



# Large-scale prevention trials could provide stronger evidence for decision-makers: Opportunities to design and report with a focus on the benefit–harm balance

Hélène E Aschmann<sup>1,2</sup> , John J McNeil<sup>3</sup> and Milo A Puhan<sup>1,4</sup>

Randomized controlled trials (RCTs) are typically designed for a single specific primary outcome, and their reporting focuses on this outcome. Large-scale trials, prevention studies in particular, could yield more valuable and relevant evidence by focusing on the benefit–harm balance. These results would be more helpful for informing guideline development, health policy and individual treatment decisions. Currently, RCTs are often inconclusive and sometimes misleading with respect to the balance of benefits (i.e. efficacy outcomes) and harms (e.g. side effects or treatment burden) of interventions, which is ultimately of interest to decision-makers (such as patients, healthcare providers or guideline developers). Little of the recent progress in methods to evaluate the balance of benefits and harms of interventions has been applied to improve the design and reporting of RCTs. There is still a lack of consensus on how to design RCTs specifically to address the benefit–harm balance and contribute better evidence for decision-making. We propose three fundamental changes to the current practice to maximize how much RCTs, in particular prevention trials, can increase the certainty when estimating the benefit–harm balance of interventions.

First, we call for defining the benefit–harm balance as the primary aim of large-scale prevention RCTs. Since the primary outcome sets the focus in how the results should be interpreted,<sup>1</sup> the common practice of selecting single benefit outcomes as primary outcomes generally emphasizes benefits and reduces the RCTs' value for interpreting the benefit–harm balance. For example, after an expert panel decided that a top priority question was whether a lower systolic blood pressure target reduces cardiovascular events more than a standard target in people with hypertension and without diabetes,<sup>2</sup> cardiovascular benefits were designated as the primary aim. Accordingly, the Systolic Blood Pressure Intervention Trial (SPRINT) was designed as

an efficacy trial and was stopped early when the primary outcome, cardiovascular events, was significantly reduced with the lower target.<sup>2</sup> As a large, high-quality, definitive study with a unique comparison of blood pressure targets, SPRINT was well-positioned to directly inform guidelines and trigger guideline updates. But a debate arose around the clinical relevance and a potentially inappropriate focus on benefits. Finally, guideline developers disagreed whether benefits outweigh increased rates of adverse events like acute kidney injury or increased treatment burden, ultimately resulting in two conflicting US guidelines.<sup>3,4</sup>

Specifying a primary outcome that more directly informs the benefit–harm balance reduces the risk of multiple testing and misleading interpretation of study results. For example, the Aspirin in Reducing Events in the Elderly (ASPREE) trial used disability-free survival as a primary outcome.<sup>5</sup> Rather than just showing a debatable benefit–harm balance of less myocardial infarctions at the cost of more gastrointestinal bleeds, ASPREE demonstrated that aspirin did not prolong disability-free survival. If such composite outcomes are the primary outcome, they need to fulfill fundamental criteria to be interpretable.<sup>6</sup> They should also be highly

<sup>1</sup>Epidemiology, Biostatistics and Prevention Institute, Department of Epidemiology, University of Zurich, Zurich, Switzerland

<sup>2</sup>Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA, USA

<sup>3</sup>School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, VIC, Australia

<sup>4</sup>Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

#### Corresponding author:

Milo A Puhan, Epidemiology, Biostatistics and Prevention Institute, Department of Epidemiology, University of Zurich, Hirschengraben 84, Zurich 8001, Switzerland.  
Email: miloalan.puhan@uzh.ch

relevant to the target population, such as disability-free survival for the elderly.

Second, we call for large-scale RCTs to have sufficient statistical power to explicitly measure clinically relevant differences in the benefit–harm balance. Current guidance proposes that studies should be powered for multiple patient-important outcomes to ensure that benefit–harm balance can be assessed,<sup>1</sup> but is otherwise non-specific. It is typically neither feasible nor necessary to power the study for *all* patient-important outcomes, or (often) for all outcomes in a core outcome set. Instead, we propose that sample size calculations should aim at estimating a metric for the benefit–harm balance (e.g. disability-free survival, survival with good function, or probability of net benefit) as precisely as needed for decision-making, based on expected treatment effects on both benefits and harms and taking into account both baseline risks and the relative importance of outcomes.

Benefit–harm metrics are useful to compare multiple outcomes on a common scale and to model the impact of additional evidence on the benefit–harm balance.<sup>7–9</sup> To inform RCT design, such benefit–harm metrics should be sensitive to key patient-important benefits and harms and be responsive to the intervention. The duration of RCTs should reflect time-frames relevant to stakeholders in which both benefits and harms occur, as treatments or preventive and screening interventions may cause some benefits or harms earlier than others. Powering for the benefit–harm balance will require a larger sample size than powering for the composite of all benefit outcomes, as is often done. However, if there is more than one benefit outcome, compared to powering for a single benefit outcome only, the sample size could both decrease or increase when powering for the benefit–harm balance instead. By powering RCTs for a benefit–harm metric, RCTs will generate more valid and precise evidence for the benefit–harm balance. This will also avoid stopping large supposedly definitive RCTs like SPRINT based on benefit alone, and instead allow formulating explicit stopping rules for net benefit, net harm or futility.

Third, we advocate for nested patient preference surveys, as preference surveys could strongly guide the interpretation of results and impact guideline development and policymaking. Patient preferences can inform the choice of the outcomes and their relative importance. Decision-makers need to consider the patients' perspective to balance benefits against harms and determine clinical relevance. In the absence of such evidence, guidelines can contradict each other, as in the blood

pressure target example above. Although some evidence on patient preferences may be available, it is unlikely that any preference survey designed and performed independently of an RCT will include all outcomes (or health states) of interest and that outcome descriptions in the survey match outcome definitions of the RCT. Furthermore, respondents of surveys performed separately may not represent the trial or target population well.

Moreover, for large definitive RCTs like SPRINT, it is feasible to perform sufficiently large nested preference surveys to additionally determine the impact of variation in preferences between individuals. In contrast, guideline panels do not typically have the resources to perform large, applicable preference surveys. In a recent research project, we performed our own preference survey in patients with hypertension using best-worst scaling, a ranking exercise where the respondent repeatedly chooses the best and the worst outcome in different combinations.<sup>10</sup> We could show that there is large variation in preferences between individuals, and that individual preferences can shift the benefit–harm balance of blood pressure targets.<sup>11</sup> Guideline developers highly valued this result and suggested shared decision-making would be appropriate.<sup>12</sup> We also found that patient preference surveys are difficult to perform for guideline developers: funding may frequently be lacking, and contacting the right patients may only be feasible through collaboration with care delivery groups with learning health systems, which can identify and contact a representative sample of members of the target population.<sup>12</sup> Therefore, nested surveys in RCTs will likely provide the most applicable, valid and precise evidence on preferences to many guideline panels.

In summary, we propose a major change in the culture for large-scale prevention RCTs to primarily aim to increase the certainty in the benefit–harm balance rather than in single benefit outcomes. In particular, definitive RCTs should aim to establish net benefit, net harm or equivalence. This approach requires thorough stakeholder engagement, in particular to ensure all patient-important outcomes are considered.<sup>12</sup> Furthermore, the ethics committee and data and safety monitoring boards would need to accept the benefit–harm metric as valid to determine equipoise. Given the disproportionate focus of RCTs on benefits and often lacking evidence on harms, it is not surprising that guideline developers frequently come to conflicting conclusions although often based on the same RCTs. Since RCTs are the main source of information for guideline development, policies and ultimately

individual decision-making, a design and reporting that focuses more on the benefit–harm balance will help RCTs to provide high-quality, actionable evidence.


### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: H.E.A. was supported by a PhD fellowship of the Béatrice Ederer-Weber foundation and a Swiss National Science Foundation Early Postdoc.Mobility fellowship.

### ORCID iD

Hélène E Aschmann  <https://orcid.org/0000-0003-1234-4321>

### References

1. Cook JA, Julious SA, Sones W, et al. Practical help for specifying the target difference in sample size calculations for RCTs: the DELTA<sup>2</sup> five-stage study, including a workshop. *Health Technol Assess* 2019; 23: 1–88.
2. SPRINT Research Group, Wright JT, Williamson JD, et al. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med* 2015; 373: 2103–2116.
3. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. *J Am Coll Cardiol* 2017; 71(19): e127–e248.
4. Qaseem A, Wilt TJ, Rich R, et al. Pharmacologic treatment of hypertension in adults aged 60 years or older to higher versus lower blood pressure targets: a clinical practice guideline from the American College of Physicians and the American Academy of Family Physicians. *Ann Intern Med* 2017; 166: 430–437.
5. McNeil JJ, Woods RL, Nelson MR, et al. Effect of aspirin on disability-free survival in the healthy elderly. *N Engl J Med* 2018; 379: 1499–1508.
6. Montori VM, Permanyer-Miralda G, Ferreira-González I, et al. Validity of composite end points in clinical trials. *Br Med J* 2005; 330: 594–596.
7. Puhan MA, Yu T, Boyd CM, et al. Quantitative benefit-harm assessment for setting research priorities: the example of roflumilast for patients with COPD. *BMC Med* 2015; 13: 157.
8. Puhan MA, Singh S, Weiss CO, et al. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med Res Methodol* 2012; 12: 173.
9. Guo JJ, Pandey S, Doyle J, et al. A review of quantitative risk–benefit methodologies for assessing drug safety and efficacy—report of the ISPOR risk-benefit management working group. *Value Health* 2010; 13(5): 657–666.
10. Aschmann HE, Puhan MA, Robbins CW, et al. Outcome preferences of older people with multiple chronic conditions and hypertension: a cross-sectional survey using best-worst scaling. *Health Qual Life Outcomes* 2019; 17: 186.
11. Aschmann HE, Boyd CM, Robbins CWR, et al. Balance of benefits and harms of different blood pressure targets in people with multiple chronic conditions: a quantitative benefit-harm assessment. *BMJ Open* 2019; 9: e028438.
12. Aschmann H, Boyd C, Robbins C, et al. Informing patient-centered care through stakeholder engagement and highly stratified quantitative benefit-harm assessments. *Value Health* 2020; 23(5): 616–624.