Received: 3 April 2017 | Revised: 5 February 2019 | Accepted: 5 February 2019

DOI: 10.1002/sim.8133

RESEARCH ARTICLE

WILEY Statistics in Medicine

Level of evidence for promising subgroup findings: The case of trends and multiple subgroups

Julien Tanniou[1,2] | Sanne C. Smid[3,4] | Ingeborg van der Tweel[3] | Steven Teerenstra[5,6] | Kit C.B. Roes[3,5]

[1]INSERM CIC 1412, CHRU Brest, Brest, France

[2]European Medicines Agency, London, UK

[3]Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, Utrecht, The Netherlands

[4]Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

[5]Medicines Evaluation Board, College ter Beoordeling van Geneesmiddelen, Utrecht, The Netherlands

[6]Department of Health Evidence, Section Biostatistics, Radboud UMC, Nijmegen, The Netherlands

Correspondence
Julien Tanniou, INSERM CIC 1412, CHRU Brest, Brest, France; or European Medicines Agency, London, UK.
Email: julien.tanniou@chu-brest.fr

Subgroup analyses are an essential part of fully understanding the complete results from confirmatory clinical trials. However, they come with substantial methodological challenges. In case no statistically significant overall treatment effect is found in a clinical trial, this does not necessarily indicate that no patients will benefit from treatment. Subgroup analyses could be conducted to investigate whether a treatment might still be beneficial for particular subgroups of patients. Assessment of the level of evidence associated with such subgroup findings is primordial as it may form the basis for performing a new clinical trial or even drawing the conclusion that a specific patient group could benefit from a new therapy. Previous research addressed the overall type I error and the power associated with a single subgroup finding for continuous outcomes and suitable replication strategies. The current study aims at investigating two scenarios as part of a nonconfirmatory strategy in a trial with dichotomous outcomes: (a) when a covariate of interest is represented by *ordered* subgroups, eg, in case of biomarkers, and thus, a trend can be studied that may reflect an underlying mechanism, and (b) when multiple covariates, and thus multiple subgroups, are investigated at the same time. Based on simulation studies, this paper assesses the credibility of subgroup findings in overall nonsignificant trials and provides practical recommendations for evaluating the strength of evidence of subgroup findings in these settings.

KEYWORDS

clinical trials, failed study, multiple testing, overall nonsignificant trial, subgroup analysis, type I error

1 | INTRODUCTION

Traditionally, patients enrolled in confirmatory clinical trials should “closely mirror the target population. Hence, in these trials it is generally helpful to relax the inclusion and exclusion criteria as much as possible within the target population, while maintaining sufficient homogeneity to permit precise estimation of treatment effects […] Subsequently, if heterogeneity of treatment effects is found, this should be interpreted with care.”[1] Tanniou et al recently published a review of marketing authorisation application (MAA) dossiers relating to medicinal products containing new active substances and evaluated by the European Medicines Agency over the period 2012-2015.[2] The review showed that subgroup-related so-called Major Objections and/or Other Concerns are prominent (68%) in the so-called day 80 assessment reports. This result emphasises the essential role of subgroup analyses in the assessment of MAAs.

When the overall treatment effect is significant, subgroup analyses are typically performed to investigate whether this effect is consistent across subgroups. On the contrary, if the overall treatment effect does not reach (a priori defined) statistical significance, any subgroup analysis finding(s) can only be seen as exploratory. This latter scenario is of particular interest as a potential new relevant treatment may only benefit a certain subpopulation.[3-5] Claiming efficacy of a drug based on an apparent positive subgroup finding in an overall nonsignificant trial and without replication is usually considered as a no-go decision because of the well-known increase of the type I error rate.

The recent EU draft guidance on the investigation of subgroups in confirmatory clinical trials reinforces that conclusions on positive subgroup findings are possible in very exceptional instances and that amongst others biological/pharmacological rational, (at least partial) replication of this finding is mandatory and of paramount importance.[6] This guidance also underlines that, in specific situations, such as for the clinical setting of high unmet medical need or for situations where trials are usually of considerable size, careful assessment of the overall available evidence could exceptionally lead to a positive (subgroup) licensing decision. In conditions where there are no or few treatments available, a new drug could be licensed for a subpopulation or conditionally approved with post-approval commitment. Similarly, when, for instance, the feasibility of an additional confirmatory trial is highly questionable, the assessment of the totality of evidence for decision-making might be less driven by statistical considerations.

Current practice in many clinical trials is that a number of subgroup analyses is pre-defined in the protocol, although often without clear statement whether they are intended for consistency checks or because of a prior expectation of differential effect, unless they are planned as part of a confirmatory testing strategy protecting the overall type I error.[7] In addition, when primary results of clinical trials are published, they often include a (large) number of subgroup analyses,[8,9] with at least some conclusions drawn based on these.

Recently, Tanniou et al investigated the statistical level of evidence for promising subgroup findings in overall nonsignificant trials with continuous outcomes.[3] They considered the overall type I error and the power associated with such findings and also suitable replication strategies. In case of a single trial, the inflation of the overall type I error is substantial, especially in relatively small subgroups. They also showed, unexpectedly, that testing a subgroup when there is a so-called "tendency" for effect in the overall population, defined as a one-sided p-value for the overall test between 0.025 and 0.05, is bad practice with substantial overall type I error inflation. The replication strategies investigated substantially improved the level of evidence.[3] In absence of other evidence, replication of promising subgroup findings in a new trial should be the standard approach if the trial is overall statistically nonsignificant. Replication of trials may incur substantial investments, importantly from participating patients, and should therefore only be undertaken if the biological plausibility as well as the statistical level of evidence is sufficiently convincing.

Some literature exists about control of the overall type I error in the specific situation where one is (a priori) interested in the effect of a given study treatment on a subgroup of patients with certain clinical or biological attributes, ie, when the subgroup of interest is part of the confirmatory testing strategy.[10-12] Especially, Song and Chi proposed a method that (strongly) controls the familywise type I error rate considering an overall pre-specified trend as expressed with a p-value.[10]

This paper, however, investigates a practical approach for two scenarios as part of a nonconfirmatory strategy (ie, no a priori control of overall type I error) in a trial with dichotomous outcomes: (a) when a covariate of interest is represented by *ordered* subgroups, eg, in case of biomarkers, and thus, a trend can be studied that may reflect an underlying mechanism, and (b) when multiple covariates, and thus multiple subgroups, are investigated at the same time. Even though not part of a confirmatory strategy setting, both situations may raise relevant questions. Practical advice on how to properly judge the level of evidence is therefore of importance in view of the conclusion and actions that may follow evaluation of subgroups.

In general, if a statistically significant trend, based on a test with a suitably small p-value, is observed across subgroups that results from a categorical covariate (eg, different age classes, biomarker levels), it could be perceived as more convincing. It should be highlighted that the purpose of this paper is not to add a "tick-box" in this decision-making process that could lead to a treatment approval for a particular subpopulation, but rather as a first step to objectively decide whether the (statistical) level of evidence associated with the observed trend is convincing enough to undertake any further investigations, such as a new (replication) trial dedicated to that subgroup. It aims subsequently to describe the data appropriately and discern if a signal could be separated from noise, keeping in mind that all those investigations are exploratory. From the perspective of efficacy ("benefit"), we are questioning whether it is justified to focus on the subgroup(s) for either establishing benefit risk in more detail and/or initiate a follow-up trial. Of note and when dealing with a quantitative variable ordered into categories, the (pre-) definition of cut-points to define the different subgroups is an important point to consider when assessing such *ordered* categories. In the eventuality of replication of the subgroup finding, this aspect should be further investigated in order to confirm the adequacy of the subgroup definition. The choice of the cut-points is, however, beyond the scope of this study.

The remainder of this paper is organised as follows. In Section 2, an empirical example is presented. In Section 3, we present our approach and results for the situation where a trend in a categorical subgrouping variable is observed, while Section 4 deals with the investigation of multiple subgroups. This paper ends with a discussion in Section 5.

2 | THE DEXAMETHASONE FOR CARDIAC SURGERY TRIAL EXAMPLE

The Dexamethasone for Cardiac Surgery trial was a multicentre, randomised, double-blind, placebo-controlled trial comparing high-dose intravenous Dexamethasone to placebo treatment in patients undergoing cardiac surgery.[13] The objective was to quantify the effect of intraoperative high-dose Dexamethasone on the incidence of major adverse events in patients undergoing cardiac surgery. The primary outcome was the composite of death, myocardial infarction, stroke, renal failure, or prolonged postoperative mechanical ventilation, within 30 days after surgery. Several subgroup analyses, as well as the categories used, were pre-planned for the primary outcome and its separate components: age with protocol defined categories (< 65, 65-74, 75-79, and ≥ 80 years), sex, diabetes, chronic pulmonary disease, high (≥ 5) versus low (≤ 4) EuroScore pre-operative risk estimate, and prolonged cardiopulmonary bypass (CPB) duration (defined as CPB duration >150 minutes). It is important to note that these analyses were pre-planned for the conventional objective of checking consistency. Logistic regression was used to assess whether any interaction existed between Dexamethasone and the subgroups mentioned above, with a pre-defined 0.10 threshold for significance of an interaction term. The composite primary study endpoint occurred in 157 of the 2235 patients (7.0%) randomised to Dexamethasone and in 191 of the 2247 patients (8.5%) randomised to placebo (RR, 0.83; 95% CI, 0.67-1.01; absolute risk reduction, −1.5%; 95% CI, −3.0% to 0.1%; $p = 0.07$). Therefore, the primary endpoint was formally not statistically significant. The subgroup analyses suggested an age-dependent effect of Dexamethasone on the primary study endpoint (test of interaction; $p = 0.08$). In patients younger than 65 years, Dexamethasone was associated with lower likelihood for the primary endpoint (RR, 0.65; 95% CI, 0.44-0.96; $p = 0.03$). For patients between 65 and 75 years of age, the RR was 0.78 (95% CI, 0.56 to 1.09), and for patients between 75 and 80 years of age, the RR was 0.88 (95% CI, 0.59 to 1.33). However, in patients older than 80 years, the direction of the effect reversed toward an increased risk with an RR of 1.69 (95% CI, 0.92-3.10; $p = 0.09$). The qualitative interaction of effect with age appeared to be predominantly caused by the *mortality* component of the primary composite endpoint, which showed a significant interaction ($p = 0.05$). In patients younger than 65 years, the RR for mortality was 0.42 (95% CI, 0.13-1.34; $p = 0.13$), but it gradually increased with age to 3.87 (95% CI, 1.10-13.6; $p = 0.02$) in patients aged 80 years or older. There was no (statistically significant) differential treatment effect in the subgroup analyses on sex, diabetes, chronic obstructive pulmonary disease, EuroScore, or prolonged CPB duration, even though point estimates were slightly different for EuroScore (with one of the subgroups being significant), COPD, and prolonged CPB duration. Hence, even though the treatment effect was not statistically significant for the overall population, this study in cardiac surgical patients might be considered to suggest a trend toward a beneficial effect in part of the population. In their publication, the authors mentioned that a subsequent study could also be considered with patient selection for the therapy, based on a larger beneficial effect in younger patients and no apparent benefit in those aged 80 years or older.

3 | OBSERVED TREND ACROSS SUBGROUPS

Practical decisions about the validity of subgroup findings have to be made on a day-to-day basis by different stakeholders, such as treating physicians, regulators, and pharmaceutical companies. If a trend in effect is observed over different subgroups, this could be assessed through various methods. Logistic regression analysis is commonly performed when analysing binary outcomes, providing (log) odds ratios (OR). Binomial and Poisson regression analysis, respectively referring to risk difference (RD) and (log) relative risk (RR), are considered more appropriate in some situations, particularly in case of randomised clinical trials. For studying interaction effects, the choice of the effect measure matters as no interaction on one scale likely implies interaction on another scale.[14] In other words, if a treatment effect truly exists, homogeneity in one measure, eg, RR, implies heterogeneity in at least one other, eg, RD. While testing for interaction is not sufficient on its own and can be criticised because of its lack of power, the concept of interaction can serve as a starting point for investigating the trend across subgroups. Based on the above, three *post hoc* strategies (all with three different scenarios) to assess the evidence for a subgroup effect were considered, all with an overall statistically nonsignificant treatment effect based on a chi-square test ($p_{OT} \geq 0.05$; OT referring to "Overall Test"). In the following, $RR_{\cdot}$, $OR_{\cdot}$, and $RD_{\cdot}$ are the estimates of the (log) relative risk, the (log) odds ratio, and the risk difference, respectively. Of note, the subscript refers to one of the three subgroups, subgroup 1 having the most promising result.

- Scenarios 1-3: observed trend in the data across subgroups:

1. $RR_1 < RR_2 < RR_3$;
2. $OR_1 < OR_2 < OR_3$;
3. $RD_1 < RD_2 < RD_3$.

The observed trend is *only* based on the ordering between subgroups of the point estimates for the effects (and not on any confidence intervals). This should constitute weak evidence.

- Scenarios 4-6: observed trend in the data across subgroups supported by an statistically significant interaction test ($p_{IT} < 0.10$; IT referring to "Interaction Test"):

4. $RR_1 < RR_2 < RR_3$ and a statistically significant interaction (Poisson regression analysis using a robust covariance matrix estimator);
5. $OR_1 < OR_2 < OR_3$ and a statistically significant interaction (logistic regression analysis);
6. $RD_1 < RD_2 < RD_3$ and a statistically significant interaction (binomial regression analysis).

The observed trend is evaluated on the chosen scale and within the same model. Results for $p_{IT} < 0.05$ and $p_{IT} < 0.20$ are provided in the supplementary material.

- Scenarios 7-9: observed trend in the data across subgroups supported by a statistically significant test for the effect in the most promising subgroup ($p_{ST_1} < 0.05$ with the chi-square test; ST referring to "Subgroup Test"):

7. $RR_1 < RR_2 < RR_3$ and a statistically significant effect for subgroup 1;
8. $OR_1 < OR_2 < OR_3$ and a statistically significant effect for subgroup 1;
9. $RD_1 < RD_2 < RD_3$ and a statistically significant effect for subgroup 1.

The observed trend is evaluated on the chosen scale and within the same model.

Finally scenario 10, where only a statistically significant subgroup without looking at a trend, is presented and compared to strategies 7-9 in order to clarify the potential added benefit of incorporating the trend in the decision strategy. For this comparison to be fair with scenarios 7-9, only subgroup 1 is tested ($p_{ST_1} < 0.05$), ie, we do not compare this with the data-dredging strategy of testing each of the three subgroups (which would, of course, perform even worse than scenario 10).

For these three assessment strategies, four different values for the size of subgroup 1 ($G_{1\cdot}$), denoted by $r_{G_{1\cdot}}$, as proportions of the total sample size were studied:

- $r_{G_{1\cdot}} = 0.1 \; \left(r_{G_{2\cdot}} = r_{G_{3\cdot}} = 0.45\right)$;
- $r_{G_{1\cdot}} = 0.33 \; \left(r_{G_{2\cdot}} = r_{G_{3\cdot}} = 0.33\right)$;
- $r_{G_{1\cdot}} = 0.5 \; \left(r_{G_{2\cdot}} = r_{G_{3\cdot}} = 0.25\right)$;
- $r_{G_{1\cdot}} = 0.9 \; \left(r_{G_{2\cdot}} = r_{G_{3\cdot}} = 0.05\right)$.

Of note, the second subscript refers to either c for control or t for treatment.

The population proportions of responders of the binary outcome are denoted by π_c and π_t for the control and treatment arms, respectively. These proportions of responders are assumed known for sample size calculation with an alpha of 0.05 (two-sided) and a power of 80%.[15] Three different proportions of responders for the control and the treatment group were investigated (Table 1) to reflect different sizes of clinical trials. Following Cohen's h effect sizes based on the arcsine transformation, they can be classified as very small, small, and medium effect sizes.[16]

Note that the subgroups are assumed to have equal sizes within the treatment and control groups, ie, being perfectly stratified. In order to assess the overall type I error and the power of each strategy, different proportions of responders have been simulated for all subgroups, ie, G_{1T}, G_{2T}, G_{3T}, G_{1C}, G_{2C}, and G_{3C}. Table 2 provides an exhaustive summary of these parameters.

TABLE 1 Parameters used at the design stage of randomized controlled trial for the sample size calculations

Parameters	Values		
Significance level of the overall test	0.05 (two-sided)		
Power of the overall test	0.80		
Proportions of responders, effect size, and sample size per arm	$\pi_c = 0.2$ $\pi_t = 0.25$	$h = 0.12$	$N = 1091$
	$\pi_c = 0.5$ $\pi_t = 0.625$	$h = 0.25$	$N = 244$
	$\pi_c = 0.2$ $\pi_t = 0.4$	$h = 0.44$	$N = 79$

Note: π_c: proportion of responders control arm; π_t: proportion of responders treatment arm; h: Cohen's effect size; N: sample size per arm.

TABLE 2 Simulated proportions of responders for all scenarios

k	$\pi_c = 0.2; \pi_t = 0.25;$ $N = 1091$	$\pi_c = 0.5; \pi_t = 0.625;$ $N = 244$	$\pi_c = 0.2; \pi_t = 0.4;$ $N = 79$
0	$p_{G_{1T}} = 0.2$ $p_{G_{2T}} = 0.2$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$	$p_{G_{1T}} = 0.5$ $p_{G_{2T}} = 0.5$ $p_{G_{3T}} = 0.5$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.5$	$p_{G_{1T}} = 0.2$ $p_{G_{2T}} = 0.2$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$
0.5	$p_{G_{1T}} = 0.225$ $p_{G_{2T}} = 0.2125$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$	$p_{G_{1T}} = 0.5625$ $p_{G_{2T}} = 0.53125$ $p_{G_{3T}} = 0.5$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.5$	$p_{G_{1T}} = 0.3$ $p_{G_{2T}} = 0.25$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$
1	$p_{G_{1T}} = 0.25$ $p_{G_{2T}} = 0.225$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$	$p_{G_{1T}} = 0.625$ $p_{G_{2T}} = 0.5625$ $p_{G_{3T}} = 0.5$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.5$	$p_{G_{1T}} = 0.4$ $p_{G_{2T}} = 0.3$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$
1.5	$p_{G_{1T}} = 0.275$ $p_{G_{2T}} = 0.2375$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$	$p_{G_{1T}} = 0.6875$ $p_{G_{2T}} = 0.59375$ $p_{G_{3T}} = 0.5$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.5$	$p_{G_{1T}} = 0.5$ $p_{G_{2T}} = 0.35$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$
2	$p_{G_{1T}} = 0.3$ $p_{G_{2T}} = 0.25$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$	$p_{G_{1T}} = 0.75$ $p_{G_{2T}} = 0.625$ $p_{G_{3T}} = 0.5$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.5$	$p_{G_{1T}} = 0.6$ $p_{G_{2T}} = 0.4$ $p_{G_{3T}} = 0.2$ $p_{G_{1C}} = p_{G_{2C}} = p_{G_{3C}} = 0.2$

Note: π_c: proportion of responders control arm; π_t: proportion of responders treatment arm; N: sample size per arm; k: parameter related to the simulated proportion of responders for each subgroup (under the null: $k = 0$; under the alternative: $k = \{0.5, 1, 1.5, 2\}$).

The parameter k relates to the simulated proportion of responders for G_{1T} ($p_{G_{1T}}$) and G_{2T} ($p_{G_{2T}}$) as regards π_c and π_t based on the following formulas:

$$p_{G_{1T}} = \pi_c + k(\pi_t - \pi_c)$$

$$p_{G_{2T}} = \pi_c + \frac{k}{2}(\pi_t - \pi_c),$$

and $p_{G_{3T}}$, $p_{G_{1C}}$, $p_{G_{2C}}$, and $p_{G_{3C}}$are always equal to π_c. For instance, with proportions of responders of the binary outcome $\pi_c = 0.2$ in the control group and $\pi_t = 0.25$ in the experimental group, a difference $(\pi_t - \pi_c)$ of 0.05 is anticipated. For $k = 2$, the simulated proportions of responders are $p_{G_{1T}} = 0.3$, $p_{G_{2T}} = 0.25$, and $p_{G_{3T}}$ (as well as $p_{G_{1C}}$, $p_{G_{2C}}$, and $p_{G_{3C}}$) $= 0.2$. The case where $k = 0$ relates to the evaluation of the overall type I error, whereas the remaining cases ($k = 0.5, 1, 1.5, 2$) relate to power.

Per proportion of the subgroup, the overall type I error and the power are assessed by means of simulations for the three combinations of success rates mentioned above. Under the null hypothesis (parameters via $k = 0$), the following events are counted as a false-positive finding: either the overall test is statically significant or the overall is not statistically significant but a trend, as described above (scenarios 1-9), is present. For instance, in scenario 1, if the overall test is statistically significant, or if the overall test is not statistically significant but $RR_1 < RR_2 < RR_3$, then this situation is counted as a false-positive finding. Similarly for scenario 4, if the overall test is statistically significant, or if the overall test is not statistically significant but $RR_1 < RR_2 < RR_3$ and the test of interaction ($p_{IT} < 0.10$) is statistically significant, this situation is counted as a false-positive finding. Regarding the power computation under the alternative hypothesis, the same principle has been applied. Under the alternative hypothesis, a positive finding is when the overall null hypothesis is *not* rejected and a trend is found as defined as per the scenario considered. All simulations (data generation and analyses) were done with R version 3.1.1.[17] To report the results with a precision of two digits, the standard error needs to be ≤ 0.005 ($1.96 \times 0.005 \approx 0.01$). To reach this precision, the number of simulations, with both a statistically nonsignificant overall result and a "positive" trend as defined for each scenario, is 10 000. In order to make this study reproducible, R code (scenarios 1, 4, and 7) is provided in the supplementary material.

3.1 | Results of the observed trend across subgroups

Results regarding the overall type I error ($k = 0$) and the power ($k = 0.5; 1; 1.5; 2$) of each strategy are presented in Table 3. For the sake of clarity, only results for subgroup proportions $r_{G_{1.}} = 0.33$ and $r_{G_{1.}} = 0.5$ are presented. As regards the power results, only $k = 1$ and $k = 2$ were considered in Table 3. All other results can be found in the supplementary material.

By looking only at the observed trend across subgroups, the overall type I error is substantial, ie, around 0.20. Irrespective of the metric used, the results are roughly similar. The overall type I error decreases to approximately 0.09 when the

TABLE 3 Overall type I error ($k = 0$) and power ($k = 1; 2$) for scenarios 1-10

			$N = 1091$		**$N = 244$**		**$N = 79$**	
Strategy	**Metric**	***k***	**$r_{G_{1.}} = 0.33$**	**$r_{G_{1.}} = 0.5$**	**$r_{G_{1.}} = 0.33$**	**$r_{G_{1.}} = 0.5$**	**$r_{G_{1.}} = 0.33$**	**$r_{G_{1.}} = 0.5$**
Observed	RR	0	0.22	0.20	0.21	0.20	0.20	0.19
		1	0.46	0.43	0.47	0.44	0.46	0.42
		2	0.74	0.69	0.80	0.75	0.75	0.68
	OR	0	0.22	0.20	0.21	0.20	0.20	0.18
		1	0.44	0.43	0.46	0.43	0.42	0.38
		2	0.72	0.68	0.78	0.73	0.69	0.61
	RD	0	0.21	0.19	0.21	0.19	0.18	0.17
		1	0.44	0.42	0.44	0.42	0.42	0.39
		2	0.73	0.68	0.75	0.71	0.70	0.64
Observed + interaction ($p < 0.10$)	RR	0	0.09	0.08	0.09	0.09	0.09	0.08
		1	0.24	0.24	0.26	0.25	0.26	0.24
		2	0.62	0.59	0.71	0.67	0.62	0.59
	OR	0	0.09	0.09	0.09	0.08	0.09	0.08
		1	0.24	0.22	0.25	0.23	0.22	0.20
		2	0.58	0.55	0.69	0.66	0.54	0.48
	RD	0	0.09	0.08	0.09	0.08	0.09	0.08
		1	0.24	0.23	0.25	0.23	0.24	0.21
		2	0.60	0.57	0.66	0.64	0.59	0.54
Observed + subgroup	RR	0	0.06	0.06	0.06	0.06	0.06	0.06
		1	0.17	0.19	0.18	0.19	0.17	0.16
		2	0.56	0.58	0.65	0.67	0.51	0.54
	OR	0	0.06	0.06	0.06	0.06	0.06	0.06
		1	0.17	0.19	0.17	0.19	0.17	0.17
		2	0.56	0.57	0.64	0.66	0.51	0.51
	RD	0	0.06	0.06	0.06	0.06	0.06	0.06
		1	0.17	0.19	0.17	0.19	0.17	0.16
		2	0.56	0.57	0.64	0.65	0.51	0.53
Subgroup	NA	0	0.09	0.09	0.10	0.08	0.10	0.08
		1	0.25	0.29	0.25	0.29	0.24	0.27
		2	0.66	0.73	0.75	0.84	0.61	0.71

Note: N: sample size per arm; $r_{G_{1.}}$: proportion (over the total sample) of the subgroup of interest $G_{1.}$; k: parameter related to the simulated proportion of responders for each subgroup (under the null: $k = 0$; under the alternative: $k = \{0.5, 1, 1.5, 2\}$); NA: not applicable; OR: odds ratio; RD: risk difference; RR: relative risk.

observed trend across subgroups is supported by an interaction test when investigating RR, OR, or RD. Depending on the choice of the test of interaction significance level, this value can be lowered to 0.07 when $p_{IT} < 0.05$ and increases to 0.12 when $p_{IT} < 0.20$. Finally, the overall type I error is almost controlled, ie, 0.06, irrespective of the metric chosen, when the observed trend across subgroups is supported by a statistically significant test for the effect in the most promising subgroup.

By looking only at the observed trend across subgroups, the power is about 0.45 when $k = 1$ and about 0.75 when $k = 2$. The power is about 0.25 ($k = 1$) and 0.60 ($k = 2$) when the observed trend across subgroups is supported by an interaction test. The power slightly decreases to around 0.18 ($k = 1$) and 0.55 ($k = 2$) when the observed trend across subgroups is supported by a statistically significant test for the effect in the most promising subgroup. Of note, it seems that the RR metric provides slightly better results than the OR and RD metrics. Moreover, and for $k = 2$, the results are slightly better when the population proportions of responders of the outcome for the control and treatment arms, π_c and π_t, are respectively equal to 0.5 and 0.625. It is also interesting to observe that the scenario where the observed trend across subgroups is supported by a statistically significant test for the effect in the most promising subgroup has a better control of the overall type I error as well as a slightly higher power as compared to the scenario where the observed trend across subgroups is supported by an interaction test with a stringent $p_{IT} < 0.05$. This latter scenario should therefore not be preferred in this setting, or only with a relaxed test of interaction significance level if power is preferred over type I error control.

4 | $m(m \geq 2)$ SUBGROUPS ARE INVESTIGATED

The number of subgroups analysed in practice is generally large. Thus, both multiplicity as well as correlation between test statistics determine the type I error. According to Pocock et al, 70% (35/50) of reported trials contained multiple subgroup analyses.[8] The total number of reported subgroup analyses per published trial varied from one to 24 with a median of four. Wang et al reviewed 59 clinical trials, of which 20 provided between one and four subgroup analyses, 17 between five and eight subgroup analyses, 17 more than eight subgroup analyses and five were unclear about the number of subgroup analyses performed.[9] As not all subgroup analyses performed are always reported nor are protocol defined, we considered the following number of investigated subgroups: 2, 3, 5, 10, and 20. For simplicity, we looked at cases where all m subgroups, each coming from a covariate with two levels, have the same proportion r for the subgroup under investigation. We first simulated the whole trial sample and then randomly resampled it in order to create the (m) subgroups "of interest." As all subgroups are based on the same overall sample, test statistics for H_i and H_j $(i \neq j; (i;j) = \{1, \ldots, m\}; m \geq 2)$ are positively correlated, such as what would happen in clinical practice where the same patients are included in different subgroups. Under the null hypothesis (no overall effect as well as no effect in any subgroup) and when the null is *not* rejected, if at least one subgroup, out of m subgroups tested, reaches statistical significance, ie, at least one $p_{ST}^{(i)} < 0.05$ $(i = 1, \ldots, m, m \geq 2)$, it is considered as a false-positive finding, hence is taken into account for the assessment of the overall type I error. The R code is provided in the supplementary material.

4.1 | Results when $m(m \geq 2)$ subgroups are investigated

Irrespective of the size of the trial, the results are roughly similar (Table 4). They only differ when the subgroups of interest are very small, ie, for $r = 0.1$. Practically, the most encountered scenario is expected when $r = 0.4$, 0.5, or 0.6. When two subgroups are investigated, the overall type I error is around 0.08. When five subgroups are investigated, this number reaches 0.13. For 10 and 20 subgroups, the associated overall type I errors are respectively around 0.17 and 0.23.

These results directly relate to the multiple testing problem. Despite a substantial inflation of the overall type I error across all scenarios, these results are, however, more conservative than it would be with independent tests, as the positive correlation between the subgroup test statistics attenuates the type I error inflation. For instance, when 10 subgroups are tested after an overall statistically nonsignificant result, the overall type I error is approximately 0.17, ie, an inflation of 0.12 based on 10 subgroup tests given the overall statistically nonsignificant setting. However, if these tests had been completely independent, the inflation would have been approximately 0.22. Therefore, given the positive correlation between the subgroup test statistics, the inflation for multiple testing is less dramatic than one would expect. Obviously, this observation is not a convincing argument to make firm conclusions about a subgroup finding. In order

TABLE 4 Overall type I errors

Sample Size Per Arm	Number of Subgroups	Proportion of the Subgroup G_1								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$N = 1091$	2	0.093	0.090	0.087	0.085	0.082	0.078	0.074	0.069	0.063
	3	0.113	0.109	0.105	0.102	0.096	0.090	0.084	0.077	0.068
	5	0.154	0.146	0.136	0.131	0.120	0.111	0.100	0.090	0.076
	10	0.239	0.226	0.205	0.188	0.171	0.149	0.132	0.112	0.089
	20	0.383	0.337	0.300	0.272	0.234	0.201	0.172	0.138	0.102
$N = 244$	2	0.109	0.096	0.097	0.090	0.081	0.083	0.073	0.070	0.065
	3	0.136	0.117	0.118	0.106	0.095	0.096	0.082	0.077	0.070
	5	0.186	0.155	0.154	0.137	0.117	0.118	0.097	0.088	0.078
	10	0.298	0.237	0.231	0.202	0.162	0.160	0.127	0.109	0.090
	20	0.455	0.360	0.338	0.283	0.230	0.214	0.160	0.134	0.104
$N = 79$	2	0.074	0.094	0.092	0.086	0.083	0.080	0.076	0.069	0.065
	3	0.085	0.113	0.112	0.103	0.098	0.093	0.087	0.075	0.069
	5	0.107	0.151	0.145	0.132	0.123	0.114	0.104	0.088	0.077
	10	0.159	0.235	0.219	0.189	0.173	0.154	0.135	0.108	0.089
	20	0.245	0.359	0.322	0.272	0.241	0.206	0.174	0.134	0.104

Note: Small inflation of the type I error: $0.05 \leq p < 0.06$; Medium inflation of the type I error: $0.06 \leq p < 0.07$; Large inflation of the type I error: $0.07 \leq p < 0.08$; Very large inflation of the type I error: $p \geq 0.08$.

TABLE 5 Adjusted overall type I errors with a Holm-Bonferroni correction

Sample Size Per Arm	Number of Subgroups	Proportion of the Subgroup G_1								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$N = 1091$	2	0.064	0.062	0.061	0.059	0.058	0.058	0.055	0.053	0.051
	3	0.065	0.064	0.062	0.060	0.059	0.057	0.055	0.052	0.051
	5	0.067	0.066	0.063	0.062	0.059	0.056	0.054	0.051	0.050
	10	0.067	0.066	0.065	0.060	0.058	0.056	0.052	0.051	0.050
	20	0.067	0.065	0.063	0.058	0.057	0.054	0.051	0.050	0.050
$N = 244$	2	0.067	0.067	0.065	0.063	0.062	0.059	0.056	0.055	0.053
	3	0.071	0.065	0.064	0.062	0.061	0.059	0.056	0.055	0.052
	5	0.068	0.066	0.065	0.063	0.060	0.057	0.056	0.053	0.052
	10	0.074	0.067	0.065	0.064	0.058	0.057	0.054	0.053	0.052
	20	0.073	0.067	0.065	0.060	0.060	0.057	0.054	0.052	0.052
$N = 79$	2	0.055	0.066	0.062	0.061	0.058	0.057	0.054	0.054	0.052
	3	0.056	0.057	0.062	0.062	0.060	0.058	0.055	0.053	0.051
	5	0.057	0.062	0.061	0.061	0.058	0.056	0.055	0.052	0.050
	10	0.053	0.056	0.064	0.059	0.057	0.054	0.053	0.051	0.051
	20	0.053	0.056	0.061	0.056	0.055	0.054	0.052	0.051	0.051

Note: Small inflation of the type I error: $0.05 \leq p < 0.06$; Medium inflation of the type I error: $0.06 \leq p < 0.07$; Large inflation of the type I error: $0.07 \leq p < 0.08$; Very large inflation of the type I error: $p \geq 0.08$.

to have more confidence in the subgroup finding(s), adjustment for multiplicity should be undertaken. The step-down Holm-Bonferroni procedure[18] could be a first option. Table 5 provides the corresponding adjusted overall type I error.

The results of the Holm-Bonferroni correction are rather good, irrespective of the number of subgroups investigated. A significant subgroup finding after Holm-Bonferroni multiplicity correction could be considered as promising and therefore be a good candidate for identifying subgroups that could be considered for a replication study, if not yet available. Of course, the number of subgroups tested should be known, or otherwise, it should be stated that the subgroup finding should not be outside the list of those mentioned in the protocol.

In previous research, Tanniou et al suggested that, if the p-value for the subgroup of interest is very low (< 0.004 one-sided), this is not consistent with the null hypothesis of no subgroup effect across a range of scenarios, despite the type I error inflation.[3] We acknowledge that pragmatic decisions based on thresholds, such as the classic "$p < 0.05$," are

TABLE 6 Overall type I errors with a subgroup significance level of 0.004 (one-sided)

Sample Size Per Arm	Number of Subgroups	Proportion of the Subgroup G_1 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$N = 1091$	2	0.053	0.053	0.053	0.052	0.051	0.051	0.050	0.050	0.050
	3	0.055	0.055	0.053	0.053	0.052	0.051	0.051	0.050	0.050
	5	0.057	0.057	0.056	0.054	0.054	0.052	0.052	0.050	0.050
	10	0.063	0.063	0.061	0.060	0.057	0.054	0.053	0.051	0.050
	20	0.080	0.079	0.072	0.069	0.063	0.058	0.054	0.052	0.050
$N = 244$	2	0.055	0.054	0.054	0.054	0.053	0.053	0.052	0.052	0.052
	3	0.056	0.056	0.055	0.055	0.053	0.053	0.053	0.052	0.052
	5	0.060	0.059	0.058	0.056	0.055	0.054	0.053	0.052	0.052
	10	0.068	0.064	0.062	0.061	0.057	0.056	0.055	0.053	0.052
	20	0.079	0.076	0.074	0.068	0.064	0.060	0.058	0.053	0.052
$N = 79$	2	0.051	0.052	0.052	0.052	0.052	0.052	0.051	0.050	0.050
	3	0.051	0.052	0.053	0.053	0.053	0.052	0.051	0.050	0.050
	5	0.051	0.055	0.054	0.054	0.053	0.053	0.051	0.051	0.050
	10	0.053	0.056	0.056	0.058	0.058	0.055	0.052	0.051	0.051
	20	0.054	0.064	0.063	0.066	0.064	0.059	0.054	0.052	0.050

Note: Small inflation of the type I error: $0.05 \leq p < 0.06$; Medium inflation of the type I error: $0.06 \leq p < 0.07$; Large inflation of the type I error: $0.07 \leq p < 0.08$; Very large inflation of the type I error: $p \geq 0.08$.

often more consensus-based than purely scientifically-based. In the same spirit, we do not recommend to base a subgroup finding only on the pragmatic boundary proposed in previous research. This pragmatic threshold should rather be seen as one of the many factors playing a role to come to a scientifically informed decision, which necessarily needs to include the biological plausibility as well as other nonstatistical considerations. We were therefore interested in studying the impact on the overall type I error by applying this very practical adjusted criterion (Table 6). In the same spirit, Benjamin et al have recently made a plea for smaller p-value thresholds, ie, 0.005.[19]

By applying the strict threshold of 0.004, the correction of the overall type I error for small subgroups ($r \leq 0.5$) provides closer results to the intended control of type I error than the Holm-Bonferroni correction. When the number of investigated subgroups increases (≥ 20), the Holm-Bonferroni provides a better correction of the type I error. Both presented multiplicity correction approaches have good results, can be implemented in practice in a straightforward way, and can be used post hoc, or even when assessing a paper in which multiple subgroup analyses are presented. Moreover, one additional advantage is that the criterion based on a stricter subgroup significance level (0.004) remains the same, regardless the number of subgroups investigated, eg, 10, 50, or 100. Both approaches could potentially be applied to decide whether a replication study should be undertaken or not. However, more powerful methods taking into account the correlation between the test statistics could apply, such as resampling-based approaches. For more details about these approaches, we refer to the work done by Bretz et al.[20] The principal drawback of these more advanced methods is the practical implementation and limitations in post hoc application.

5 | DISCUSSION

When no significant overall treatment effect is found in a clinical trial, subgroup analyses could be conducted to investigate whether a treatment might be beneficial for particular subgroups. Following strict (frequentist) principles, if the overall effect is not statistically significant, drawing confirmatory conclusions from any significant subgroup findings is not possible as the type I error is exhausted. In case subgroup analyses are conducted in an exploratory way, such as for checking consistency, they are associated with an increased overall type I error. Moreover, in case of no significant overall treatment effect and the assumption of homogeneity, we are logically forced to retain the null hypothesis of no effect for the whole study population, irrespective of (promising) subgroups. However, if the treatment effect is heterogeneous across the study population, with part of the population showing benefit and another part not, a subgroup might still benefit from the treatment under investigation. Subgroup analyses can thus be in the interest of patients. This study expands the results of the previous work done by Tanniou et al[3] by quantifying the inflation of the overall type I error for a range of new scenarios with dichotomous outcomes. It presents strategies when the covariate of interest is divided into three *ordered* subgroups and where a potential treatment effect trend over subgroups might be anticipated. Assuming an overall nonsignificant result, the strategy providing the best results (amongst those investigated) is when an observed trend across subgroups, based on point estimates only and not on any confidence intervals, and a statistically significant subgroup treatment effect is observed. This strategy, irrespective of the metric chosen, ie, (log) ORs, RDs, or (log) RRs, provides an overall type I error close to the initial significance level and outperforms the commonly used practice of an interaction test. The concept of credibility is of paramount importance here. If a different (nonmonotonic) response was to be expected, such as a U-shaped response, the concept of trends as well as the scenarios discussed in this paper would not be appropriate and would hypothetically not outperform the test of interaction. The associated power can be up to 65% if the subgroup effect is twice the overall anticipated effect at the design stage. It should be noted that these results are based on simulations where one single covariate is investigated. It should also be highlighted that the scientific (clinical) plausibility should play a major role when assessing signals based on subgroup analyses. Finally, to assess a potential differential effect between subgroups, a test of interaction, a test of heterogeneity, or a test for trend could be used. As these tests have different purposes, the choice of the test to be used should be properly justified by researchers. In order to remain as close as possible to current clinical practice, we have only considered the test of interaction. However, it is important to emphasise that the choice of the test may provide inconsistent conclusions.[21]

The next focus was on the investigation of scenarios where the interest is not only in one covariate, but in multiple covariates, from 2 to 20, at the same time. This directly relates to a multiple testing problem with dependent subgroups, ie, the test statistics of the subgroup analyses will be positively correlated, because all patients will usually appear in each subgroup of interest. This strategy is close to the performance of data dredging when the overall test failed to show any potential effect for the investigated sample. As expected, the higher the number of subgroups, the larger the inflation of the overall type I error, although attenuated by the positive correlation of subgroup test statistics. Therefore, the significance of within-subgroup treatment effects should be adjusted for multiplicity when multiple subgroup analyses are performed simultaneously. We propose pragmatic solutions such as the Holm-Bonferroni adjustment method[18] and an adjusted criterion for a false-positive finding based on our results with continuous outcomes.[3] These solutions perform adequately, but we acknowledge that more efficient methods could be used, for example, those that exploit the positive correlation between subgroup tests. However, these might encounter some reluctance in practice as they cannot be easily implemented nor be applied post hoc when evaluating trials results as reported.

Based on the simulation results, it is interesting to return to the example about the Dexamethasone for Cardiac Surgery trial presented in Section 2.[13] No overall significant effect was observed, but a trend across age-related subgroups was observed in patients younger than 65 years of age; a protective effect of prophylactic Dexamethasone on the primary endpoint appeared statistically significant with an RR of 0.65 (95% CI, 0.44 to 0.96). Furthermore, the publication reported six investigations related to subgroups, only one with *ordered* subgroups. It is interesting to note that the authors provided the results in RRs while a logistic regression analysis, providing (log) ORs, was performed to assess if heterogeneity was present or not. It should be emphasised that even though the metric chosen, ie, (log) ORs, RDs, or (log) RRs, provides on average roughly the same results in the presented simulations, researchers should be aware that the choice of the metric used should be consistent to perform a proper evaluation, as it does impact the interaction test results, ie, within one specific clinical trial.[14] According to the results presented in this study, the investigations of five subgroups would provide an overall type I error of about 0.10-0.15. Moreover, observing a trend over the RR metric as well as a statistically significant subgroup has a low inflation of the type I error, ie, about 0.06 (with only one subgroup investigated). The latter result is, however, based on scenarios where ordered subgroups all have a nonnegative effect. It should also be emphasised that other aspects, such as the definition of cut-points to define the different subgroups, were not considered, even though those were pre-specified. Finally and based on our study, we would not discard the trend across age-related subgroups directly. The best strategy based on our results is a trend based on point estimates and a statistically significant subgroup, which is the case here. Hence, based on what we investigated and following our own recommended approach, the finding on age would warrant further investigation. Also, and because we are dealing with ordered subgroups, the credibility of this subgroup finding may be reinforced.

Our study, together with previous research, provides an understanding of subgroup findings in “failed” trials. Some limitations need to be addressed. The main one relates to the multiple testing problem and how to evaluate the subgroup finding knowing (or not) the number of subgroup analyses performed. Also, the model we used to investigate subgroup findings assumes that there is a positive effect in the subgroups (or not) and that the effect in the rest of the trial population is nonnegative, so the situation of an overall failed trial due to a positive effect in some subgroups paired with a negative effect, ie, detrimental, in the rest of the trial population is not covered.

To conclude, subgroup analyses in overall nonsignificant trials are likely to cause statistical problems but should still be considered as a potential new relevant treatment may benefit a subpopulation. Therefore, particular attention should be given to plausibility and replication in order to reassure decision makers. This study provides practical results to better evaluate whether the level of evidence associated with subgroup findings is sufficient to perform further research.

DECLARATION OF CONFLICTING INTERESTS

The authors declare that there is no conflict of interest.

DISCLAIMER

The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties.

ORCID

Julien Tanniou https://orcid.org/0000-0002-3806-2413

REFERENCES

1. International Conference on Harmonisation E9 Expert Working Group. ICH harmonised tripartite guideline: statistical principles for clinical trials. *Statist Med*. 1999;18(15):1905-1942.
2. Tanniou J, Teerenstra S, Hassan S, et al. European regulatory use and impact of subgroup evaluation in marketing authorisation applications. *Drug Discov Today*. 2017;22(12):1760-1764.
3. Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. Level of evidence for promising subgroup findings in an overall non-significant trial. *Stat Methods Med Res*. 2016;25(5):2193-2213.
4. Koch A, Framke T. Reliably basing conclusions on subgroups of randomized clinical trials. *J Biopharm Stat*. 2014;24(1):42-57.
5. Hemmings R. An overview of statistical and regulatory issues in the planning, analysis, and interpretation of subgroup analyses in confirmatory clinical trials. *J Biopharm Stat*. 2014;24(1):4-18.
6. EMA/CHMP/539146/2013, Committee for Medicinal Products for Human Use CHMP. Guideline on the investigation of subgroups in confirmatory clinical trials. 2014. http://www.ema.europa.eu.proxy.library.uu.nl/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf
7. Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC Med Res Methodol*. 2016;16(1):20. https://doi.org/10.1186/s12874-016-0122-6
8. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statist Med*. 2002;21(19):2917-2930.
9. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357(21):2189-2194.
10. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. *Statist Med*. 2007;26(19):3535-3549.
11. Spiessens B, Debois M. Adjusted significance levels for subgroup analyses in clinical trials. *Contemp Clin Trials*. 2010;31(6):647-656.
12. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. *Statist Med*. 2009;28(1):3-23.
13. Dieleman JM, Nierich AP, Rosseel PM, et al. Intraoperative high-dose dexamethasone for cardiac surgery: a randomized controlled trial. *JAMA*. 2012;308(17):1761-1767.
14. White IR, Elbourne D. Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusions? *BMC Med Res Methodol*. 2005;5:15
15. Van der Tweel I. *Sample Size Determination* [Technical Report]. Utrecht, Netherlands: Department of Biostatistics, Julius Centre for Health Sciences and Primary Care; 2006.
16. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
17. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN:3-900051-07-0. http://www.R-project.org
18. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65-70.
19. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2017. https://www.nature.com/articles/s41562-017-0189-z
20. Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*. New York, NY: Chapman and Hall; 2010.
21. Dehbi HM, Hackshaw A. Investigating subgroup effects in randomized clinical trials. *J Clin Oncol*. 2017;35(2):253-254.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Tanniou J, Smid SC, van der Tweel I, Teerenstra S, Roes KCB. Level of evidence for promising subgroup findings: The case of trends and multiple subgroups. *Statistics in Medicine*. 2019;38:2561–2572. https://doi.org/10.1002/sim.8133