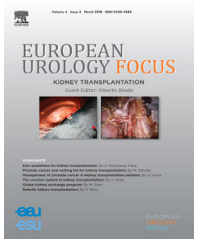
available at www.sciencedirect.com
journal homepage: www.europeanurology.com/eufocus



Mini Review – Epidemiology

Evidence-based Urology: Subgroup Analysis in Randomized Controlled Trials

Tuomas P. Kilpeläinen [a], *Kari A.O. Tikkinen* [a,b], *Gordon H. Guyatt* [c,d], *Robin W.M. Vernooij* [e,f,*]

[a] *Department of Urology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland;* [b] *Department of Surgery, South Karelia Central Hospital, Lappeenranta, Finland;* [c] *Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada;* [d] *Department of Medicine, McMaster University, Hamilton, Canada;* [e] *Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht, The Netherlands;* [f] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*

Article info

Article history:
Accepted October 5, 2021
Available online 20 October 2021
Associate Editor: Richard Lee

Keywords:
Evidence-based medicine
Subgroup analyses
Effect modification
Interaction

Abstract

In randomized controlled trials, investigators often explore the possibility that the treatment effects differ between subgroups (eg, women vs men, old vs young, more versus less severe disease). Investigators often inappropriately claim subgroup effects (also called "effect modification" or "interaction") when the likelihood of a true effect modification is low. Criteria for assessing the credibility of subgroup analyses, nicely summarized in a formal Instrument for Assessing the Credibility of Effect Modification Analyses (ICEMAN), include investigator postulation of a priori hypotheses with a specified direction; support from prior evidence; a low likelihood that chance explains the apparent subgroup effect; and only testing a small number of subgroup hypotheses.

Patient summary: Randomized clinical trials often use subgroup analyses to explore whether a treatment is more or less effective in a particular patient subgroup (eg, women vs men, old vs young). In this mini-review, we explore the common pitfalls of subgroup analyses.

Of all study designs, randomized controlled trials (RCTs) provide the best evidence regarding treatment efficacy [1,2]. Randomized trials often enroll diverse populations (the old and young; severe and mild conditions), raising the possibility that the treatment effect may differ in subpopulations (eg, treatment is effective in the old but not in the young). Investigators therefore often conduct analyses to explore such possible subgroup effects, also referred to as "effect modification" or "interaction".

Despite the best intentions, investigators often fail to conduct subgroup analyses adequately and to optimally interpret the results of such analyses. Claims of subgroup effects that are in fact spurious have the potential to compromise patient care [3,4]. In this mini-review, we explore common limitations and pitfalls of subgroup analyses in RCTs.

The first question that arises when treatment appears to work better in one subgroup than another is whether chance can explain the difference. To address this issue, investigators must execute a statistical test, usually referred to as a "test of interaction" [5]. The interaction test generates *p* values: if $p > 0.05$, chance remains a likely explanation of an apparent subgroup effect; only very low *p* values (≤ 0.005) for the interaction test provide high confidence that chance cannot explain the apparent subgroup effect [6].

Aside from the *p* value for the test of interaction, other criteria can help in distinguishing between a credible and less credible subgroup claim. A claim is more credible (1) if it is supported by an a priori hypothesis with an accurately prespecified direction; (2) if prior evidence of the subgroup effect exists; (3) if investigators have tested only

* Corresponding author at: University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, The Netherlands.
E-mail address: r.w.m.vernooij-2@umcutrecht.nl (R.W.M. Vernooij).

https://doi.org/10.1016/j.euf.2021.10.001

Table 1 – Instrument for Assessing the Credibility of Effect Modification Analyses (ICEMAN) questions for randomized controlled trials [6].

1: Was the direction of the effect modification correctly hypothesized a priori?			
[ ] Definitely no *Clearly post hoc or results inconsistent with hypothesized direction or biologically very implausible*	[**X**] Probably no or unclear *Vague hypothesis or hypothesized direction unclear*	[ ] Probably yes *No prior protocol available but unequivocal statement of a priori hypothesis with correct direction of effect modification*	[ ] Definitely yes *Prior protocol available and includes correct specification of direction of effect modification, eg, based on a biologic rationale*
2: Was the effect modification supported by prior evidence?			
[ ] Inconsistent with prior evidence *Prior evidence suggested a different direction of effect modification*	[ ] Little or no support or unclear *No prior evidence or consistent with weak or very indirect prior evidence (eg, animal study at high risk of bias) or unclear*	[**X**] Some support *Consistent with more limited or indirect prior evidence (eg, large observational study, non-significant effect modification in prior RCT, or different population)*	[ ] Strong support *Consistent with strong prior evidence directly applicable to the clinical scenario (eg, significant effect modification in related RCT)*
3: Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification? (consider irrespective of number of effect modifiers)			
[**X**] Chance a very likely explanation *Interaction p value > 0.05*	[ ] Chance a likely explanation or unclear *Interaction p value ≤ 0.05 and >0.01, or no test of interaction reported and not computable*	[ ] Chance may not explain *Interaction p value ≤ 0.01 and >0.005*	[ ] Chance an unlikely explanation *Interaction p value ≤ 0.005*
4: Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?			
[ ] Definitely no *Explicitly exploratory analysis or large number of effect modifiers tested (eg, greater than 10) and multiplicity not considered in analysis*	[**X**] Probably no or unclear *No mention of number or 4–10 effect modifiers tested and number not considered in analysis*	[ ] Probably yes *No protocol available but unequivocal statement of 3 or fewer effect modifiers tested*	[ ] Definitely yes *Protocol available and 3 or fewer effect modifiers tested or number considered in analysis*
5: If the effect modifier is a continuous variable, were arbitrary cut points avoided? [ ] not applicable: not continuous			
[ ] Definitely no *Analysis based on exploratory cut point (eg, picking cut point associated with highest interaction p value)*	[**X**] Probably no or unclear *Analysis based on cut point(s) of unclear origin*	[ ] Probably yes *Analysis based on pre-specified cut points, eg, suggested by prior RCT*	[ ] Definitely yes *Analysis based on the full continuum, eg, assuming a linear or logarithmic relationship*
How would you rate the overall credibility of the proposed effect modification? The overall rating should be driven by the items that decrease credibility. The following provides a sensible strategy: •All responses definitely or probably reduced credibility or unclear => very low •Two or more responses definitely reduced credibility => maximum usually low even if all other responses satisfy credibility criteria •One response definitely reduced credibility => maximum usually moderate even if all other responses satisfy credibility criteria •Two responses probably reduced credibility => maximum usually moderate even if all other responses satisfy credibility criteria •No response options definitely or probably reduced credibility => high very likely			

a small number of subgroup hypotheses; and (4) if the subgroup effect is a continuous variable, investigators have avoided cut points driven by the data (eg, choosing a threshold of age 50 yr rather than 40 or 60 yr because 50 is threshold that suggests a subgroup effect). To facilitate clinician judgments regarding subgroup effects, investigators have developed a simply applied tool called Instrument for Assessing the Credibility of Effect Modification Analyses (ICEMAN) (Table 1) that summarizes these criteria [6].

Previous studies have demonstrated that trialists are often suboptimal in planning and conducting the appropriate statistical test for interaction [7,8]. Spurious or false-positive results are especially common when investigators test a plethora of hypotheses. Defining subgroups post hoc, as evidenced by failure to report the subgroup test in the original trial protocol, may be particularly problematic. In such instances, when the subgroups are not preplanned, spurious subgroup inferences are common.

Another serious error is defining subgroups after randomization, when treatment might have already influenced patient characteristics. Therefore, clinicians should reject any subgroup analysis that does not focus on the variables defined at baseline [4].

All these concerns highlight why clinicians cannot necessarily trust authors' interpretation of subgroup effects, which is why the ICEMAN instrument is needed. To illustrate the use of the ICEMAN tool, we selected the well-known Prostate Cancer Intervention Versus Observation Trial (PIVOT) as an example of assessing subgroup credibility in the light of current evidence [9].

To summarize, during 1994–2002, PIVOT recruited 731 men (age ≤ 75 yr, life expectancy ≥ 10 yr, fit for surgery) with localized prostate cancer (prostate-specific antigen [PSA] level < 50 μg/l, clinical stage T1–2, any grade). The men were randomized to radical prostatectomy (n = 364) or observation (n = 367). At 22 yr of follow-up, the risk of any-cause death was 68% for men randomized to surgery and 73% for men in the observation group (relative risk 0.92, 95% confidence interval 0.84–1.01). In their abstract, the authors state: "Results did not significantly vary by patient or tumor characteristics, although differences were larger favoring surgery among men aged < 65 yr, of white race, and having better health status, fewer comorbidities, ≥34% positive prostate biopsy cores, and intermediate-risk disease." [9]. The latest European Association of Urology prostate cancer guideline appears to consider these subgroup inferences credible: the guideline states that patients with intermediate-risk cancer benefit more from surgery than men with low-risk or high-risk cancer [10].

Should urologists thus recommend radical prostatectomy to younger White men with good overall health status and fewer comorbidities and large-volume intermediate-risk cancers, but perhaps not to Black older men with comorbidities and low-volume, high-grade cancer?

The first ICEMAN question asks if the direction of the effect modification was correctly hypothesized a priori (Table 1). The first PIVOT paper in 2012 includes a study protocol as a supplementary file [11]. The protocol does predefine nine subgroups, including age, race, tumor stage, tumor grade, family history, PSA level, and Charlson comor-

bidity index. However, they do not specify the direction of any hypotheses (eg, were the authors thinking that surgery would have a greater effect on Black or White men) and therefore the answer is "probably no" (Table 1).

The second question is: Was the effect modification supported by prior evidence? If we look at the evidence accumulated thus far from the three RCTs concerning prostatectomy versus observation, namely PIVOT, SPCG-4 [12], and ProtecT [13], as well as the large observational PCBase Sweden study [14], we find some support for a subgroup effect in some subgroups in the SPCG-4 trial and the PCBase Sweden study (eg, age < 65 vs ≥65 yr), but not in the ProtecT trial. These findings from SPCG-4 and the PCBase Sweden study are probably of low credibility (SPCG-4: no formal interaction tests, no preplanned subgroups; PCBase Sweden: observational study). The answer to the question is thus "some support" (Table 1).

The third question asks if the interaction test suggests that chance is an unlikely explanation for subgroup differences. The *p* values for interaction tests in the PIVOT subgroups are between 0.1 and 0.8, meaning that chance is a very likely explanation for the effect modification observed. The answer is "chance a very likely explanation" (Table 1).

The fourth question asks: Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis? At least nine subgroups were listed in the original protocol, of which seven are presented in the latest publication. Therefore, the answer is "probably no" (Table 1).

The last question is: If the effect modifier is a continuous variable, were arbitrary cut points avoided? In the original protocol the cut points were not predefined [11]. The continuous variables in subgroup analyses in the latest PIVOT publication are age (<65 vs ≥65 yr), PSA (≤10 vs >10 ng/ml), performance score (0 vs 1–4), and Gleason score (<7 vs 7 vs 8–10). These thresholds are based on clinical relevance; however, the positive biopsy core subgroups were selected according to <34% versus ≥34 positive cores. The choice for this threshold remains unclear and therefore the answer for this ICEMAN item is also "probably no".

The responses to all individual items of ICEMAN suggest low to very low credibility of the subgroup effects. The clinician should thus anticipate that the overall relative effect would apply to all patients and should thus not recommend differential therapy on the basis of subgroup effects.

Of course, our discussion has focused on the relative subgroup effect. The absolute effect is a different matter. With the same relative effect, any absolute reduction in mortality would be small in low-risk patients and larger in intermediate- or high-risk patients owing to the higher absolute (baseline) risk of death with higher-risk cancer [15]. For example, let us assume that radical prostatectomy provides a substantial relative risk reduction (in prostate cancer mortality) of 35% in all subgroups [16]. If the absolute risk of prostate cancer death with low-risk cancer is 3%, with the 35% relative risk reduction the absolute risk reduction is 1% or one in 100. In very high-risk cancer, however, if the absolute risk of prostate cancer death is 60%, with the same 35% relative risk reduction the absolute risk reduction is approximately 20% or 20 in 100. When we look at effect modification, however, we are focusing on relative effects. Differences in absolute effects across subgroups will be present for any effective treatment in which patients differ in their risk of adverse outcomes (in contrast to true differences in relative effects that are rare, differences in baseline risk are extremely common).

When reading a clinical trial that includes a claim of a subgroup effect, asking the ICEMAN questions is crucial for any clinician, reviewer, or editor. It is not rare that the results of the subgroup analyses alter interpretation of RCTs and guide treatment choices. False inferences may lead to the use of ineffective treatments or deny patients effective treatment. Exploratory subgroup analyses may lead to important findings and guide further research, but clinicians should consider results for which the credibility is low as merely hypothesis-generating and not a finding that should influence their practice.

Conflicts of interest: The authors have nothing to disclose.

References

[1] Sackett DL. Clinician-trialist rounds: 16. Mind your explanatory and pragmatic attitudes! — part 1: what? Clin Trials 2013;10:495–8.
[2] Califf RM, Sugarman J. Exploring the ethical and regulatory issues in pragmatic clinical trials. Clin Trials 2015;12:436–41.
[3] Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176–86.
[4] Sun X, Ioannidis JP, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. JAMA 2014;311:405–11.
[5] Brankovic M, Kardys I, Steyerberg EW, et al. Understanding of interaction (subgroup) analysis in clinical trials. Eur J Clin Invest 2019;49:e13145.
[6] Schandelmaier S, Briel M, Varadhan R, et al. Development of the Instrument to Assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. Can Med Assoc J 2020;192:E901–6.
[7] Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. JAMA Intern Med 2017;177:554–60.
[8] Kasenda B, Schandelmaier S, Sun X, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. BMJ 2014;349:g4539.
[9] Wilt TJ, Vo TN, Langsetmo L, et al. Radical prostatectomy or observation for clinically localized prostate cancer: extended follow-up of the Prostate Cancer Intervention Versus Observation Trial (PIVOT). Eur Urol 2020;77:713–24.
[10] Mottet N, van den Bergh RCN, Briers E, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer—2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. Eur Urol 2021;79:243–62.
[11] Wilt TJ, Brawer MK, Jones KM, et al. Radical prostatectomy versus observation for localized prostate cancer. N Engl J Med 2012;367:203–13.
[12] Bill-Axelson A, Holmberg L, Garmo H, et al. Radical prostatectomy or watchful waiting in early prostate cancer. N Engl J Med 2014;370:932–42.
[13] Hamdy FC, Donovan JL, Lane JA, et al. 10-Year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. N Engl J Med 2016;375:1415–24.
[14] Sooriakumaran P, Nyberg T, Akre O, et al. Comparative effectiveness of radical prostatectomy and radiotherapy in prostate cancer: observational study of mortality outcomes. BMJ 2014;348:g1502.
[15] Witte LPW, Tikkinen KAO, Guyatt GH, Malde S. Evidence-based urology: importance of relative versus absolute effect. Eur Urol Focus [in press].
[16] Kilpeläinen TP, Järvinen P, Tikkinen KAO. Randomized trials show a consistent benefit of radical prostatectomy on mortality outcomes. J Urol 2019;202:1106–8.