631 (885–3,154)	1,594 (886–3,305)
NYHA functional class		
I	180 (4.3)	209 (5.0)
II	2,998 (71.6)	2,921 (69.3)
III	969 (23.1)	1,049 (24.9)
IV	33 (0.8)	27 (0.6)
Missing data	7 (0.2)	6 (0.1)
Medical history		
Hypertension	2,969 (70.9)	2,971 (70.5)
Diabetes	1,451 (34.7)	1,456 (34.6)
Atrial fibrillation	1,517 (36.2)	1,574 (37.4)
Hospitalization for heart failure	2,607 (62.3)	2,667 (63.3)
Myocardial infarction	1,818 (43.4)	1,816 (43.1)
Stroke	355 (8.5)	370 (8.8)
Pre-trial use of ACE inhibitor	3,266 (78.0)	3,266 (77.5)
Pre-trial use of ARB	929 (22.2)	963 (22.9)
Treatments at randomization		
Diuretic agent	3,363 (80.3)	3,375 (80.1)
Digitalis	1,223 (29.2)	1,316 (31.2)
Beta-blocker	3,899 (93.1)	3,912 (92.9)
Mineralocorticoid antagonist	2,271 (54.2)	2,400 (57.0)
Implantable cardioverter-defibrillator	623 (14.9)	620 (14.7)
Cardiac resynchronization therapy	292 (7.0)	282 (6.7)

Values are mean ± SD, n (%), or median (interquartile range). Table summarizing the characteristics at the baseline visit for patients in the PARADIGM-HF trial by treatment allocation. Adapted with permission from McMurray et al. (1).

ACE = angiotensin-converting enzyme; ARB = angiotensin receptor blocker; NYHA = New York Heart Association; PARADIGM-HF = Prospective Comparison of Angiotensin Receptor-Neprilysin Inhibitor with Angiotensin-Converting-Enzyme Inhibitor to Determine Impact on Global Mortality and Morbidity in Heart Failure.

One good practice that is, sadly, rarely done is to convey the extent of statistical uncertainty in the estimates over time by plotting standard error (SE) bars at regular time points. In this case, the SEs would be much tighter at 6 months compared with 1 year, reflecting the substantial proportion of patients not followed out to 1 year.

Sometimes, Kaplan-Meier plots are inverted, thereby showing the declining percentage of patients over time that are event free. This can be particularly misleading if there is a break in the vertical axis (which readers may not spot). In general, we feel it is more informative to have the curves going up (not down), thereby focusing on cumulative incidence, with a sensible range (up to 12% in this case) rather than a full vertical axis up to 100%, so that relevant details, especially regarding treatment differences, can be clearly seen. The choice of vertical scale is an important ingredient in interpreting these plots; not so wide (0% to 100%) as to cramp the visual effect, but not so tight as to exaggerate any small differences that may occur.

REPEATED MEASURES OVER TIME. For quantitative or symptom-related outcomes, repeated measures over time are usually obtained at planned visits. Consequent treatment comparisons of means (or % with symptoms) are usually best presented in a figure. See **Figure 2** for mean systolic blood pressure in the PARADIGM-HF trial (1), both in the build-up to randomization and over the subsequent 3 years. Each mean by treatment group should have SE bars around it. In this case, the large numbers of patients make the tiny SEs hard to see. With such precise estimation it is obvious without formal testing that mean systolic blood pressure is consistently around 2.5 mm Hg lower on LCZ696 compared with enalapril, but this secondary finding was peripheral to the trial's main aims concerning clinical events.

TRIAL PROFILE. As part of the CONSORT guidelines for clinical trial reports (7), it is recommended that every trial publication should have a trial profile that shows the flow of patients through the trial from the pre-randomization build-up to the post-randomization follow-up. **Figure 3** is an example from the HEAT PPCI trial (How Effective Are Antithrombotic Therapies in Primary PCI) (8). It nicely shows the high proportion of eligible patients who were randomized, the small number not getting their randomized treatment (but still included in intention to treat analysis), the controversial delayed consent, and the consequent small numbers removed from analysis or lost to follow-up. The use of delayed consent meant 17 patients died before consent could be obtained, and for a further 17 surviving patients, no consent was obtained. **Figure 3** shows how 2,499 patients were identified, 1,829 were randomized, and 1,812 were included in the analysis, with all

steps along the way documented, each with patient numbers. Note that with a more conventional patient consent prior to randomization, the trial profile would become somewhat simplified.

The next most common figure is the forest plot for subgroup analyses, but more on that in next week's paper.

ESTIMATES OF TREATMENT EFFECTS AND THEIR CIs

Now we get down to the serious business of estimating the magnitude of the difference between treatments on patient outcomes. First, we wish to obtain a *point estimate*, that is, the actual difference observed. Then we need to express the degree of uncertainty present in the data, that is, the bigger the trial, the more precise the point estimate will be. Such uncertainty is usually expressed as a *95% CI*.

Exactly what type of estimate is required depends on the nature of the patient outcome of interest. There are 3 main types of outcome data:

1. A binary (yes/no) response, for example, success or failure, dead or alive, or in the trial we pursue, below the composite of death, myocardial infarction, ischemia-driven revascularization, or stent thrombosis (i.e., did any of these occur within 48 h of randomization in percutaneous coronary intervention [PCI] patients, yes or no)?
2. A time to event outcome, for example, time to death, time to symptom relief, or in the trial we pursue, below the time to first hospitalization for heart failure or cardiovascular death, whichever (if either) happens first.
3. A quantitative outcome, for example, change in systolic blood pressure from randomization to 6 months later.

What follows are the standard estimation methods for these 3 types of data. In the process, we also explain what a CI actually means.

ESTIMATES BASED ON PERCENTAGES. In acute disease, the comparative efficacy of 2 treatments is often assessed by "success or failure" in terms of "absence or presence" of a serious clinical event. For instance, in the CHAMPION-PHOENIX trial (Cangrelor versus Standard Therapy to Achieve Optimal Management of Platelet Inhibition PHOENIX) (9), the primary outcome was the composite of death, myocardial infarction, ischemia-driven revascularization, or stent thrombosis within 48 h of randomization. Patients undergoing PCI were randomized to cangrelor or clopidogrel (n = 5,470 and n = 5,469, respectively) and the numbers (%) experiencing

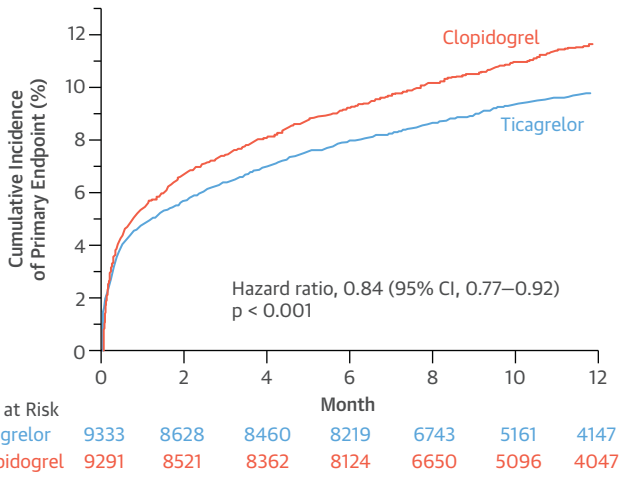
TABLE 2 Pre-Specified Clinical Endpoints in the SAVOR-TIMI 53 Trial

	Saxagliptin (n = 8,280)	Placebo (n = 8,212)	Hazard Ratio (95% CI)	p Value
Cardiovascular death, myocardial infarction, or stroke: primary efficacy endpoint	613 (7.3)	609 (7.2)	1.00 (0.89–1.12)	0.99
Cardiovascular death, myocardial infarction, stroke, hospitalization for unstable angina, heart failure, or coronary revascularization: secondary efficacy endpoint	1,059 (12.8)	1,034 (12.4)	1.02 (0.94–1.11)	0.66
Death from any cause	420 (4.9)	378 (4.2)	1.11 (0.96–1.27)	0.15
Death from cardiovascular causes	269 (3.2)	260 (2.9)	1.03 (0.87–1.22)	0.72
Myocardial infarction	265 (3.2)	278 (3.4)	0.95 (0.80–1.12)	0.52
Ischemic stroke	157 (1.9)	141 (1.7)	1.11 (0.88–1.39)	0.38
Hospitalization for unstable angina	97 (1.2)	81 (1.0)	1.19 (0.89–1.60)	0.24
Hospitalization for heart failure	289 (3.5)	228 (2.8)	1.27 (1.07–1.51)	0.007
Hospitalization for coronary revascularization	423 (5.2)	459 (5.6)	0.91 (0.80–1.04)	0.18
Doubling of creatinine level, initiation of dialysis, renal transplantation, or creatinine >6.0 mg/dl (530 μmol/l)	194 (2.2)	178 (2.0)	1.08 (0.88–1.32)	0.46
Hospitalization for hypoglycemia	53 (0.6)	43 (0.5)	1.22 (0.82–1.83)	0.33

Values are n (%), unless otherwise indicated. 2-year Kaplan-Meier estimates and hazard ratios (95% confidence intervals [CIs]) for pre-specified clinical endpoints in the SAVOR-TIMI 53 trial. Percentages are 2-year Kaplan-Meier estimates. Adapted from Scirica et al. (2).

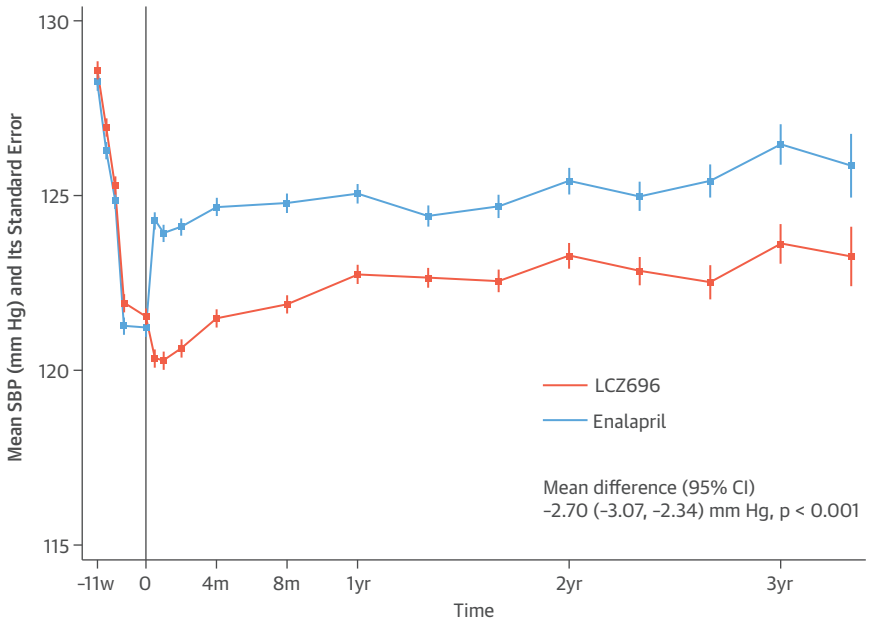
SAVOR-TIMI 53 = Saxagliptin Assessment of Vascular Outcomes Recorded in Patients with Diabetes Mellitus–Thrombolysis In Myocardial Infarction 53.

FIGURE 1 Kaplan-Meier Estimates of the Cumulative Incidence Over Time of the First Adjudicated Occurrence of the Primary Efficacy Endpoint in the PLATO Trial



Cumulative incidence of the primary endpoint—a composite of death from vascular causes, myocardial infarction, or stroke—was significantly lower in the ticagrelor group than in the clopidogrel group (9.8% vs. 11.7% at 12 months; hazard ratio: 0.84; 95% confidence interval [CI]: 0.77 to 0.92; p < 0.001). PLATO = The Study of Platelet Inhibition and Patient Outcomes. Reprinted with permission from Wallentin et al. (5).

FIGURE 2 Systolic Blood Pressure During Run-In And After Randomization in the PARADIGM-HF Trial



Mean systolic blood pressure (SBP) by visit and treatment group, and overall mean difference (95% confidence interval [CI]) in the PARADIGM-HF trial. PARADIGM-HF = Prospective Comparison of Angiotensin Receptor–Neprilysin Inhibitor with Angiotensin-Converting-Enzyme Inhibitor to Determine Impact on Global Mortality and Morbidity in Heart Failure. Reprinted, with permission, from McMurray et al. (1).

the primary composite outcome were 257 (4.7%) and 322 (5.9%), respectively. The various estimates of comparative treatment efficacy based on these 2 percentages are displayed in **Table 3**, with each estimate accompanied by its 95% CI.

Relative risk is the ratio of 2 percentages, here 0.798, and can be converted to the *relative risk reduction*, which on a percentage scale is 20.2%. A common alternative to relative risk is *relative odds*, here 0.788. This is less readily understandable, because except for those who gamble on horses, the concept of odds is harder to grasp. However, as explained later, relative odds are linked to logistic regression, which permits adjustment for baseline variables. Relative risk and relative odds are sometimes called risk ratio and odds ratio instead. If event rates are small then the 2 give quite similar estimates, with the odds ratio always slightly further away from 1.

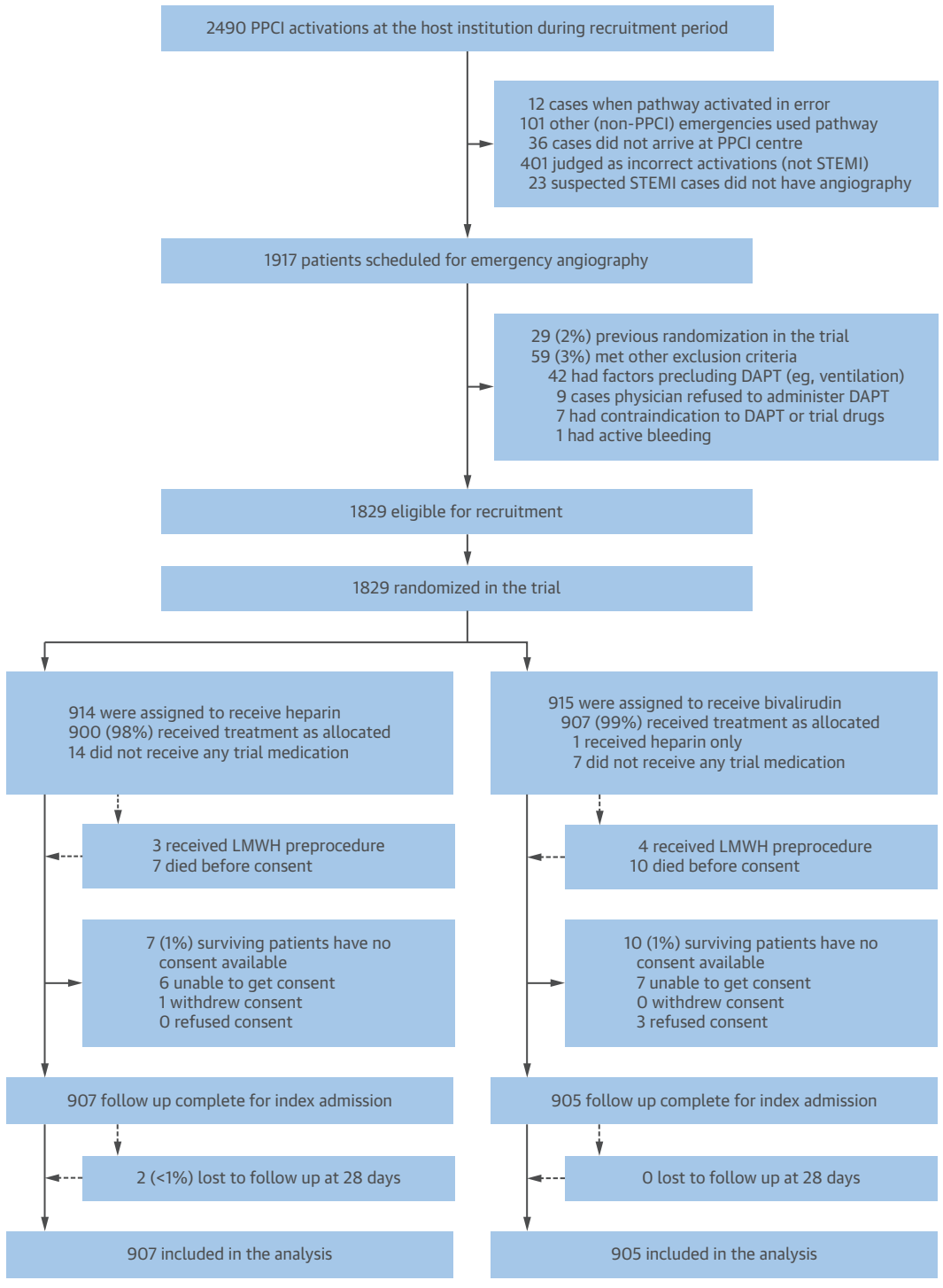
The *absolute difference in percentages*, here 1.19%, is another important statistic. It is sometimes called the absolute risk reduction. In trial reports, it is useful to present both the absolute and relative risk reduction. The former expresses the estimated absolute benefit across all randomized patients in avoiding the primary endpoint by giving cangrelor instead of clopidogrel. The latter expresses in relative terms what estimated percentages of primary events on clopidogrel would have been prevented by using cangrelor instead.

The difference in percentages can be converted into the *number needed to treat* (NNT), here 84.0. This means that to prevent 1 primary event by using cangrelor instead of clopidogrel we need to treat an estimated 84 patients. For NNT, it is important to note the relevant timeframe: here it is 48 h post-randomization.

EXPRESSING UNCERTAINTY USING CIs. All estimates based on percentages, such as in **Table 3** (and, indeed, other types of estimates to follow in the next 2 sections) are not to be trusted at face value. Any estimate has a built-in imprecision because of the finite sample of patients studied, and indeed the smaller the study, the less precise the estimate will be. The extent of such statistical uncertainty is best captured by use of a 95% CI around any estimate (10,11).

For instance, the observed relative risk reduction of 20.2% has a 95% CI from 6.4% to 32.0%. What does this mean? In simple terms, we are 95% sure that the true reduction with cangrelor versus clopidogrel lies between 6.4% and 32.0%. However, the frequentist principles of statistical inference, which underpin all use of confidence intervals and p values, give a more

FIGURE 3 **Trial Profile of the HEAT-PPCI Trial**



Trial profile for HEAT-PPCI summarizing the flow of patients through the trial from the pre-randomization recruitment period to the post-randomization follow-up and analysis. Reprinted with permission from Shahzad et al. (8). DAPT = dual antiplatelet therapy; HEAT-PPCI = How Effective Are Antithrombotic Therapies in Primary PCI; LMWH = low-molecular-weight heparin; PPCI = primary percutaneous coronary intervention; STEMI = ST-segment elevation myocardial infarction.

TABLE 3 Estimates Based on the Comparison of 2 Percentages, Illustrated by the Primary Outcome* of the CHAMPION-PHOENIX Trial

	Cangrelor	Clopidogrel
Randomized patients, n	5,470	5,469
Patients with primary outcome, n (%)	257 (4.698)†	322 (5.888)†

Estimate	Formula	Result
Relative risk (95% CI)	$\frac{4.698}{5.888}$	= 0.798 (0.680 to 0.936)
Relative risk reduction (95% CI)	$(1 - 0.798) \times 100$	= 20.2% (6.4% to 32.0%)
Relative odds (95% CI)	$\frac{4.698/(100 - 4.698)}{5.888/(100 - 5.888)}$	= 0.788 (0.666 to 0.932)
Difference in percentages (95% CI)	$4.698 - 5.888$	= −1.19% (−0.35% to −2.03%)
Number needed to treat (95% CI)	$\frac{100}{1.19}$	= 84.0 (49.3 to 285.7)

Number and percentage of patients with a primary outcome (death, myocardial infarction, ischemia driven revascularization, or stent thrombosis within 48 h of randomization) in the CHAMPION-PHOENIX trial along with various estimates of treatment effect. *Primary composite outcome is death, myocardial infarction, ischemia-driven revascularization, or stent thrombosis within 48 h of randomization. †In the middle of all numerical calculations, any values (e.g., percentages) should be precise (e.g., to ≥3 decimal places). Only at the final step should values be rounded for convenience of expression.

CHAMPION-PHOENIX = Cangrelor versus Standard Therapy to Achieve Optimal Management of Platelet Inhibition PHOENIX trial; CI = confidence interval.

precise meaning as follows. If we were to repeat the whole clinical trial many, many times using an identical protocol we would get a slightly different confidence interval each time. Of those CIs, 95% would contain the true underlying relative risk reduction. But, whenever we calculate a 95% CI, there is a 2.5% chance that the true effect lies below and a 2.5% chance that the true effect lies above the interval.

What matters here is that the whole 95% CI indicates a clear relative risk reduction. This is reinforced by 95% CI for the difference that is from −0.35% to −2.03% (**Table 3**). These relatively tight CIs, each wholly in a direction substantially favoring cangrelor, provides strong evidence that cangrelor reduces the risk of the primary endpoint compared with clopidogrel. Later, we achieve the same message by use of a p value. Note that **Table 3** also gives a 95% CI for the NNT. Some trials report the NNT but not its 95% CI, a practice to be avoided because readers can be led astray by thinking that the NNT is precisely known.

An important obvious principle is that larger studies (more patients and hence more events) produce more precise estimation and tighter CIs. Specifically, to halve the width of a CI, one needs 4× as many patients. This logic feeds into statistical power calculations when designing a clinical trial (see future paper in this series).

Another issue is: why choose 95% confidence, and why not 90% or 99%? Well, there is no universal wisdom that says 95% is the right thing to do. It is just a convenient fashion that, for consistency's sake, virtually all papers follow. It also has a link to $p < 0.05$, as discussed in the following text. It is worth noting that "confidence" is not evenly distributed over a 95% CI. For instance, there is around a 70% chance that the true treatment effect lies in the inner one-half of the 95% CI. Also, 1.5 × (95% CI width) = 99.9% CI.

ESTIMATES FOR TIME-TO-EVENT OUTCOMES. Many major clinical trials have a primary outcome, which is time to an event. In the PLATO trial (5) (**Figure 1**), the Kaplan-Meier plot is for time to the primary composite outcome of death, myocardial infarction, and stroke. The curves diverge in favor of ticagrelor, but do not in themselves provide a simple estimate summarizing the treatment difference. One can read off the Kaplan-Meier estimate at the end of plotted time, 1 year in this instance, and they are 1-year cumulative rates of 9.8% and 11.7% for ticagrelor and clopidogrel, respectively. If everyone had been followed for 1 year this would have merit (and a 95% CI for the treatment difference 1.9% could be calculated), but given that only around one-half of the patients have been followed for a year, this is far from ideal.

Instead, the most common approach is to use a Cox proportional hazards model to obtain a *hazard ratio* and its 95% CI, which in this case is 0.84 (95% CI: 0.77 to 0.92). Technically, the instantaneous hazard rate at any specific time point is the probability of the outcome occurring exactly at that time for patients who are still outcome-free. The hazard ratio can be thought of as the hazard rate in one group (ticagrelor) divided by the hazard rate in the other group (clopidogrel) averaged over the whole follow-up period. Conceptually, it is similar to the relative risk (risk ratio), except that follow-up time is taken into account (12).

In this example, the hazard ratio and its 95% CI are all substantially <1 indicating strong evidence of fewer primary outcomes on ticagrelor compared to clopidogrel. For those who like a more straightforward statistical life, we note that the numbers of patients having the primary outcome are 864 and 1,014 in the ticagrelor and clopidogrel groups, respectively. The simple ratio 864/1,014 = 0.852 is very similar to the hazard ratio, as will usually be the case for trials with equal randomization and relatively low event rates.

The hazard ratio is best suited to data where the Kaplan-Meier plot shows a steady divergence between treatment groups. But, in some trials, especially with surgical intervention, one might anticipate an early excess risk followed by a subsequent gain in efficacy as time goes by. For instance, the FREEDOM trial (Future Revascularization Evaluation in Patients with Diabetes Mellitus: Optimal Management of Multivessel Disease) (13) of coronary artery bypass grafting (CABG) versus PCI has a primary

composite endpoint of death, myocardial infarction, or stroke (**Figure 4**). An early excess event rate for CABG (mainly due to stroke) is followed by a lower event rate after the first 6 months. The Kaplan-Meier curves cross at around 1 year. Here, a hazard ratio would be a peculiar average of early bad news followed by later good news for CABG, and so is not particularly useful. The focus on the 5-year composite event rate (18.7% on CABG, 26.6% on PCI) is informative, but suffers from the fact that only around one-third of patients have so far been followed for 5 years. A more complete 5-year follow-up is required to clarify this.

Another problem with hazard ratios is that they focus on a "vertical interpretation" of the Kaplan-Meier plot. But, in chronic diseases, a more "horizontal interpretation" focusing on event-free time gained may be more appropriate. The *accelerated failure time model* (14) is unfortunately rarely used, but it can nicely capture this concept. In a nutshell, it estimates a time ratio whereby if a new treatment helps to delay the occurrence of events, the time ratio will be >1. For instance, a time ratio of 1.5 means that, on average, it takes 50% longer for an event to occur in patients on the new treatment compared with control subjects.

We illustrate the use of the accelerated failure time model in a post hoc analysis of the EMPHASIS-HF (Eplerenone in Mild Patients Hospitalization and Survival Study in Heart Failure) trial (15) of eplerenone versus placebo in patients with systolic heart failure and mild symptoms. The primary composite endpoint, heart failure hospitalization or cardiovascular death, is plotted in **Figure 5**. Reading off horizontally from this plot, the eplerenone and placebo groups reach 10% incidence at 0.84 and 0.40 years, respectively, a time ratio of 2.10. The 20% incidence occurs at 2.02 and 1.09 years, respectively, a time ratio of 1.86. The accelerated failure time model averages these time ratios across all possible cut-offs on the vertical scale of **Figure 4**. The end result is a time ratio of 1.71 in favor of eplerenone, with 95% CI: 1.38 to 2.11.

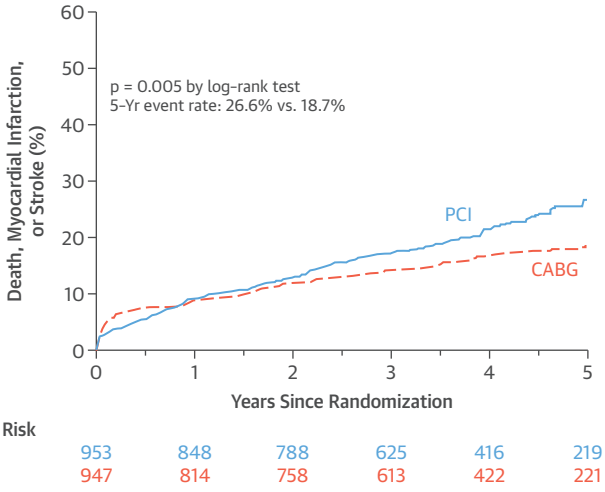
An alternative simpler approach is to calculate the incidence rate of the primary endpoint over all follow-up: for eplerenone and placebo groups this is 10.60 and 15.47 per 100 patient-years, respectively. This gives a rate ratio of 0.69, the inverse of which is the time ratio 1.47. This crude approach works well, provided incidence rates are fairly steady throughout follow-up. But, in general, the Kaplan-Meier plot in many diseases shows a much higher incidence rate in early follow-up.

ESTIMATES FOR QUANTITATIVE OUTCOMES. In analyzing a quantitative outcome, one usually compares the means in 2 treatment groups. However, as the same outcome is usually measured at baseline, it is more efficient to compare mean changes from baseline. Yet, this still misses the fact that changes tend to depend on the baseline value, based on the concept of regression to the mean. That is, patients with high baseline value tend to have a bigger fall in value than patients with lower baseline values. This requires an analysis of covariance (ANCOVA), in which one compares mean changes adjusted for baseline value (16).

We illustrate these issues using results for the primary endpoint, 6-month change in systolic blood pressure (SBP) in the SYMPLICITY HTN-3 trial (Renal Denervation in Patients With Uncontrolled Hypertension) (17) comparing renal denervation with a sham procedure in a 2:1 randomization (n = 350 and n = 169 patients for renal denervation and sham, respectively) (**Table 4**).

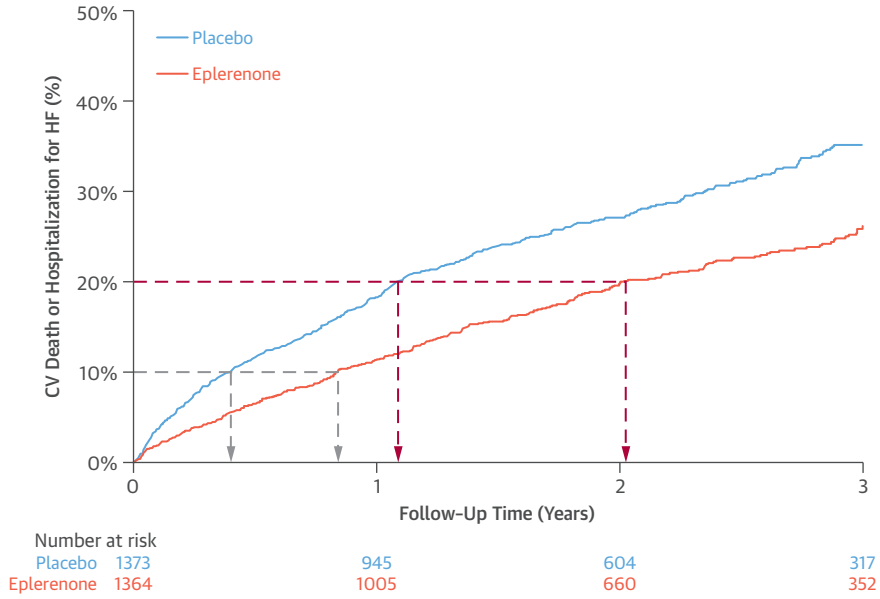
First, note the relatively poor showing of an analysis of the 6-month SBP only, ignoring baseline: this fails to account for the marked variation in patients' baseline blood pressure, and hence yields a wider 95% CI. The comparison of the 2 analyses of mean changes, with and without adjustment for baseline, is more subtle. Results are fairly similar, but it is statistically inevitable that ANCOVA produces a slightly more precise estimate of the

FIGURE 4 Kaplan-Meier Estimates of the Cumulative Incidence Over Time of the Primary Efficacy Endpoint in the FREEDOM Trial



Kaplan-Meier estimates of the cumulative incidence of the primary efficacy endpoint (death, myocardial infarction, or stroke) by treatment group in the FREEDOM trial. FREEDOM = Future Revascularization Evaluation in Patients with Diabetes Mellitus: Optimal Management of Multivessel Disease. Reprinted with permission from Tepel et al. (25). CABG = coronary artery bypass graft; PCI = percutaneous coronary intervention.

FIGURE 5 Time to Primary Event in the EMPHASIS-HF Trial



Kaplan-Meier cumulative incidence of cardiovascular (CV) death or hospitalization for heart failure (HF) in the EMPHASIS-HF trial. The eplerenone and placebo groups reach 10% incidence at 0.84 and 0.40 years, respectively, a time ratio of 2.10. The 20% incidence occurs at 2.02 and 1.09 years, respectively, a time ratio of 1.86. EMPHASIS-HF = Eplerenone in Mild Patients Hospitalization and Survival Study in Heart Failure.

treatment effect, that is, its 95% CI is a bit tighter (18). Even so, in this case, the 95% CI still includes zero treatment difference, meaning that there is insufficient evidence that renal denervation lowers SBP in this population.

What ANCOVA is doing is illustrated in **Figure 6**, which plots individual 6-month change in SBP by baseline SBP using different symbols for the 2 treatment groups. The 2 drawn parallel regression lines show the anticipated regression to the mean, that is, patients near to the minimum eligible baseline SBP of 160 mm Hg have (in both treatment groups) a tendency to have less blood pressure reduction compared with those starting at higher levels. The vertical distance between the 2 regression lines is 4.11 mm Hg, the mean treatment effect adjusted for baseline. Note that this kind of scatter diagram is a useful reminder as to the huge individual variation in SBP over time (with or without treatment), which is why we need clinical trials of several hundred patients to detect realistic treatment effects.

One issue is whether to choose absolute change (as here) or percentage change from baseline. Statistically, it depends on which gives the better model fit using ANCOVA.

When a quantitative outcome is measured repeatedly over time at planned visits, there are various options for statistical analysis, depending on what estimate of treatment effect one wishes to focus on. It could be: 1) the mean treatment difference averaged over time, as in the PARADIGM-HF trial (1) SBP data in **Figure 2**; 2) the differing rates of decline (slopes) in, say, forced expiratory volume in a study of deteriorating respiratory function; or 3) a mean treatment effect at a specific point of follow-up, for example, glycated hemoglobin at 18 months in a trial evaluating glycemic efficacy of antidiabetic drugs. In each case, the correlation structure in each within-patient trajectory is used in a repeated measures analysis, often with variation in the extent of patient

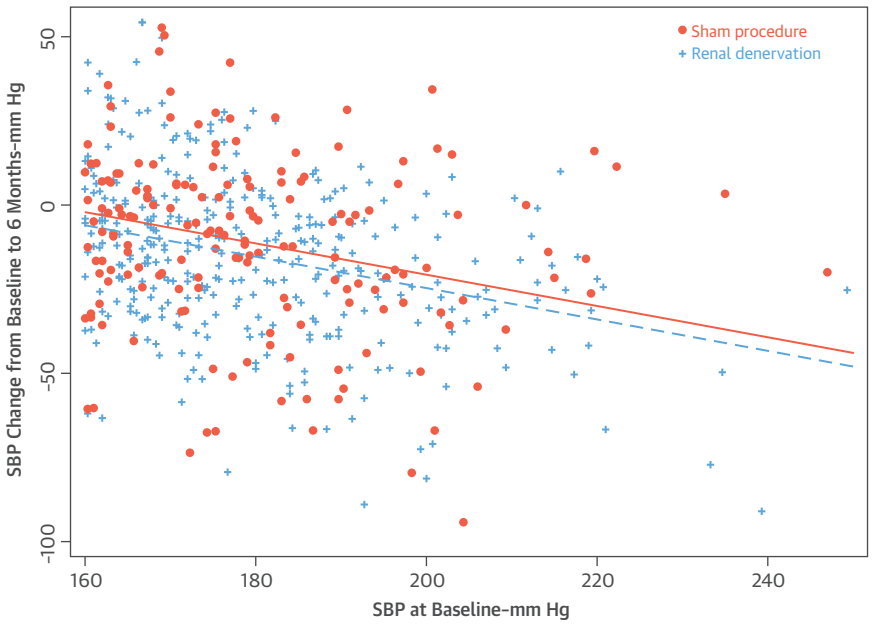
TABLE 4 6-Month Results From the SYMPLICITY HTN-3 Trial

SBP at 6 months	−4.20 (−9.17 to +0.77)
6-month change in SBP	−4.07 (−8.63 to +0.49)
6-month change in SBP adjusted for baseline SBP using ANCOVA	−4.11 (−8.44 to +0.22)

Values are mean treatment difference (95% confidence interval) in mm Hg. Three different methods of analyzing 6-month systolic blood pressure (SBP) results from the SYMPLICITY HTN-3 trial of renal denervation versus a sham procedure.

ANCOVA = analysis of covariance; SYMPLICITY HTN-3 = Renal Denervation in Patients With Uncontrolled Hypertension.

FIGURE 6 Analysis of Covariance for Change in SBP in the SYMPLICITY HTN-3 Trial



Individual 6-month change in systolic blood pressure (SBP) by baseline SBP by treatment group from the SYMPLICITY 3 trial. The **2 drawn parallel regression lines** show the anticipated regression to the mean, that is, patients with a higher baseline SBP (in both treatment groups) tend to have a greater blood pressure reduction compared with those nearer the minimum inclusion criteria of 160 mm Hg. The vertical distance between the 2 regression lines is 4.11 mm Hg, the mean treatment effect adjusted for baseline. SYMPLICITY HTN-3 = Renal Denervation in Patients With Uncontrolled Hypertension.

follow-up, to provide the most valid estimate based on the totality of patient data.

Sometimes, a quantitative outcome has a highly skewed distribution so that a conventional analysis of means becomes unstable because of its dependence on a few extreme values. Options then are: 1) to use a suitable transformation (e.g., natural logarithm leading to comparison of geometric means); 2) to use nonparametric analyses, often focusing on a comparison of medians; or 3) to focus on a particular cut-off value(s) (e.g., the upper limit of normal in liver function tests) with a consequent comparison of percentages.

p VALUES AND THEIR INTERPRETATION

We have deliberately delayed explaining p values until after covering descriptive statistics, estimation, and CIs. This is an attempt to counter the obsessive tendency for people to classify a clinical trial into "positive" or "negative" depending on whether or not the primary endpoint achieves $p < 0.05$. This oversimplification is an abuse of the p value, which can be a valuable statistical tool when interpreted appropriately.

Alongside an estimate of treatment difference and its 95% CI, the corresponding p value is the most succinct direct route to expressing the extent to which it looks plausibly like a real treatment effect, or rather could readily have arisen by chance. At the heart of any significance test is the null hypothesis that the 2 treatments are identical in their effect on the outcome of interest. The p value is the probability of obtaining a treatment difference at least as great (in either direction) as that actually observed if the null hypothesis were true. The smaller the p value, the stronger the evidence against the null hypothesis, that is, the more convincing the evidence is that a genuine treatment difference exists.

Let us consider some recent trials to elucidate the range of their p values for the primary endpoint. In doing so, our aim is to translate statistical evidence into plain English (19). The PARADIGM-HF trial (1) compared a new drug, LCZ696, with enalapril in patients with chronic heart failure. For the primary

endpoint, heart failure hospitalization or cardiovascular death over a median 27 months of follow-up, the hazard ratio was 0.80 with 95% CI: 0.73 to 0.87 and p = 0.0000004. Such a small p value means that if LCZ696 were truly no better than enalapril, the chances of getting this magnitude of treatment difference (or greater) is <1 in a million.

Such a small p value provides overwhelming evidence of a treatment difference. Such proof beyond reasonable doubt means one can confidently assert that LCZ696 is superior to enalapril with regard to the incidence of the primary endpoint.

The CHAMPION-PHOENIX trial (9) of cangrelor versus clopidogrel had an odds ratio of 0.788 with 95% CI: 0.666 to 0.932 for its primary endpoint (**Table 3**). Here, p = 0.005 means that there is a 1 in 200 chance of such a difference (or greater) arising by chance; this is not as thoroughly convincing as the PARADIGM-HF trial, but is still strong evidence of a treatment benefit (i.e., cangrelor appears to be superior to clopidogrel). Note that the trial report (9) gave an adjusted odds ratio and so on; we discuss covariate adjustment in next week's paper.

IMPROVE-IT (Improved Reduction of Outcomes: Vytorin Efficacy International trial) (20) compared ezetimibe with placebo in 18,144 post-ACS patients receiving simvastatin 40 mg. The primary composite endpoint over a mean 5.4 years was cardiovascular death, myocardial infarction, stroke, unstable angina, and coronary revascularization, with hazard ratio of 0.936 (95% CI: 0.888 to 0.988). Here, p = 0.016 means that there is a <1 in 50 chance of such a difference (or greater) arising by chance. This provides some evidence of a treatment benefit: it reaches the oft-used guideline of p < 0.05, that is, statistically significant at the 5% level. There is a 6.4% relative risk reduction and an absolute treatment difference of 2.0%, both with wide CIs. This suggests a modest treatment benefit that is imprecisely estimated.

In the SYMPLICITY HTN-3 trial (17) (**Figure 5**), the mean difference between renal denervation and sham procedure in 6-month change in SBP adjusted for baseline was −4.11 mm Hg, with 95% CI: −8.44 to +0.22 mm Hg, and p = 0.064. Under the null hypothesis that renal denervation is ineffective, this observed magnitude of treatment difference has more than a 1 in 20 chance of occurring. Because p > 0.05 (i.e., 5% significance is not achieved), it is customary to declare that there is insufficient evidence that renal denervation reduces SBP. This should not be interpreted dogmatically that renal denervation has no effect (i.e., the null hypothesis is not necessarily true). Rather, we should declare there is insufficient evidence that renal denervation lowers SBP compared with a sham procedure. It may be that renal denervation has a modest effect OR it may have no effect: the data are inconclusive.

Now, to a more clearly neutral finding. The ASTRONAUT trial (Aliskiren Trial on Acute Heart Failure Outcomes) (21) randomized 1,639 patients with hospitalized heart failure to aliskiren or placebo with a median 11.3 months of follow-up. The primary endpoint, rehospitalization for heart failure or cardiovascular death, had a hazard ratio of 0.92 (95% CI: 0.76 to 1.12), with p = 0.41. With such a clearly nonsignificant p value, there is no evidence that aliskiren has an effect on the primary endpoint. However, we still cannot assert definitively that aliskiren has no effect: the hazard ratio is in the direction of slightly fewer primary events on aliskiren and the wide CI extends a substantial distance from neutrality (hazard ratio: 1) in both directions.

Thus, we may think of p values not as a "black and white" significant/nonsignificant dichotomy, but more in terms of "shades of gray" (22). This analogy to a recent movie is not to make statistics sexy, nor is it to suggest that statisticians are sadists, but it is more in the spirit of the expression's original meaning. The smaller the value of p, the stronger the evidence to contradict the null hypothesis of no true treatment difference. We can think of p < 0.000001 as "pure white" and p = 0.99 as "pure black," with a trend of increasingly darkening grayness in-between those extremes. **Table 5** summarizes a useful vocabulary that might be applied to interpreting p values.

A brief history of the p value and its variety of interpretations is provided in the Online Appendix.

"A p VALUE IS NO SUBSTITUTE FOR A BRAIN." This quote from Stone and Pocock (23) is to remind us all that interpretation of a seemingly "positive" trial rests on more than just a significant p value:

1. It is good practice to give the *actual p value* (i.e., p = 0.042 rather than p < 0.05 or crudely "significant," *or* p = 0.061 rather than "not significant").
2. It is useful to recognize the *link between the p value and the 95% CI* for the treatment difference. If the latter includes no difference, that is, 0 on an absolute scale (e.g., % or mean difference) or 1 on a

TABLE 5 A Useful Language for Interpreting p Values

p < 0.001	Overwhelming evidence
0.001 ≤ p < 0.01	Strong evidence
0.01 ≤ p < 0.05	Some evidence
0.05 ≤ p < 0.10	Insufficient evidence
p ≥ 0.10	No evidence

TABLE 6 The Simplest Statistical Test*

z	p Value
1.64	0.1
1.96	0.05
2.58	0.01
2.81	0.005
3.29	0.001
3.48	0.0005
3.89	0.0001

$\frac{a-b}{\sqrt{a+b}}$ is approximately a standardized normal deviate, z; a and b are the numbers having an outcome event in the 2 treatment groups. This table displays some useful z values. The larger the z, the smaller the p value. *Only suitable for trials with 1:1 randomization. Most reliable when proportions having events are small. Should be confirmed by the more complex test (e.g., log-rank).

TABLE 7 4 Examples Using z

Trial Name (Ref. #)	Patients With an Event (n)		z $\frac{a-b}{\sqrt{a+b}}$	p Value	Interpretation
	Control (a)	New Treatment (b)			
PARADIGM-HF (1)	1,117	914	4.50	<0.00001	Overwhelming evidence
CHAMPION-PHOENIX (9)	322	257	2.70	0.007	Strong evidence
IMPROVE-IT (20)	2,742	2,572	2.33	0.02	Some evidence
ASTRONAUT (21)	214	201	0.64	0.52	No evidence

ASTRONAUT = Aliskiren Trial on Acute Heart Failure Outcomes; IMPROVE-IT = Examining Outcomes in Subjects With Acute Coronary Syndrome: Vytorin (Ezetimibe/Simvastatin) vs Simvastatin (P04103); other abbreviations as in Tables 1 and 3.

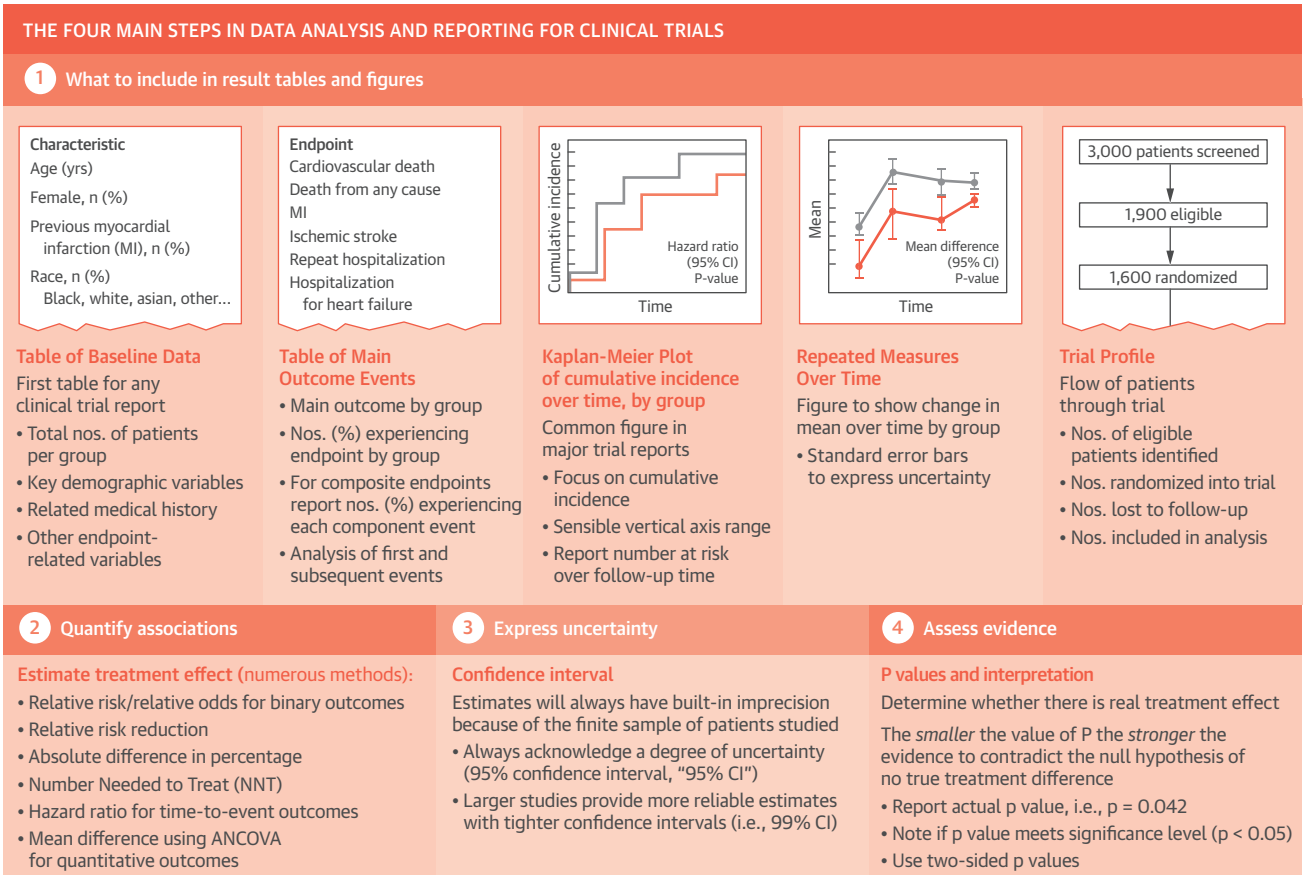
ratio scale (e.g., relative risk or hazard ratio), then we know $p > 0.05$. Conversely, if the 95% CI is wholly 1 side of the null value, then we know $p < 0.05$.

3. It is best if we always use *2-sided p values*. That is, under the null hypothesis, p is the probability of getting a difference *in either direction* as big as (or bigger) than that observed. Occasionally, people will argue that they are only interested in 1 direction of treatment effect (new treatment superior) and, hence, should be allowed to halve the p value in a 1-sided test. For instance, the CoreValve trial (24) claimed 1-sided $p = 0.04$ for lower mortality on transcatheter aortic valve replacement versus surgery, rather than the conventional 2-sided $p = 0.08$. This practice is to be avoided, because it produces an inconsistency across trial reports and makes it a bit too easy to achieve $p < 0.05$.
4. A small p value clarifies that an observed treatment difference appears greater than what could be attributed to chance, but this does not automatically mean that a real treatment effect is occurring. There may be *biases in the study design and conduct* (e.g., randomization could be absent or flawed, lack of appropriate blinding, or incomplete follow-up), which contribute wholly or in part to the apparent treatment difference. These issues contribute to why regulators often require 2 trials be conducted to demonstrate a reassuring consistency of findings in 2 different settings.
5. There is an important distinction between statistical significance and clinical relevance of a treatment effect. Here, the magnitude of treatment difference and its CI are a guide as to whether the benefit of a new treatment is sufficiently great to merit its use in clinical practice.
6. For a small trial to reach a statistically significant treatment effect, the magnitude of treatment difference needs to be very large. For instance, a trial of acetylcysteine versus placebo to prevent contrast-induced nephropathy (25) reported 1 of 41 and 9 of 42 acute reductions in renal failure ($p = 0.01$). This finding has a risk ratio of 0.11 with a very wide 95% CI: 0.015 to 0.859. The observed result is "too good to be true." A comparable small trial with a nonsignificant finding would doubtless not have been published in a major journal. Thus, publication bias, that is, the tendency for published trials to exaggerate treatment effects, is accentuated when trials are small.
7. In this paper, we concentrate on interpreting p values (and CIs) for trials whose purpose is to determine if one treatment is superior to another. For noninferiority trials, with the goal of seeing if a new treatment is as good as the control, interpretation is somewhat different, as explained in the last paper in this series.

THE SIMPLEST STATISTICAL TEST. This paper does not provide the statistical calculations or programs required to obtain p values. Suffice it to say that for the 3 types of outcome data, binary, time-to-event, and quantitative, the most commonly used tests are chi-square, log-rank, and 2-sample Student *t* test or ANCOVA, respectively. But, for trials with 1:1 randomization and a binary or time-to-event outcome, there does exist a quick alternative (26) that can be used by the inquisitive reader who is "dying to know" if a result is statistically significant. The test is so simple that most statisticians do not know about it!

All you need to use is the number of patients in each treatment group who have the primary endpoint. Then, the difference divided by the square root of the sum is approximately a standardized normal deviate, which can readily be converted into a p value (Table 6). See Table 7 for how it works for 4 of the trials we have already

CENTRAL ILLUSTRATION **Practical Guide to the Essentials for Statistical Analysis and Reporting of Randomized Clinical Trials**



Pocock, S.J. et al. J Am Coll Cardiol. 2015; 66(22):2536–49.

ANCOVA = analysis of covariance; CI = confidence interval; NNT = number needed to treat.

discussed. In each case, this simple test agrees well with the more complex calculations used in the trial publication.

CONCLUSIONS

We have covered the essentials of statistical analysis and reporting in this paper, and the key aspects are summarized in the **Central Illustration**. Next week we tackle a variety of more complex statistical challenges that are often faced in the reporting of clinical trials. These include multiplicity of data, covariate adjustment, subgroup analysis, assessing individual benefits and risk, analysis by intention to treat and alternatives, the interpretation of surprises (both good and bad), and enhancing the overall quality of clinical trial reports.

REPRINT REQUESTS AND CORRESPONDENCE: Dr. Stuart J. Pocock, Department of Medical Statistics, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom. E-mail: Stuart.Pocock@LSHTM.ac.uk.

REFERENCES

1. McMurray JJV, Packer M, Desai AS, et al. Angiotensin-neprilysin inhibition versus enalapril in heart failure. N Engl J Med 2014;371:993–1004.

2. Scirica BM, Bhatt DL, Braunwald E, et al. Saxagliptin and cardiovascular outcomes in patients with type 2 diabetes mellitus. N Engl J Med 2013;369:1317–26.

3. Kjekshus J, Apetrei E, Barrios V, et al. Rosuvastatin in older patients with systolic heart failure. N Engl J Med 2007;357:2248–61.

4. Rogers JK, Jhund PS, Perez A-C, et al. Effect of rosuvastatin on repeat heart failure hospitalizations: the CORONA trial (Controlled Rosuvastatin Multinational Trial in Heart Failure). J Am Coll Cardiol HF 2014;2:289–97.

5. Wallentin L, Becker RC, Budaj A, et al. Ticagrelor versus clopidogrel in patients with acute coronary syndromes. N Engl J Med 2009;361:1045–57.

6. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. Lancet 2002;359:1686–9.

7. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

8. Shahzad A, Kemp I, Mars C, et al. Unfractionated heparin versus bivalirudin in primary percutaneous coronary intervention (HEAT-PPCI): an open-label, single centre, randomised controlled trial. Lancet 2014;384:1849–58.

9. Bhatt DL, Stone GW, Mahaffey KW, et al. Effect of platelet inhibition with cangrelor during PCI on ischemic events. N Engl J Med 2013;368:1303–13.

10. Bulpitt CJ. Confidence intervals. Lancet 1987;329:494–7.

11. Altman D, Machin D, Bryant T, Gardner M. Statistics with confidence: confidence intervals and statistical guidelines. Hoboken, NJ: John Wiley & Sons, 2013.

12. Clark T, Bradburn M, Love S, Altman D. Survival analysis part I: basic concepts and first analyses. Br J Cancer 2003;89:232.

13. Farkouh ME, Domanski M, Sleeper LA, et al. Strategies for multivessel revascularization in patients with diabetes. N Engl J Med 2012;367:2375–84.

14. Wei L. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med 1992;11:1871–9.

15. Zannad F, McMurray JJV, Krum H, et al. Eplerenone in patients with systolic heart failure and mild symptoms. N Engl J Med 2011;364:11–21.

16. Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. BMJ 2001;323:1123–4.

17. Bhatt DL, Kandzari DE, O'Neill WW, et al. A controlled trial of renal denervation for resistant hypertension. N Engl J Med 2014;370:1393–401.

18. Senn S. Change from baseline and analysis of covariance revisited. Stat Med 2006;25:4334–44.

19. Pocock SJ, Ware JH. Translating statistical findings into plain English. Lancet 2009;373:1926–8.

20. Cannon CP, Blazing MA, Giugliano RP, et al. Ezetimibe added to statin therapy after acute coronary syndromes. N Engl J Med 2015;372:2387–97.

21. Gheorghiade M, Böhm M, Greene SJ, et al. Effect of aliskiren on postdischarge mortality and heart failure readmissions among patients hospitalized for heart failure: the ASTRONAUT randomized trial. JAMA 2013;309:1125–35.

22. Sterne JA, Davey Smith G. Sifting the evidence—what's wrong with significance tests? BMJ 2001;322:226–31.

23. Stone GW, Pocock SJ. Randomized trials, statistics, and clinical inference. J Am Coll Cardiol 2010;55:428–31.

24. Adams DH, Popma JJ, Reardon MJ, et al. Transcatheter aortic-valve replacement with a self-expanding prosthesis. N Engl J Med 2014;370:1790–8.

25. Tepel M, van der Giet M, Schwarzfeld C, et al. Prevention of radiographic-contrast-agent-induced reductions in renal function by acetylcysteine. N Engl J Med 2000;343:180–4.

26. Pocock SJ. The simplest statistical test: how to check for a difference between treatments. BMJ 2006;332:1256–8.

KEY WORDS clinical trials, interpretation, publication, statistics

APPENDIX For supplemental material, please see the online version of this article.