Journal of Clinical Epidemiology 120 (2020) 59–66

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

An approach to explore for a sweet spot in randomized trials

Donald A. Redelmeier[a,b,c,d,e,*], Robert J. Tibshirani[f,g]

[a]Department of Medicine, University of Toronto, Toronto, Ontario, Canada
[b]Evaluative Clinical Sciences Department, Sunnybrook Research Institute, Toronto, Ontario, Canada
[c]Population and Global Health Department, Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
[d]Division of General Internal Medicine, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada
[e]Center for Leading Injury Prevention Practice Education & Research, Toronto, Ontario, Canada
[f]Department of Biomedical Data Sciences, Stanford University, Stanford, CA, USA
[g]Department of Statistics, Stanford University, Stanford, CA, USA

Accepted 12 December 2019; Published online 23 December 2019

Abstract

Objective: The objective of the study was to demonstrate how a conventional randomized trial can be analyzed through a stratified or a matched approach to identify a potential sweet spot where observed differences might be accentuated in the mid range of disease severity.

Design and Setting: We review a landmark randomized trial of heart failure patients that tested whether implantable defibrillators reduce mortality (n = 2,521).

Results: Overall, 22% (182/829) of the patients in the defibrillator group died compared with 29% (484/1,692) of patients in the control group. Proportional hazards analysis yielded a modest 25% survival benefit (hazard ratio = 0.75, 95% confidence interval: 0.63 to 0.89). Stratified analysis of the trial yielded a larger 52% survival benefit for those in the middle quintile of disease severity (hazard ratio = 0.48, 95% confidence interval: 0.29 to 0.79). In contrast, little of the survival benefit was explained by patients with the greatest disease severity (hazard ratio = 0.89, 95% confidence interval: 0.69 to 1.15). The discrepancy between crude and stratified analyses could be visualized by graphical displays and replicated with matched comparisons.

Conclusion: Our approach for analyzing a randomized trial could help identify a potential sweet spot of an accentuated treatment effect.

Keywords: Clinical trials; Sudden death; Heterogeneous treatment effect; Cardiac defibrillator; Patient diversity; Precision medicine

1. Introduction

Randomized trials are the gold standard for clinical research, but precious metals sometimes need polishing. A large strength of randomized trials is their simplicity in design, analysis, and reporting. Nothing matches the sublime logic and ease of interpreting a randomized trial [1]. A major weakness of randomized trials is the diverse set of biases that slant results toward the null. These include insufficient statistical power, fallible patient adherence, brief duration of follow-up, faulty dose selection, outcome ascertainment error, and the play of random chance [2]. Moreover, randomized trials that focus on time-to-event outcomes (also termed survival analysis) are plagued by even rudimentary debates such as whether to calculate adjusted or unadjusted estimates of treatment effect sizes to account for baseline patient characteristics [3].

Randomized trials often require a large sample size to achieve adequate statistical power for an infrequent

Funding: This project was supported by a Canada Research Chair in Medical Decision Sciences, the Canadian Institutes of Health Research, and the BrightFocus Foundation. The views expressed are those of the authors and do not necessarily reflect the Ontario Ministry of Health and Long-term Care.

Conflict of interest: The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. All authors have no financial or personal relationships or affiliations that could influence the decisions and work on this manuscript.

Data sharing: The deidentified data collected for this study are available in an appendix included at the time of original manuscript submission and also are available following publication for researchers whose proposed use of data has been approved by an independent review committee.

Accountability: The lead author (D.A.R.) had full access to all the data in the study, takes responsibility for the integrity of the data, and is accountable for the accuracy of the analysis.

* Corresponding author. Sunnybrook Health Sciences Centre, G-151 2075 Bayview Ave, Toronto, Ontario, Canada M4N 3M5, Tel.: (+416) 480-6999; fax: (+416) 480-6048.

E-mail address: dar@ices.on.ca (D.A. Redelmeier).

https://doi.org/10.1016/j.jclinepi.2019.12.012
0895-4356/

What is new?

- Randomized trials recruit diverse patients where extreme cases of disease severity may be unresponsive to the treatment and render inconclusive overall statistical results.

Key findings

- We propose estimating disease severity through predilection scores of the natural history derived from the control patients.

What this adds to what was known?

- Methods for identifying a potential sweet spot of patients responsive to treatment can then be obtained by stratification or by matching based on disease severity.
- The methods can work automatically in any randomized trial and require no additional information, data collection, computer software, or investigator judgment.

What is the implication and what should change now?

- Such methods for identifying a potential sweet spot can also help check whether a negative trial truly excludes a meaningful effect.

outcome. A strategy of recruiting a broad range of patients also helps bolster generalizability by presuming all patients will be similarly responsive to treatment (albeit with different baseline risks). However, recruitment may include some patients who have self-limited illnesses where treatment is superfluous. In addition, recruitment may enroll other patients who have lethal combinations where treatment is futile [4]. Both of these extreme groups will be unresponsive to treatment and undercut statistical power. The net consequence means an analysis can be biased toward the null because the overall trial results hinge on a subgroup of patients in the middle range of disease severity who are responsive to care (sweet spot) [5].

Personalized precision medicine presumes the relative effectiveness of a treatment might vary substantially in diverse patients. Here we provide an approach to identifying such diversity based on the assumption that treatment responsiveness is related to disease severity. The strategy is to explore differential relative responsiveness in a randomized trial by applying stratification or matching [6]. The main drawback is the increased sophistication needed to avoid misinterpreting an unfamiliar approach. To do so, we examine a published landmark randomized trial to help demonstrate this approach [7]. Except where noted, we use the proportional hazards model throughout as the accepted approach for time-to-event outcomes [8]. Our approach explores how randomized trials might underestimate effectiveness and is generalizable to binary or continuous outcomes.

2. Methods

2.1. Background trial

The Sudden Cardiac Death in HEart Failure Trial (SCD-HEFT) was a study of adults (age $\geq$ 18 years) diagnosed with heart failure (New York Heart Association class II or III) from impaired cardiac function (ejection fraction $\leq$ 35%) [7]. The study was conducted between September 16, 1997 and July 18, 2001, patients were followed until October 31, 2003, the analysis was by intent-to-treat principles, and the main outcome was all-cause mortality. By randomized assignment, one-third of patients received an implantable defibrillator and the remaining two-thirds received medical management only (ClinicalTrials.gov: NCT00000609). The trial found that defibrillator treatment led to a modest reduction in overall mortality. The accompanying editorial affirmed the role of defibrillator therapy, cautioned the benefit was smaller than observed in earlier studies, and raised concerns about cost-effectiveness [9,10].

2.2. Diversity and the sweet spot

The SCD-HEFT study exemplifies the diversity of patients in clinical research. In this study, for example, half were younger and half were older than age 60 years. Similarly, cardiac ejection fraction, renal function, and many other characteristics showed wide ranges of baseline characteristics. Presumably, the unmeasured disease determinants were also variable. This means that some patients, regardless of care, might have been prone to poor outcomes and others might have been destined to do well. As a consequence, the primary statistical analysis relied on a subgroup of patients where overall disease severity was neither too high nor too low, hereafter defined as the ‘‘sweet spot.’’ [11]. This subgroup can be hard to determine in advance and may become potentially outnumbered by efforts to recruit sufficient sample size or maximize generalizability [12,13].

2.3. Primary analysis

Patient diversity is not usually emphasized in clinical research. Instead, statistical tests rely on the principle that the mean baseline value is nearly identical in groups of randomized patients. This leads to a straightforward analysis where diversity is expressed as a standard deviation estimate and minimized as a standard error equivalent. The statistical method is often declared in advance to avoid a proliferation of spurious comparisons; for example, a standard t-test can be remarkably robust to latent measurement errors when the main outcome is a continuous variable

(such as six-minute walking distances) [14]. Unfortunately, the proportional hazards approach (similar to logistic regression) is a nonlinear model and may be biased to the null by patient diversity when the main outcome is a time-to-event measurement (such as survival) [15].

2.4. Baseline predilection

A different way to consider patient diversity is by introducing an outcome predilection score (not to be confused with a treatment propensity score) [16]. Unlike a standard risk index [17], a prediction score is tailored to a specific trial, requires no external validation, and can span beyond a range of 0.0 to 1.0. To do so, fit a proportional hazards model to the control patients and calculate each patient's individual baseline prognosis from the coefficients (akin to estimating their predilection to the outcome). This predilection score needs to be estimated solely from the control group because an effective intervention may otherwise change the natural history. The resulting predilection score can then characterize each patient (defibrillator or control) according to their severity of disease (and might also appear in a technical appendix or website application). The primary benefit is to ultimately assemble clusters of defibrillator and control patients who have similar baseline predilection [18].

2.5. Subgroup stratification

One approach to forming clusters of similar patients is to stratify based on baseline predilection score. For example, examine the full distribution on predilection scores among patients and create progressive quintiles of nearly equal size (denoted as "least," "lesser," "middle," "greater," "greatest"). This allows testing the apparent survival advantage from defibrillator treatment in each quintile separately, with particular attention to the middle, least, and greatest quintiles. In addition, separate survival comparisons can be visualized according to predilection quintile. Such stratified approaches are a classic method for addressing a single predictor that does not satisfy the proportionate hazards assumption and can be a feasible method to identify a potential sweet-spot hypothesis in a randomized trial [19].

2.6. Individually matched sets

A different approach to forming clusters of similar patients involves converting the randomized trial into a matched randomized trial. For example, apply a matching algorithm to assemble matched sets where some patients received a defibrillator, other patients did not receive a defibrillator, and all patients had a similar baseline predilection score. In this example due to the 1:2 randomization, a natural approach is to use a matching algorithm to form triplets (not duets) where 1 patient received a defibrillator, 2 did not receive a defibrillator, and all 3 had similar predilection scores [20]. The strength is that the matching extends beyond quintiles (five-level stratification) to yield hundreds of matched sets (finer stratification). The matched sets can then test the observed reduction in mortality from defibrillator treatment across a range of disease severity.

3. Results

3.1. Summary data

The SCD-HEFT randomized trial tested heart failure patients (n = 2,521) who were followed for a median duration of 3.8 years. One-third received a defibrillator and the remaining two-thirds did not. Additional characteristics were also recorded at baseline with reasonable symmetry between the two groups and substantial variation within each group (Table 1). For example, all patients underwent angiography, about half had significant stenosis of at least one major coronary artery, and the remaining half had no major coronary artery stenosis. Subsequent results showed a lower final patient mortality in the defibrillator group (182/829 = 22%) compared with the control group (484/1,692 = 29%). This difference was easily visualized (Figure 1) and equal to a one-quarter relative reduction (hazard ratio = 0.75, 95% confidence interval: 0.63 to 0.89).

Many baseline characteristics independently predicted mortality. Proportional hazards regression based on the control patients identified and quantified separate risk factors and subsequently derived a predilection score for individual patients (Appendix). Overall, the predilection scores showed plausible patterns and moderate goodness-of-fit (C-statistic = 0.71); for example, the mean predilection score was higher for patients who had diabetes and a major coronary stenosis than for patients who had no diabetes and no major coronary stenosis ($P < 0.001$). This same equation and coefficients then yielded a predilection score for individual defibrillator patients. As a check of randomization, the distribution of predilection scores showed excellent overall symmetry comparing the group of defibrillator patients to the group of controls (Figure 2). A further check of randomization is also possible by recreating Table 1 for each stratum.

3.2. Stratified approach

The predilection scores next generated stratified analyses that identified how the observed survival benefit varied across different quintiles of disease severity. The rationale was to examine whether some patients were highly responsive to treatment (sweet-spot analysis) and whether other patients were relatively unresponsive to treatment (and undercut statistical power). We found that defibrillator therapy resulted in a substantial relative risk reduction in mortality for patients in the middle quintile of disease severity (hazard ratio = 0.48, 95% confidence interval: 0.29 to 0.79). In contrast, defibrillation therapy yielded a negligible survival benefit for patients at the greatest disease severity (Figure 3). Together, this pattern

Table 1. Baseline patient characteristics

Defibrillator patients	Defibrillator (n = 829)	Control patients (n = 1,692)
Demographic features		
Mean age (y)	59.4 (11.9)	58.8 (11.9)
Sex (male)	639 (77)	1,294 (77)
Race (Caucasian)	640 (77)	1,292 (76)
Clinical characteristics		
New York Heart Association (class II) [a]	566 (68)	1,195 (71)
Mean ejection fraction	23.6 (7.0)	24.0 (6.9)
Major coronary stenosis [b]	430 (52)	875 (52)
Diabetes mellitus	253 (31)	514 (3)
Chronic lung disease	175 (21)	305 (1)
Hyperlipidemia	431 (52)	898 (53)
Hypertension	453 (55)	947 (56)
Syncope history	52 (6)	110 (7)
Atrial fibrillation	141 (17)	249 (15)
Ventricular tachycardia	210 (25)	373 (22)
Prescribed medications		
ACE inhibitor [c]	783 (94)	1,649 (97)
Beta blocker	576 (69)	1,162 (69)
Digoxin	552 (67)	1,203 (71)
Thiazide diuretic	63 (8)	112 (7)
Loop diuretic	676 (82)	1,388 (82)
Potassium sparing diuretic	168 (20)	339 (20)
Statin [d]	312 (38)	653 (39)
ASA [e]	477 (58)	938 (55)
Warfarin	266 (32)	591 (35)
Physical examination		
Mean weight (pounds) [f]	194.3 (44.3)	193.7 (44.0)
Mean heart rate (beats/minute)	74.9 (13.7)	74.6 (13.9)
Mean systolic blood pressure (mm Hg)	119 (20)	120 (19)
Mean diastolic blood pressure (mm Hg)	71 (12)	71 (11)
Mean serum sodium (mmol/L)	139 (3)	139 (3)
Mean creatinine (mg/dL)	1.2 (0.4)	1.2 (0.7)

Data are count (percentage) of column unless indicated otherwise as mean (standard deviation).

[a] Class II for mild dyspnea with ordinary activity and no marked limitations in activities.

[b] Major stenosis denotes >75% narrowing of at least one major coronary artery.

[c] ACE for angiotensin-converting enzyme agent and includes angiotensin receptor blocker.

[d] Statin for HMG-CoA reductase inhibitor prescribed as a lipid-lowering agent.

[e] ASA for acetylsalicylic acid as a platelet aggregation inhibitor.

[f] To convert pounds to kilograms, multiply by 0.454.

suggested the relative effects of defibrillator therapy on patient survival depend on baseline characteristics.

The survival advantage with defibrillator treatment was sufficiently clear that it could be confirmed by simply classifying patients as dead or alive at study termination.

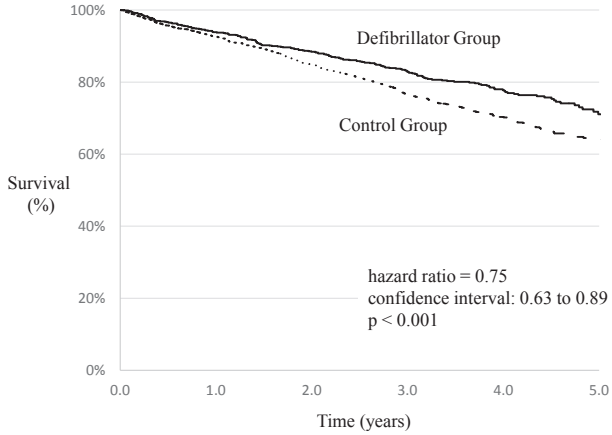

Fig. 1. All-cause mortality for full cohort. Kaplan-Meier plot of survival from day of randomization. X-axis shows time in years spanning to a maximum of 5 years. Y-axis shows proportion surviving at each time point. *P*-value based on logrank test. Results show greater all-cause mortality in control compared with the defibrillator group.

Overall, 22% (n = 182) of the defibrillator patients died, whereas 29% (n = 484) of the control patients died, indicating an absolute survival advantage of 7% (n = 55) for the defibrillator patients. The absolute survival advantage was mostly due to patients with the middle disease severity (Figure 4). None of the survival advantage was explained

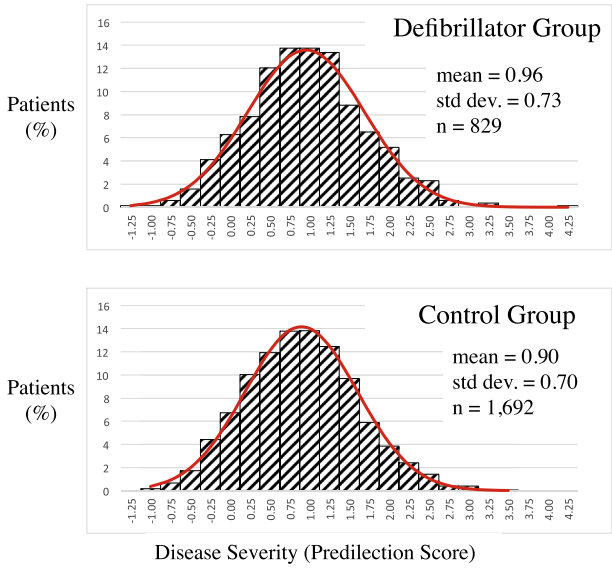

Fig. 2. Distribution of predilection scores. Histograms of predilection scores for patients. Upper panel shows defibrillator group (n = 829) and lower panel shows control group (n = 1,692). X-axis denotes predilection score calculated for each patient at baseline (predilection equation coefficients and structure based on proportional hazards model derived from control patients). Y-axis shows count of patients with corresponding predilection scores. Additional curve based on mean and standard deviation shown from normal approximation. Results show good similarity in predilection scores comparing the defibrillator group to the control group with reasonable congruence to normal distribution for each group.

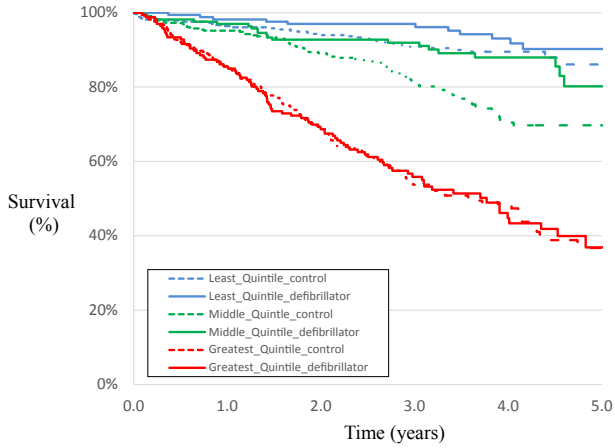

Fig. 3. Survival in different predilection quintiles. Stratified Kaplan-Meier plots from different predilection quintiles. X-axis shows time spanning from day of randomization to a maximum of 5 years. Y-axis shows proportion surviving at each time point. Least hazardous quintile in blue, greatest hazardous quintile in red, middle quintile in green, and intermediate quintiles not shown for simplicity. Solid lines denote the defibrillator group and dashed lines denote the control group. Results show modest survival benefit from defibrillator in least hazardous quintile, negligible survival benefit in greatest hazardous quintile, and accentuated survival benefit in middle quintile. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

by patients with the greatest disease severity, and little was explained by patients with the least disease severity. This latter analysis that ignored the timing of death could be subjected to pairwise tests of statistical significance and generalized tests for interactions based on individual patient data.

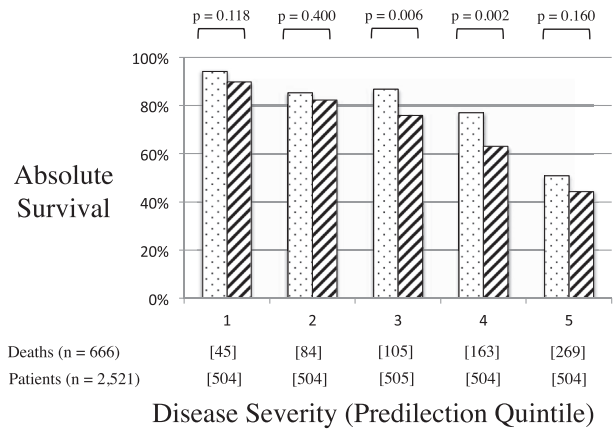

Fig. 4. Survival in different predilection quintiles. Stratified analysis from different predilection quintiles. Data show proportion of each group alive at study termination. Speckled bars for the defibrillator group, striped bars for the control group, and p-values for direct comparison. Square brackets denote total number of patients and total number of deaths observed in each quintile. Least hazardous quintile on left, most hazardous quintile on right, and intermediate quintiles in intermediate position. Results show reduced greater survival with defibrillator treatment for each quintile, accentuated benefits in middle quintile, and more modest benefits in greatest quintile.

3.3. Matched approach

We also formed matched triplets composed of one patient randomized to defibrillator therapy and two patients randomized to control treatment, of whom all three had similar predilection scores. To do so, we used the greedy matching algorithm with a caliper width of 0.2 so that complete triplets were created for most patients (total patients = 2,487). In particular, we retained 100% of the defibrillator patients (n = 829), 98% of the control patients (n = 1,658), and each triplet was 100% complete (no incomplete sets). Results for the matched patients again showed a lower overall patient mortality in the defibrillator group (182/829 = 22%) compared with the control group (479/1,658 = 29%), equal to a one-quarter relative reduction (hazard ratio = 0.75, 95% confidence interval: 0.63 to 0.89).

The survival advantage with defibrillator treatment could also be confirmed again by simply classifying patients as dead or alive at study termination in each triplet. Overall, 661 of the 2,487 matched patients had died after a median of 3.4 years, indicating a net survival advantage of 57 defibrillator patients (95% confidence interval: 32 to 80). The cumulative survival advantage was mostly due to patients with the middle range of disease severity (Figure 5). None of the survival advantage was explained by patients with the greatest disease severity and little was explained by patients with the least disease severity. This latter analysis that ignored the timing of death could be subjected to a global test of statistical significance following sigmoid models for biology growth and tests of statistical significance [21].

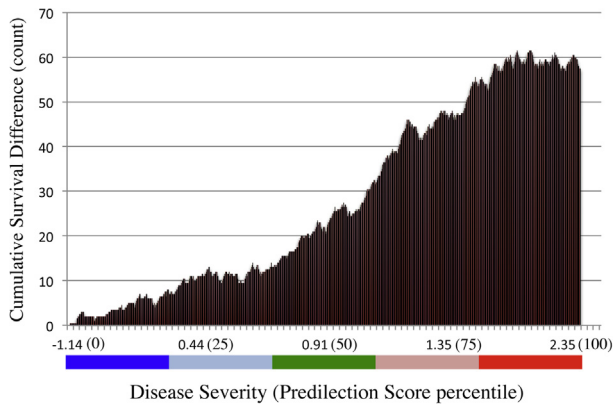

Fig. 5. Cumulative survival advantage according to predilection score. Histogram of cumulative survival advantage comparing defibrillator patients to control patients. X-axis shows consecutive triplets of patients matched on predilection score and sequenced by ordinal rank of increasing baseline tendency of mortality. Y-axis shows cumulative count of survival advantage for defibrillator patients. Cumulative rather than marginal data plotted to display trend. Display contains 829 bars for 829 matched triplets (one defibrillator patient and two control patients in each triplet, all with similar predilection score). X-axis scaled so predilection scores spaced by percentile with corresponding quintiles also shown by color. Results show survival advantage of defibrillator patients mostly explained by individuals with intermediate predilection scores.

The analysis of matched patients also allowed a more nuanced analysis that accounted for the duration of observation and censoring when testing the increased survival for defibrillator patients. Specifically, matched triplets were scored as ambiguous if all patients were alive. Conversely, matched triplets where all patients died were accordingly scored as superior for the defibrillator patient, inferior for the defibrillator patient, or mixed for the defibrillator patient (longer than one control and shorter than the other control). Other patterns were scored depending on which and when patients died. This scored analysis accounting for the timing of death again showed the survival advantage was mostly due to patients with middle disease severity and was again difficult to attribute to chance ($P < 0.001$).

4. Discussion

In this study, we introduce a predilection score for disease severity to explore a potential sweet spot in a randomized trial. We illustrate the approach using a randomized trial of heart failure patients treated with a defibrillator to reduce mortality. Our main finding is that analyses based on disease severity stratifying patients tended to yield a greater estimate of effectiveness than basic analyses that did not account for patient diversity. Under the null condition when treatment is ineffective, a sweet-spot analysis and a basic analysis will tend to yield identical point estimates (yet different precisions). When patients show an important degree of diversity, however, a sweet-spot analysis will generally yield results that are more extreme than a basic analysis.

The economics literature provides other methods for identifying heterogeneous treatment effects [22–24]. These methods tend to use a supervised learning approach that randomly divides the data into a training set and a validation set. Statistical models are then fitted to the training set and checked in the validation set to further adjust the parameters. Of course, sample splitting weakens the power for identifying a sweet spot. Moreover, an effective treatment may reduce power further by changing the patient's outcome. Our approach requires no data splitting because the fitting is essentially unsupervised and does not involve adjusting parameters from data in the treatment arm. Our approach is also fully automatable for machine learning algorithms.

Randomized trials sometimes include a risk score to explore differential treatment effects [25]. This approach, however, is not feasible in many important settings because it requires identifying ex-ante an established, validated, and accepted risk score [26]. When available, furthermore, a risk score may not be updated to reflect modern care, extend to stringently selected trial patients, map exactly to the trial end point, or aligned to the observed length of trial follow-up [27]. In addition, a risk score approach can be unreliable because of post hoc cut-points, group stratification, and underpowered Mantel-Haenszel interaction tests [28]. A matched analysis using predilection scores avoids these limitations and yields a more powerful method for identifying a potential sweet spot [29].

Our approach to identifying a sweet spot has several limitations regardless of whether a stratified or a matched analysis is followed. The most important limitation is the assumption that treatment responsiveness is correlated with disease severity: this assumption is difficult to establish before the study is completed. A related downside is the general preference for statistical simplicity expressed by granting agencies, medical journals, government regulators, and practicing clinicians [30]. A further weakness is that the statistical analysis will not identify (or correct) a failure of randomization or other fundamental flaw [31,32]. Together, these theoretic reasons may explain why future randomized trials may remain hesitant when checking for a sweet spot.

The extra work for a sweet-spot analysis of a randomized trial may not be worth the effort unless three conditions apply. First, the patient sample must have substantial underlying clinical diversity, unlike animal experiments that involve genetically identical laboratory mice. Second, salient information must be available on each individual patient, unlike anonymous surveys that recruit undergraduate students or internet volunteers. Third, plentiful sample size and outcome counts must be available among controls to support meaningful multivariable modeling, unlike small trials of surgical techniques. The importance of these three conditions may explain why sweet-spot analyses of a randomized trial have been rarely conducted despite having positive potential [33,34].

A sweet-spot analysis of a randomized trial also has practical drawbacks. Secondary analyses can be prone to misinterpretation because of the potential for multiple hypothesis testing and capitalizing on chance [35–37]. This concern about spurious *P*-values can be extreme if a sweet spot is clinically implausible [38]. The available covariates must be sufficiently important so the matching accounts from important prognostic factors and, therefore, is informative. An easy visual display of summary data can become cluttered because of the lost simplicity of a two-group contrast. The team running a trial, moreover, may lack sufficient resources for statistical analysis after devoting primary attention toward patient recruitment, intervention implementation, and data collection.

A sweet-spot analysis also has advantages relative to some alternative approaches to randomized trials. Forest plots for multiple separate patient characteristics raise a proliferation of type 1 and type 2 statistical errors, whereas a predilection score is a single stratification with potentially fewer spurious findings [39]. Highly restrictive patient selection criteria can lead to reduced external generalizability, whereas stratification is easy to conduct and explain. Adaptive randomization or Bayesian analysis raises worries about complexity and tampering, whereas stratification can be set in advance, audited in retrospect, and blinded for the entire trial [40]. A stratification analysis is also more rigorous than editorial claims about possible "Goldilocks" patients [41–43].

Another advantage of sweet-spot analysis is the ease of rechecking a negative trial. For example, ICE-PACS (ClinicalTrials.gov: NCT01528475) was a randomized trial of prehospital cooling for adults after return of spontaneous circulation after a cardiac arrest (n = 582) [44]. Primary analysis showed no significant increase in the frequency of achieving target temperature (relative increase = 1.17; 95% confidence interval: 0.91 to 1.52; P = 0.22). Sweet-spot analysis yielded similar findings for those in the middle quintile of severity (relative increase = 1.25, 95% confidence interval: 0.53 to 2.92). Similarly, no significant benefit was observed at the extremes. Together, this sweet-spot analysis confirms the robustness of the primary findings.

A conservative analysis in randomized trials is often defended to compensate for other biases that slant studies toward positive conclusions. Examples of biases include skewed patient recruitment, inattention to adverse events, mismatched ascertainment over time, differential early or late effects, and fallible follow-up. An unmatched analysis of randomized trials, therefore, yields a conservative result that helps avoid exaggerations. Yet two wrongs do not make a right because patient diversity could mask a key nuance [45]. Moreover, a proliferation of noninferiority trials means faulty analyses may reinforce faulty study designs [46]. We suggest, therefore, a sweet-spot analysis may be useful as a secondary analysis before deciding whether a randomized trial is really inconclusive.

Acknowledgments

The authors thank Peter Austin, Paul Dorian, Michael Fralick, Gordon Guyatt, Daniel Kahneman, Lauren Lapointe-Shaw, Fizza Mazoor, Ruxandra Pinto, Sharon Reece, Damon Scales, Therese Stukel, Stefan Wager, and Jonathan Zipursky for helpful suggestions on specific points.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.12.012.

References

[1] Kroenke K, Monahan PO, Kean J. Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. J Clin Epidemiol 2015;68:1085−92.
[2] Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. J Clin Epidemiol 2018;98:89−97.
[3] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment andbaseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002;21:2917−30.
[4] Peto R, Baigent C. Trials: the next 50 years. Large scale randomised evidence of moderate benefits. BMJ 1998;317:1170−1.
[5] Redelmeier DA, Tibshirani RJ. Methods for analyzing matched designs with double controls: excess risk is easily estimated and misinterpreted when evaluating traffic deaths. J Clin Epidemiol 2018;98:117−22.
[6] Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. BMJ 2010;340:c2289.
[7] Bardy GH, Lee KL, Mark DB, Poole JE, Packer DL, Boineau R, et al, For the Sudden Cardiac Death in Heart Failure Trial (SCD-HeFT) Investigators. Amiodarone or an implantable cardioverter-defibrillator for congestive heart failure. N Engl J Med 2005;352:225−37.
[8] Tibshirani RJ. A plain man's guide to the proportional hazards model. Clin Invest Med 1982;5(1):63−8.
[9] Kadish A. Prophylactic defibrillator implantation–toward an evidence-based approach. N Engl J Med 2005;352:285−7.
[10] Mark DB, Nelson CL, Anstrom KJ, Al-Khatib SM, Tsiatis AA, Cowper PA, et al. Cost-effectiveness of defibrillator therapy or amiodarone in chronic stable heart failure: results from the Sudden Cardiac Death in Heart Failure Trial (SCD-HeFT). Circulation 2006;114:135−42.
[11] Penston J. Large-scale randomised trials–a misguided approach to clinical research. Med Hypotheses 2005;64(3):651−7.
[12] Drake MT. Vitamin D and the goldilocks principle: too little, too much, or just right? J Clin Endocrinol Metab 2014;99:1164−6.
[13] Pinto DS, Grandin EW. Risk prediction in AMI shock: goldilocks and the search for "just right. J Am Coll Cardiol 2017;69:1921−3.
[14] Redelmeier DA, Bayoumi AM, Goldstein RS, Guyatt GH. Interpreting small differences in functional status: the Six Minute Walk test in chronic lung disease patients. Am J Respir Crit Care Med 1997;155:1278−82.
[15] Tibshirani RJ, Ciampi A. A family of proportional- and additive-hazards models for survival data. Biometrics 1983;39:141−7.
[16] Hansen BB. The prognostic analogue of the propensity score. Biometrika 2008;95(2):481−8.
[17] Stuart EA, Lee BK, Leacy FP. Prognostic score−based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J Clin Epidemiol 2013;66:S84−90.
[18] Rubin DB. Matching to remove bias in observational studies. Biometrics 1973;29:159−83.
[19] Lesko CR, Henderson NC, Varadhan R. Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 2018;100:22−31.
[20] Austin C. A comparison of 12 algorithms for matching on the propensity score. Stat Med 2014;33:1057−69.
[21] Yin X, Goudriaan J, Lantinga EA, Vos J, Spiertz HJ. A flexible sigmoid function of determinate growth. Ann Bot 2003;91(3):361−71.
[22] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 2018;113:1228−42.
[23] Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci U S A 2016;113:7353−60.
[24] Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A 2019;116(10):4156−65.
[25] Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. Stat Methods Med Res 2009;18(1):67−80.
[26] Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G, et al. The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. JAMA 2000;284:835−42.
[27] Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol 2016;45:2075−88.
[28] Schmidt AF, Groenwold RH, Knol MJ, Hoes AW, Nielen M, Roes KC, et al. Exploring interaction effects in small samples increases rates of false-positive and false-negative findings: results

from a systematic review and simulation study. J Clin Epidemiol 2014;67:821−9.
[29] Redelmeier DA, Tibshirani RJ. A simple method for analyzing matched designs with double controls: McNemar's test can be extended. J Clin Epidemiol 2017;81:51−5.
[30] Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. J Clin Epidemiol 2010;63:142−53.
[31] Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. Lancet 1990;335:149−53.
[32] Senn S. Testing for baseline balance in clinical trials. Stat Med 1994;13:1715−26.
[33] Sussman JB, Kent DM, Nelson JP, Hayward RA. Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of Diabetes Prevention Program. BMJ 2015;350:h454.
[34] Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med 2019.
[35] Senn SJ. Covariate imbalance and random allocation in clinical trials. Stat Med 1989;8:467−75.
[36] Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004;57:229−36.
[37] Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010;11:85.
[38] Parker AB, Naylor CD. Interpretation of subgroup results in clinical trial publications: insights from a survey of medical specialists in Ontario, Canada. Am Heart J 2006;151:580−8.
[39] Fan J, Song F, Bachmann MO. Justification and reporting of subgroup analyses were lacking or inadequate in randomized controlled trials. J Clin Epidemiol 2019;108:17−25.
[40] Bhatt DL, Mehta C. Adaptive designs for clinical trials. N Engl J Med 2016;375:65−74.
[41] Jhun P, Jhun K, Wei D, Herbert M. The neonatal airway and the goldilocks phenomenon. Ann Emerg Med 2017;69(2):167−70.
[42] Chappell R, Branch LG. The Goldilocks dilemma in survey design and its solution. J Clin Epidemiol 1993;46:309−12.
[43] Oliver JD, Chaudhry A, Vyas KS, Manrique OJ, Martinez-Jorge J. Aesthetic Goldilocks mastectomy and breast reconstruction: promoting its use in the ideal candidate. Gland Surg 2018;7(5):493−5.
[44] Scales DC, Cheskes S, Verbeek PR, Pinto R, Austin D, Brooks SC, et al. Strategies for Post-Arrest Care SPARC Network. Prehospital cooling to improve successful targeted temperature management after cardiac arrest: a randomized controlled trial. Resuscitation 2017;121:187−94.
[45] Kent DM, van Klaveren D, Paulus JK, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement: Explanation and Elaboration. Ann Intern Med 2019.
[46] Hong J, Tung A, Kinkade A, Tejani AM. Noninferiority drug trials fail to report adequate methodological detail: an assessment of noninferiority trials from 2010 to 2015. J Clin Epidemiol 2019;108:144−6.