Three simple rules to ensure reasonably credible subgroup analyses

James F Burke,[1,2] Jeremy B Sussman,[2,3] David M Kent,[4,5] Rodney A Hayward[2,3]

[1]Department of Neurology, University of Michigan School of Medicine, Ann Arbor, MI 48109-2800, USA
[2]VA Center for Clinical Management and Research, Ann Arbor
[3]Department of Internal Medicine, University of Michigan School of Medicine
[4]Department of Internal Medicine, Tufts University School of Medicine, Boston, MA, USA
[5]Predictive Analytics and Comparative Effectiveness Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Tufts University School of Medicine

Correspondence to: J F Burke
jamesbur@umich.edu

Cite this as: *BMJ* 2015;351:h5651
doi: 10.1136/bmj.h5651

Accepted: 2 October 2015

The limitations of subgroup analyses are well established—false positives due to multiple comparisons, false negatives due to inadequate power, and limited ability to inform individual treatment decisions because patients have multiple characteristics that vary simultaneously. In this article, we apply Bayes's rule to determine the probability that a positive subgroup analysis is a true positive. From this framework, we derive simple rules to determine when subgroup analyses can be performed as hypothesis testing analyses and thus inform when subgroup analyses should influence how we practice medicine.

A table or figure reporting about a dozen subgroup analyses is a near ubiquitous feature of major clinical trial publications.[1 2] The motivation behind these analyses is clear and compelling—to determine which patients most benefit from treatment, based on specific risk factors. However, the limitations of these analyses are well established—false positives due to multiple comparisons, false negatives due to inadequate power, and limited ability to inform individual treatment decisions because patients have multiple characteristics that vary simultaneously.[3] When, if ever, should subgroup analyses, tested using subgroup treatment interactions, influence how we practice medicine?

SUMMARY POINTS

Limitations of subgroup analyses are well established—false positives due to multiple comparisons, false negatives due to inadequate power, and limited ability to inform individual treatment decisions because patients have multiple characteristics that vary simultaneously. It remains uncertain when subgroup analyses should influence clinical practice

Categorical subgroup analyses should not be part of a typical clinical trial's hypothesis testing analysis unless the prior probability for a subgroup effect being present is at least 20% and preferably higher than 50%

Rarely should more than one to two primary categorical subgroup analyses be performed

In trials with exceptional power to identify subgroup effects, hypothesis testing analyses of subgroups should be justified a priori

Contrary to common belief, the well documented unreliability of subgroup analyses are not inherent; the same problems would arise for clinical trials themselves if we routinely performed underpowered trials examining haphazardly selected interventions. If properly selected (based on previous empirical evidence and current scientific theory), an adequately powered subgroup analysis can be a valid hypothesis testing endeavour. Trialists, reviewers, and editors should carefully consider such issues when making the essential scientific distinction between primary (that is, hypothesis testing) and secondary (that is, hypothesis generating) subgroup analyses.[4] A positive, hypothesis testing analysis can directly influence patient care whereas a positive hypothesis generating analysis only calls for confirmatory research.

There are excellent general discussions of subgroup analyses and checklists to evaluate their credibility,[3 5-7] but in this paper, we will quantitatively explore the key pragmatic question of when a subgroup analysis should be considered hypothesis testing versus hypothesis generating. We used simulation modelling to derive simple quantitative rules of thumb that can be applied by trialists, reviewers, and editors to ensure that subgroup analyses are properly contextualised and used by readers to quickly evaluate the credibility of a specific subgroup finding.

Predictive value of subgroup analyses as diagnostic tests: an analogy

Interpretation of a subgroup analysis is analogous to rigorously interpreting a diagnostic test. Before ordering a diagnostic test, a clinician considers the probability the person has the condition (the prior probability) and the accuracy of the test (often measured with sensitivity and specificity). With this information, the probability that a positive test is a true positive versus false positive can be estimated using Bayes's rule: posterior odds=sensitivity÷(1–specificity)×prior odds.

Bayes's rule can be seamlessly applied to the context of subgroup analysis, and informs why a shotgun approach to subgroup analysis fails. The sensitivity of a subgroup analysis is its statistical power: the probability of finding a true difference between groups if one exists. Most large clinical trials are powered to find a clinically meaningful difference between treatment and control groups around 80-90% of the time. Compared with the power for the trial's main effect, most subgroup analyses have much less statistical power to identify subgroup effects. Power might often be closer to 20-30% for subgroup effect sizes similar in magnitude to the main treatment effect sizes (that is, a relative odds ratio for a subgroup treatment that is equal to the

odds ratio for the overall treatment)[8 9] Thus, the sample size needed to adequately contrast treatment effects measured in two different subgroups is much larger than the sample needed to distinguish an overall treatment effect from the null. Just as statistical power can be thought of as the sensitivity of a trial, the specificity of clinical trials is generally set to be 95%, based on the conventional significance threshold of P<0.05.

Finally, an estimate of the prior probability is needed to interpret a subgroup analysis. In both diagnostic testing and subgroup analyses, prior probability estimates are often unsettling given their inherent uncertainty and subjectivity, but failing to grapple with this tends to bias us towards falsely accepting new evidence as truth.[10] Existing criteria to judge the credibility of subgroup analyses emphasise the importance of prior probability and specifically require that a hypothesis and its direction of effect are specified a priori and that the subgroup effect is supported by within-study empirical and biological evidence.[6] In most cases, prior probability can be roughly estimated by thinking about the strength of previous theoretical or empirical evidence that the factor in question is likely to modify the relative treatment effect.[11] Just as power for subgroups is usually much lower than for the main effect, so are the prior probabilities.

Expensive trials can only be justified if there is a reasonable probability of success based on prior data. In contrast, subgroup analyses with low priors are commonly conducted, perhaps because they are perceived as being essentially free, but as is shown below, conducting multiple subgroup analyses is statistically costly. We can also use empirical data as a rough starting point for thinking about prior probability. Of roughly 1200 subgroup analyses of recent clinical trials published in high impact journals, 83 (7%) were reportedly positive.[1] Assuming a 5% false positive rate, only a fraction of these analyses were likely true positives. This observation is supported by the observation that less than 15% of these subgroup analyses met four of 10 criteria for credibility. So, a high-end starting point for the prior probability for the average published subgroup analysis is probably around 5%, which can be adjusted on a case by case basis, based on the prior empirical and theoretical evidence.

As per Bayes's rule, low prior probabilities greatly increase the chance of a positive result being a false positive finding, and low power greatly increases this problem. Back to our analogy, the same phenomenon explains why ordering insensitive diagnostic tests with a low pretest probability leads to most positive test results being false positives (table 1).[12]

Table 1 | Comparison between diagnostic tests and subgroup analyses

	Diagnostic testing	Subgroup analyses
Prior probability	Based on population prevalence and clinical factors	Based on previous clinical evidence and pathophysiological rationale
Test accuracy	Sensitivity	Statistical power
	Specificity	1–α

Application to subgroup analyses

Using this framework that has been applied in related contexts,[13] we can calculate the probability of a positive subgroup analysis (treatment subgroup interaction) being a true positive versus false positive using Bayes's rule.[13] A positive analysis, in this context, refers to a significant difference in treatment effect between groups such that some groups can be demonstrated to have a greater or lesser relative treatment effect than other groups. Figure 1 illustrates the association between prior probability and positive predictive value (the chance that a trial reporting a statistically significant result is not reporting a false positive) when subgroup statistical power is varied. In rare scenarios where there is excellent subgroup power, positive results can be highly reliable, but even in this ideal situation the positive predictive value drops precipitously when prior probability drops below 20% to 30% or when multiple subgroup analyses are performed without adjusting for multiple comparisons.

Table 2 quantifies the positive predictive value for statistically significant subgroup findings (that is, treatment subgroup interaction effects) over a wide range of prior probabilities and number of subgroup comparisons. Although we know of no established conventions for acceptable positive predictive values, for illustrative purposes we apply a minimum threshold

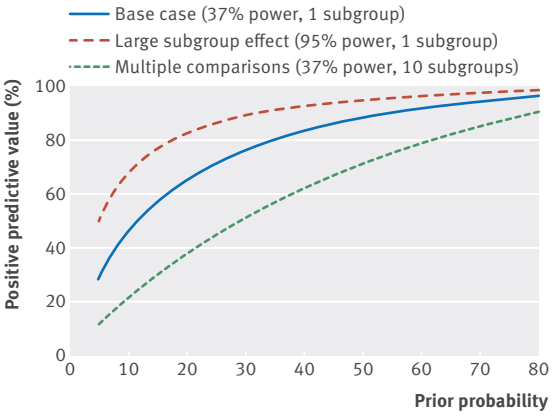

Fig 1 | Association between prior probability and positive predictive values for subgroup analyses. The base case represents an uncommonly well powered, categorical subgroup analysis (that is, evenly divided subgroups, the effect the overall trial is powered for is entirely present in one subgroup, there is no effect is present in the other subgroup, and the trial has 90% power overall to find its primary effect), resulting in 37% power to find the subgroup effect. The probability of a true positive finding can be either reduced or improved by changing these assumptions. The lower line (multiple comparisons) illustrates how positive predictive values decline if 10 subgroup analyses are performed with 37% power to identify each subgroup effect and with no adjustment for multiple comparisons (as opposed to the one analysis illustrated in the base case). Conversely, the higher line (large subgroup effect) illustrates a scenario where the positive predictive values increase as power is improved—the effect size in the subgroup is twice as large as the base case (subgroup power=95%)

doi: 10.1136/bmj.h5651 | *BMJ* 2015;351:h5651 | the bmj

of an 80% positive predictive value (that is, accepting a one in five chance (20%) of a positive finding being false). As this threshold is arbitrary and should differ depending on context,[14] our results are presented so that readers can choose whatever threshold they wish. These analyses are based on P values of 0.05. When P values for subgroups are lower than 0.05, the positive predictive values will be higher. Like all rules of thumb, these three rules are proposed as only a starting point for thinking through individual cases.

Rules of thumb for performing primary one-at-a-time subgroup analyses

Rule of thumb 1

Categorical subgroup analyses should not be part of a typical clinical trial's primary (hypothesis testing) analysis unless the prior probability for a subgroup effect being present is at least 20% and preferably higher than 50%. Even under optimal circumstances, a subgroup analysis of a categorical variable will rarely have greater than 50% statistical power to detect a moderate subgroup effect, and more often is closer to 20%.[9] In general, for a modestly powered subgroup effect (that is, 20% power) unless there is a strong prior probability (that is, 50%), the chance of a positive subgroup effect being a true positive finding will be less than 80% (table 2). Even when subgroup effect power is excellent (that is, 80%), to reach an 80% true positive threshold requires prior probability of 20%.

One seeming exception to this rule is that we recommend routinely conducting multivariable risk based analyses of trials with positive overall results and of negative trials in which the intervention has known harms (such as major surgery). We have so far only considered relative subgroup effects (for example, one group benefits from treatment while another group does not benefit). Risk based analyses examine how both the relative and absolute benefit of interventions varies between patients at high, medium, and low pretreatment risk. Even if a credible relative subgroup exists, it does not imply that treatment decisions should be different across subgroup levels, since treatment could still be worthwhile (or not worthwhile) in both groups. Further, risk based analyses can inform care even when the prior probability for a risk based subgroup effect is low.[11] Even if no such effect exists, the absolute risk reduction with treatment will be higher in patients at high pretreatment risk and these analyses can quantify the magnitude of those differences.

Table 2 | Positive predictive values (%)* for significant subgroup findings according to prior probability and number of subgroup comparisons

Prior probability (%)	Power of subgroup comparison and no of comparisons								
	20% power			50% power			80% power		
	1	5	10	1	5	10	1	5	10
5	17	14	11	35	18	12	46	19	12
10	31	25	20	53	32	22	64	33	22
20	50	43	36	71	52	38	80	53	38
30	63	56	49	81	65	52	87	65	52
40	73	67	60	87	74	62	91	75	62
50	80	75	69	91	81	71	94	82	71
60	86	82	77	94	87	79	96	87	79
70	90	87	84	96	91	85	97	91	85
80	94	92	90	98	95	91	99	95	91

*Positive predictive values=probability that all reported positives analyses are true positives for a trial reporting at least one positive subgroup effect (that is, no false positives) for a given prior probability and power in the context of conducting one, five, or 10 subgroup comparisons without adjustment for multiple comparisons, assuming α=5% (0.05). In formal Bayesian statistical analyses, the priors and posteriors are generally presented as probability distributions, but we have represented both as fixed values for simplicity. Estimated using approach of Ioannidis.[13]

Rule of thumb 2

Rarely should more than one to two primary categorical subgroup analyses be performed. When one study examines treatment differences across multiple risk factors (multiple comparisons), the likelihood that a study with a reported subgroup effect has at least one false positive result increases, thereby eroding the reliability of all positive findings. As shown in figure 1, the reliability of a positive finding falls considerably across the spectrum of prior probability when increasing the number of subgroup analyses. Even with an above average prior probability for a given subgroup comparison (that is, 20%) and excellent statistical power to find a subgroup effect (that is, 80%), the positive predictive value would fall from 80% if one comparison is made to 38% if 10 comparisons are made (table 2). Another, potentially less appreciated, consequence of performing multiple underpowered subgroup comparisons is that readers can be misled into believing that these negative results provide reliable evidence that the treatment is similarly effective in all patients, potentially masking important subgroup effects owing to inadequate power.[9]

Correction for multiple comparisons decreases the risk of false positives, but does not eliminate the broader problem. Firstly, correcting for multiple comparisons decreases statistical power, thus increasing the risk of false negative findings further. Secondly, as investigators add subgroup analyses with less and less evidence or theory to support them, the prior probability for the average subgroup analysis inevitably falls, further reducing the positive predictive value (fig 2).

Rule of thumb 3

In trials with exceptional power to identify subgroup effects, hypothesis testing subgroup analyses should be justified a priori. Various trial circumstances can lead to increased power to identify subgroup effects (table 3). If, for example, the effect size difference between subgroups is larger than that found for the overall trial population, the probability of a true positive subgroup effect increases (fig 1). Although such effects are likely uncommon, they are not unheard of—and they are the effects that are of most clinical importance. For example, large subgroup effects were anticipated and then

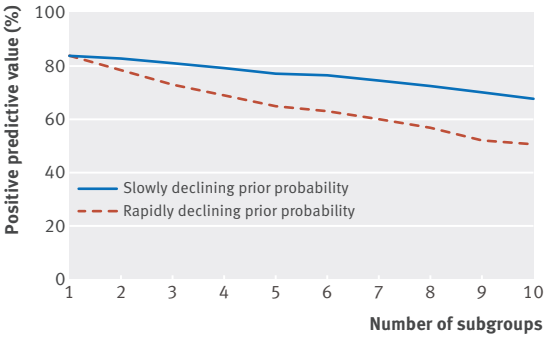

Fig 2 | **Effect of decreasing prior probability of each additional subgroup on positive predictive value (the likelihood that a positive finding is a "true positive") adjusting for multiple comparisons. The two lines represent how the positive predictive value decreases under different assumptions about how prior probability decreases. In the blue line (slowly declining prior probability), prior probability starts at 0.4 and falls per extra subgroup either linearly (0.04 per subgroup). In the dotted line (rapidly declining prior probability), prior probability falls inversely to the number of subgroups**

Table 3 | **Changes in power to identify subgroup effects after changes in specification of subgroup effect**

Change in subgroup effect and description	Power to detect subgroup effect (%)
Base case only (no change)	
Trial with 90% power for main effect; binary subgroup with half the trial population in each population; treatment effect, equal to the trial's powered effect, exists in one subgroup and no effect exists in the other subgroup	38
Base case with change in main trial power	
Decrease overall trial power to 80%	30
Increase overall trial power to 95%	45
Increase overall trial power to 99%	58
Base case with change in subgroup size	
25% of trial population in subgroup	27
10% of trial population in subgroup	16
5% of trial population in subgroup	11
Base care with change in subgroup effect size	
50% of base case effect size	12
150% of base case effect size	74
200% of base case effect size	95

Power was estimated for each of the outlined scenarios using the methodology of Brookes and colleagues.[19]

discovered for higher degrees of carotid stenosis in symptomatic endarterectomy trials,[15] and in time to reperfusion for stroke[16] and myocardial infarction.[17 18] Similarly, in biologically directed treatments, such as for cancer treatments, it is possible that a treatment will be effective in one subgroup and ineffective in another.

In addition to anticipated subgroup effect size, equalising the proportion of individuals in each subgroup and testing a continuous variable (such as baseline blood pressure level) along its continuum will substantially improve statistical power.[8 19] But even with excellent power, prior probability is an important determinant of whether a significant subgroup effect is a true positive. Just as an excellent diagnostic test (sensitivity 99%, specificity 95%) results in a 20% positive predictive value if the prior probability of disease is 1%, only 20% of statistically significant subgroup effects will be true positives if the subgroup effect only had a 1% chance of occurring to begin with, even with 99% power.

This demonstrates why looking at multiple variables might be fine for hypothesis generating analyses, but that primary (hypothesis testing) analyses must be based on previous empirical evidence or theory, generally be few in number, and be specified a priori. For unique cases with improved power to detect subgroup effects, after formally estimating subgroup effect power before the trial and considering the prior probability, table 2 can inform whether power is adequate to consider such an analysis as a hypothesis testing analysis versus a hypothesis generating analysis.

Conclusion

By not following a sound scientific process, conventional subgroup analyses increase the risk of both false positive and false negative findings. Careful consideration of the likelihood of a subgroup effect being present (prior probability) and the statistical power of the subgroup analysis (sensitivity) need to inform whether a subgroup analysis should be part of the primary (hypothesis testing) or exploratory (hypothesis generating) analyses. We recommend that hypothesis testing analyses include no more than one or two prespecified subgroup analyses founded on adequate prior probability and power so that positive findings are more reliable and can thus be used to target treatment.

This approach would substantially restrict the number of subgroup analyses performed. It can be argued that subgroup analyses that do not meet these criteria should never be performed because false positives will greatly outnumber true positives and could be integrated into clinical decisions in spite of the best intentions of researchers. However, there is also a reasonable argument supporting a limited role for exploratory hypothesis generating analyses of subgroups. For example, there was little reason to think that diabetics would fare better with coronary artery bypass than with percutaneous interventions before an exploratory subgroup analysis of the BARI trial.[20] Although still somewhat controversial,[21] the balance of evidence argues that this is a real subgroup effect that would not have been discovered without an exploratory analysis.[22 23] At least, if such analyses are performed, they should be specifically designated as exploratory, broadly understood to be hypothesis generating and reported separately from hypothesis testing analyses. For clinicians, these exploratory analyses should be ignored until confirmed or refuted by subsequent studies.[11]

Subgroup analyses have historically misinformed as much as they have informed.[3] The three simple rules outlined here can help guide more meaningful and accurate analyses and reporting of randomised controlled trials, better guide clinical implementation, and avoid repeating mistakes of the past.

doi: 10.1136/bmj.h5651 | *BMJ* 2015;351:h5651 | the bmj

We thank Tim Cole for a careful reading and useful comments on an earlier draft of this manuscript.

Contributors: The authors have experience designing and implementing research methods to individualise treatment decision making, including simulation analyses. RAH conceptualised the article and revised the manuscript. JBS performed the analyses and wrote the initial draft. JBS and DMK provided substantial intellectual feedback and revised the manuscript.

Competing interests: We have read and understood the BMJ Group policy on declaration of interests and declare no competing interests.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ* 2012;344:e1553.
2 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis) uses of baseline data in clinical trials. Lancet. *Elsevier* 2000;355:1064-9.
3 Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;365:176-86.
4 Tukey JW. Exploratory data analysis. 1st ed. Addison-Wesley, 1977.
5 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;116:78-84.
6 Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117.
7 Sun X, Ioannidis JPA, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis. *JAMA* 2014;311:405.
8 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.
9 Weiss CO, Varadhan R, Puhan MA, et al. Multimorbidity and evidence generation. *J Gen Intern Med* 2014;29:653-60.
10 Kahnemann D. Thinking, fast and slow. Farrar, Straus, and Giroux, 2013.
11 Kent DM, Rothwell PM, Ioannidis JPA, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010;11:85.
12 Sox HC. Probability theory in the use of diagnostic tests. An introduction to critical study of the literature. *Ann Intern Med* 1986;104:60-6.
13 Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2005;2:e124.
14 Chopra V, Hayward RA. Why P is not perfect. *Am J Med* 2012;125:e1-2.
15 Barnett H, Taylor D, Eliasziw M, et al. Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. *N Engl J Med* 1998;339:1415-25.
16 Lees KR, Bluhmki E, Kummer von R, et al. Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. *Lancet* 2010;375:1695-703.
17 De Luca G, Suryapranata H, Ottervanger JP, Antman EM. Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: every minute of delay counts. *Circulation* 2004;109:1223-5.
18 Cannon CP, Gibson CM, Lambrew CT, et al. Relationship of symptom-onset-to-balloon time and door-to-balloon time with mortality in patients undergoing angioplasty for acute myocardial infarction. *JAMA* 2000;283:2941-7.
19 Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess* 2001;5:1-56.
20 Bypass Angioplasty Revascularization Investigation (BARI). Comparison of coronary bypass surgery with angioplasty in patients with multivessel disease. Bypass Angioplasty Revascularization Investigation (BARI) Investigators. *N Engl J Med* 1996;335:217-25.
21 Daemen J, Boersma E, Flather M, et al. Long-term safety and efficacy of percutaneous coronary intervention with stenting and coronary artery bypass surgery for multivessel coronary artery disease: a meta-analysis with 5-year patient-level data from the ARTS, ERACI-II, MASS-II, and SoS trials. *Circulation* 2008;118:1146-54.
22 Hlatky MA, Boothroyd DB, Bravata DM, et al. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. *Lancet* 2009;373:1190-7.
23 Kapur A, Hall RJ, Malik IS, et al. Randomized comparison of percutaneous coronary intervention with coronary artery bypass grafting in diabetic patients: 1-year results of the CARDia (Coronary Artery Revascularization in Diabetes) trial. *J Am Coll Cardiol* 2010;55:432-40.