It has been widely accepted that most randomized control trials (RCTs) include patient groups that are not a representative sample of the patients who will receive the intervention in daily practice [1–3]. This has raised concerns about the generalizability of RCT results. Recently, Pressler and Kaizar [4] enriched the discussion by asserting that the bias that results from a lack of generalizability can be quantified. They define two populations: the first consisting of subjects fulfilling the inclusion criteria of an RCT (in their notation group I) and the second group (group E) comprising subjects not meeting the inclusion criteria. They propose to estimate the treatment effects in both groups \( \hat{\pi} \) (I) and \( \hat{\pi} \) (E), respectively) using nonrandomized (i.e., observational) data. Assuming equal amounts of confounding in both groups, \( \hat{\rho} = \hat{\pi} (I) - \hat{\pi} (E) \) provides an unbiased estimate of how much treatment effect modification there exists between included and excluded subjects. If the interest is in estimating the ‘population average treatment effect’ (PATE), weighing \( \hat{\rho} \) by the proportion of E among the total population of interest \( \hat{n} = \frac{n_E}{n} \) provides an estimate \( \hat{\gamma} = \hat{\pi} \hat{\rho} \) of how much ‘generalizability bias’ is created by relying on \( \hat{\pi} (I) \) to estimate the PATE. This weighing of \( \hat{\rho} \) is necessary because if there is treatment effect modification between groups I and E, the PATE is dependent on the proportionate size of both groups. While we acknowledge the relevance of the approach suggested by Pressler and Kaizar, we wish to touch upon some concerns and discuss alternative strategies for exploring generalizability.

First, Pressler and Kaizar fail to address why one would be interested in the treatment effect in group E. For example, if we explore the effectiveness of a new antihypertensive drug and E comprises subjects without hypertension, it seems illogical to try to estimate treatment effect modification between groups I and E. Second, using nonrandomized data to estimate \( \hat{\gamma} \) or \( \hat{\rho} \) only results in an unbiased estimate if the amount of confounding is equal in both groups E and I. This assumption is not testable, as the authors acknowledge, and results in a problem encountered in virtually all nonrandomized studies, that...
is, not knowing whether estimates are unbiased or biased by confounding. Third, when \( \hat{y} \) is estimated, an implicit assumption is made that the effects within groups I and E are homogenous; otherwise, \( \hat{y} \) does not necessarily reflect a lack of generalizability due to excluding group E. Imagine an RCT in which only subjects younger than 50 years are enrolled and let there be treatment effect modification by diabetes status. Hence, group E would consist of subjects older than 50 years. In that case, \( \hat{y} \) might deviate from 0 simply because age increases the number of diabetic subjects; that is, the magnitude of \( \hat{y} \) becomes dependent on the proportion of diabetic subjects. To show that this is not a lack of generalizability between excluded and included patients, note that while there is effect modification between diabetic and non-diabetic subjects \( (\hat{\beta} = \hat{\Delta}_{DM} - \hat{\Delta}_{NDM} \neq 0) \) within diabetic and non-diabetic subgroups, the difference between groups I and E equals 0 \( (\hat{y} = \hat{\Delta}_{E} \neq \hat{\Delta}_{I}) \). This brings us to the final issue. If there is treatment effect modification between groups I and E (and assuming homogenous effects within groups I and E), the PATE will depend on the proportion of excluded patients \( \hat{n} \). For example, let there be treatment effect modification between groups I and E caused by age. Specifically, the relative risk for the outcome under treatment is 0.4 in subjects younger than 50 years (group I) and 1 among subjects older than 50 (group E). Furthermore, let the proportion E differ from 0.1 to 0.9 between certain regions of a country. Consequently, the PATE will range from 0.44 to 0.91. Reporting numerous region-specific PATE estimates is at the very least inefficient compared to the alternative of reporting two age-specific estimates. Furthermore, in the presence of treatment effect modification, the PATE is not applicable to any (group of) subject(s), making this an inappropriate effect estimate. Instead, the age specific effect estimates are applicable to their respective group members. If estimating the PATE is inappropriate when there is treatment effect modification, it is equally inappropriate to interpret \( \hat{y} \) as the amount of generalizability bias. Therefore, we suggest that instead of focusing on generalizability bias, it is more helpful to simply indicate if treatment effect modification is present or absent (i.e., if \( \hat{\beta} \neq 0 \)). This can be estimated between groups I and E, as Pressler and Kaizar advise, but also within subgroups of I or E.

Before discussing alternative strategies, we want to recognize that in settings in which all trials exclude the same kind of subjects, the approach suggested by Pressler and Kaizar to estimate \( \hat{\beta} \) can indeed provide valuable insights generalizability. However, at the potential cost of confounding bias because the effect estimates in E and I are based on nonrandomized data. Alternatively, if some RCTs include patients excluded by other RCTs, comparisons of the effect estimates between trials can also provide information on generalizability [5]. While this latter approach prevents confounding bias within trials, estimates may still differ due to differences between studies, for example, concomitant drug use, which could also affect conclusions regarding generalizability [6]. Ideally, \( \hat{\beta} \) should be estimated within RCTs, which guard both against bias due to study-specific effects and confounding bias. Therefore, we suggest that researchers explore generalizability by focusing on treatment effect modification using individual patient data from multiple RCTs, which allows \( \hat{\beta} \) to be estimated within studies.

References