Treating Individuals 2

Subgroup analysis in randomised controlled trials: importance, indications, and interpretation

Lancet 2005; 365: 176–86

Peter M Rothwell

Stroke Prevention Research Unit, University Department of Clinical Neurology, Radcliffe Infirmary, Oxford OX2 6HE, UK (P M Rothwell FRCP)

peter.rothwell@cneuro.ox.ac.uk

Large pragmatic trials provide the most reliable data about the effects of treatments, but should be designed, analysed, and reported to enable the most effective use of treatments in routine practice. Subgroup analyses are important if there are potentially large differences between groups in the risk of a poor outcome with or without treatment, if there is potential heterogeneity of treatment effect in relation to pathophysiology, if there are practical questions about when to treat, or if there are doubts about benefit in specific groups, such as elderly people, which are leading to potentially inappropriate undertreatment. Analyses must be predefined, carefully justified, and limited to a few clinically important questions, and post-hoc observations should be treated with scepticism irrespective of their statistical significance. If important subgroup effects are anticipated, trials should either be powered to detect them reliably or pooled analyses of several trials should be undertaken. Formal rules for the planning, analysis, and reporting of subgroup analyses are proposed.

Introduction

> "The essence of tragedy has been described as the destructive collision of two sets of protagonists, both of whom are correct. The statisticians are right in denouncing subgroups that are formed post hoc from exercises in pure data dredging. The clinicians are also right, however, in insisting that a subgroup is respectable and worthwhile when established a priori from pathophysiological principles."
>
> A R Feinstein, 1998[1]

Randomised controlled trials (RCTs) and systematic reviews are the most reliable methods of determining the effects of treatments.[2–5] However, when trials were first developed for use in agriculture, researchers were presumably concerned about the effect of interventions on the overall size and quality of the crop rather than on the wellbeing of any individual plant. Clinicians have to make decisions about individuals, and how best to use results of RCTs and systematic reviews to do this has generated considerable debate.[6–22] Unfortunately, this debate has polarised, with statisticians and predominantly non-clinical (or non-practising) epidemiologists warning of the dangers of subgroup analysis and other attempts to target treatment, and clinicians warning of the dangers of applying the overall results of large trials to individual patients without consideration of pathophysiology or other determinants of individual response. This rift, described by Feinstein as a "clinicostatistical tragedy",[1] has been widened by some of the more enthusiastic proclamations on the extent to which the overall results of trials can properly inform decisions at the bedside or in the clinic.[23–25]

The results of small explanatory trials with well-defined eligibility criteria should be easy to apply, but generalisability is often undermined by highly selective recruitment, resulting in trial populations that are unrepresentative even of the few patients in routine practice who fit the eligibility criteria.[26] Recruitment of a higher proportion of eligible patients is a major strength of large pragmatic trials, but deliberately broad and sometimes ill-defined entry criteria mean that the overall result can be difficult to apply to particular groups,[27] and that subgroup analyses are necessary if heterogeneity of treatment effect is likely to occur. Yet despite the adverse effects on patient care that can result from misinterpreted or inappropriate subgroup analyses (table 1), there are no reviews or guidelines on the clinical indications for subgroup analysis and no consensus on the implications for trial design, analysis, and interpretation of subgroup effects, and the CONSORT statement on reporting of trials includes only a few lines on subgroup analysis.[28] This article discusses arguments for and against subgroup analyses, the clinical situations in which they can be useful, and rules for their performance and interpretation. Illustrative examples are taken mainly from treatments

Observation	Refutation
Aspirin is ineffective in secondary prevention of stroke in women[29,30]	31
Antihypertensive treatment for primary prevention is ineffective in women[32,33]	34
Antihypertensive treatment is ineffective or harmful in elderly people[35]	36
Angiotensin-converting enzyme inhibitors do not reduce mortality and hospital admission in patients with heart failure who are also taking aspirin[37]	38
β blockers are ineffective after acute myocardial infarction in elderly people,[39] and in patients with inferior myocardial infarction[41]	40
Thrombolysis is ineffective >6 hours after acute myocardial infarction[42]	43
Thrombolysis for acute myocardial infarction is ineffective or harmful in patients with a previous myocardial infarction[42]	44
Tamoxifen citrate is ineffective in women with breast cancer aged <50 years[45]	46
Benefit from carotid endarterectomy for symptomatic stenosis is reduced in patients taking only low-dose aspirin due to an increased operative risk[47]	48
Amlodipine reduces mortality in patients with chronic heart failure due to non-ischaemic cardiomyopathy but not in patients with ischaemic cardiomyopathy[49]	50

Table 1: **Examples of subgroup analyses that have shown apparently clinically important heterogeneity of treatment effect which has subsequently been shown to be false**

for cerebrovascular or cardiovascular disease but the principles are relevant to all areas of medicine and surgery.

Arguments against subgroup analysis

> " . . . it would be unfortunate if desire for the perfect (ie, knowledge of exactly who will benefit from treatment) were to become the enemy of the possible (ie, knowledge of the direction and approximate size of the effects of treatment of wide categories of patient)."
>
> S Yusuf et al, 1984[4]

The main argument against subgroup analysis is that qualitative heterogeneity of relative treatment effect (defined as the treatment effect being in different directions in different groups of patients, ie, benefit in one subgroup and harm in another) is very rare.[2–5] However, this observation is much less reassuring than it seems. First, it automatically excludes most treatments because they do not have a substantial risk of harm and can only be effective or ineffective. Yet use of an ineffective treatment can be highly detrimental if this prevents the use of a more effective alternative or if adverse effects impair quality of life. Second, the

Panel 1: **Rules of subgroup analysis: a proposed guideline for design, analysis, interpretation, and reporting**

Trial design

- Subgroups analyses should be defined before starting the trial and should be limited to a small number of clinically important questions.
- Expert clinical input into the design of subgroup analyses is needed to ensure that all relevant baseline clinical and other data are recorded.
- The direction and magnitude of anticipated subgroup effects should be stated at the outset.
- The exact definitions and categories of the subgroup variables should be defined explicitly at the outset in order to avoid post hoc data-dependent variable or category definitions. For continuous or hierarchical variables the cut-off points for analysis should be predefined.
- Stratification of randomisation by important subgroup variables should be considered.
- If important subgroup-treatment effect interactions are anticipated, trials should ideally be powered to detect them reliably.
- Trial stopping rules should take into account anticipated subgroup-treatment effect interactions and not simply the overall effect of treatment.
- If relative treatment effect is likely to be related to baseline risk, the analysis plan should include a stratification of the results by predicted risk. The risk score or model should be selected in advance so that the relevant baseline data can be recorded.

Analysis and reporting

- The above design issues should be reported in the methods section along with details of how and why subgroups were selected.
- Significance of the effect of treatment in individual subgroups should not be reported; rates of false negative and false positive results are extremely high. The only reliable statistical approach is to test for a subgroup-treatment effect interaction.
- All subgroup analyses that were done should be reported—ie, not only the number of subgroup variables but also the number of different outcomes analysed by subgroup, different lengths of follow-up etc.
- Significance of pre hoc subgroup-treatment effect interactions should be adjusted when multiple subgroup analyses are done.
- Subgroup analyses should be reported as absolute risk reductions and relative risk reductions. Where relevant the statistical significance of differences in absolute risk reductions should be tested.
- Ideally, only one outcome should be studied and this should usually be the primary trial outcome, irrespective of whether this is one outcome or a clinically important composite outcome.
- Comparability of treatment groups for prognostic factors should be checked within subgroups.
- If multiple subgroup-treatment effect interactions are identified, further analysis is needed to check whether their effects are independent.

Interpretation

- Reports of the significance of the effect of treatment in individual subgroups should be ignored, especially reports of lack of benefit in a particular subgroup in a trial in which there is overall benefit, unless there is a significant subgroup treatment effect interaction
- Genuine unanticipated subgroup-treatment effect interactions are rare (assuming that expert clinical opinion was sought in order to pre-define potentially important subgroups) and so apparent interactions that are discovered post hoc should be interpreted with caution.
  No test of significance is reliable in this situation.
- Pre hoc subgroup analyses are not intrinsically valid and should still be interpreted with caution. The false positive rate for tests of subgroup-treatment effect interaction when no true interaction exists is 5% per subgroup.
- The best test of validity of subgroup-treatment effect interactions is their reproducibility in other trials.
- Few trials are powered to detect subgroup effects and so the false negative rate for tests of subgroup-treatment effect interaction when a true interaction exists will usually be high.

observation refers only to so-called unanticipated heterogeneity.[2–5] As outlined below, there are many examples in which qualitative heterogeneity of relative treatment effect has been correctly anticipated. Third, the observation only applies to single outcome events; it is argued that subgroup analyses based on composite outcomes are inappropriate.[2–5,51] However, since qualitative heterogeneity of relative treatment effect is only possible for treatments that have a risk of harm, and such treatments almost always need a composite outcome to express the balance of both risk and benefit, qualitative heterogeneity as defined will inevitably be rare—a Catch-22, in fact.

There are several other arguments against attempts to target treatment. First, it is said that clinicians already tend to undertreat patients,[52] and we should not risk effective treatments being further restricted. However, one of the main purposes of subgroup analysis is to extend the use of treatments to subgroups that are not currently treated in routine practice. Subgroup analyses in epidemiological studies and trials often show that benefit from treatment is likely to be more universal than expected and that current indications for treatments in routine clinical practice are inappropriately narrow, as is now clear, for example, with treatment thresholds for blood pressure lowering or lipid lowering.[53,54] Second, it is argued that subgroup analyses are almost always underpowered,[55–60] but this is simply an argument for larger trials and for meta-analysis of individual patient data. Third, it has also been argued that false positive subgroup effects might be more common than genuine heterogeneity,[2–5,55–60] and these false observations might harm patients—"subgroups kill people."[61] Subgroup analyses have certainly led to mistaken clinical recommendations (table 1), but these analyses would not have satisfied the rules suggested in panel 1. Moreover, not doing subgroup analysis can also be harmful. Properly powered subgroup analyses most commonly show that relative treatment effect is consistent across subgroups and, or, that treatments should be used more extensively than is currently the case.[53,62,63] Without such evidence, unfounded clinical concerns about possible heterogeneity or inappropriately narrow indications for treatment would reduce the use of effective treatments in routine practice.[26] Not doing subgroup analyses has very probably killed more people.

***Panel 2:* The four main clinical indications for subgroup analysis**

Potential heterogeneity of treatment effect related to risk

- Differences in risks of treatment
- Differences in risk without treatment

Potential heterogeneity of treatment effect related to pathophysiology

- Multiple pathologies underlying a clinical syndrome
- Differences in the biological response to a single pathology
- Genetic variation

Clinically important questions related to the practical application of treatment

- Does benefit differ with severity of disease?
- Does benefit differ with stage in the natural history of disease?
- Is benefit related to the timing of treatment after a clinical event?
- Is benefit dependent on comorbidity?

Underuse of treatment in routine clinical practice due to uncertainty about benefit

- Underuse of treatment in specific groups of patients eg, elderly people
- Confinement of treatment according a narrow range of values of a relevant physiological variable—eg, treatment thresholds for cholesterol level or blood pressure

Situations in which subgroup analyses should be considered

> "The tragedy of excluding cogent pathophysiologic subgroup analyses merely because they happen to be subgroups will occur if statisticians do not know the distinction, and if clinicians who do know it remain mute, inarticulate or intimidated."
>
> A R Feinstein, 1998[1]

Subgroup analyses should be predefined and carefully justified. Feinstein and others have emphasised the need for determination of pathophysiological heterogeneity, but there are three other indications for subgroup analysis (panel 2), each of which are discussed below, which are probably more important.

Heterogeneity related to risk

Clinically important heterogeneity of treatment effect is common when different groups of patients have very different absolute risks with or without treatment. The need for reliable data about risks and benefits in subgroups and individuals is greatest for potentially harmful interventions, such as warfarin or carotid endarterectomy, which are of overall benefit but that kill or disable a proportion of patients. However, evidence-based guidelines usually recommend these treatments in all cases similar to those in the relevant RCTs.[64–66] In considering this approach, it is useful to draw an analogy with the criminal justice system. Suppose that research showed that individuals charged by the police with specific crimes were usually guilty. Few would argue that they should therefore be sentenced without trial. Automatic sentencing would, on average, do more good than harm, with most criminals correctly convicted, but any avoidable miscarriages of justice are widely regarded as unacceptable. In contrast, relatively high rates of

treatment-related death or disability (miscarriages of treatment) are tolerated by the medical scientific community precisely because, on average, treatment will do more good than harm. In both situations systems need to be in place to avoid doing harm. Yet the contrast between the effort that is put into the defence of the accused in order to avoid wrongful conviction and the very limited efforts of the medical scientific community to identify patients at high risk of harm is obvious. Admittedly, determination of guilt in a criminal trial is based on knowledge of past events, which can often be established with certainty, whereas probable benefit or harm from medical treatment depends on future events, which are usually less certain. However, the probable balance of risk and benefit in individual patients can be predicted to some extent with subgroup analysis and risk models, as has been shown, for example, with carotid endarterectomy.[67–70] In view of the fact that treatment complications are now a leading cause of death in developed countries,[71] effort is needed to more effectively target potentially harmful interventions.

Differences in the risk of a poor outcome without treatment can also lead to clinically important heterogeneity of treatment effect. Trial populations are often skewed in terms of control group risk, with a few individuals contributing much of the observed risk,[72] and treatment may be ineffective or harmful in the low risk majority. In vascular medicine, this is the case with endarterectomy for symptomatic carotid stenosis,[69] anticoagulation for uncomplicated non-valvular atrial fibrillation,[73] coronary artery bypass grafting,[74] and anti-arrhythmic drugs after myocardial infarction.[75] Clinically important heterogeneity of relative treatment effect by baseline risk has also been shown for blood pressure lowering,[76] aspirin,[77] and lipid lowering[78] in primary prevention of vascular disease, and in treatment of acute coronary syndromes with clopidogrel,[79] and with enoxaparin versus unfractionated heparin.[80,81] There are many similar examples in other areas of medicine,[82,83] and this issue is the subject of the next article in this series.

Pathophysiological heterogeneity

Differences between groups of patients in underlying pathology, biology, or genetics can each lead to clinically important heterogeneity of treatment effects. Examples will probably be identified more frequently as our understanding of the molecular mechanisms of disease is enhanced.

Multiple underlying pathologies

Clinicians often have to treat patients with ill-defined clinical syndromes, which probably have many underlying pathologies, rather than one disease. Primary generalised epilepsy is a typical example in which treatment effects differ between patients, probably because of the different underlying molecular pathologies. In vascular disease, clinically important heterogeneity of treatment effect in relation to underlying pathology is seen with thrombolysis for acute ischaemic stroke,[84,85] with aspirin in primary prevention of vascular disease (in which benefit may be largely confined to men with elevated levels of C-reactive protein,[86] probably indicating underlying atherosclerosis), and with blood pressure-lowering in secondary prevention of transient ischaemic attack and stroke, in which guidelines suggest that all patients be treated.[87–89] However, there is clinical concern about patients with carotid stenosis or occlusion in whom cerebral perfusion is often severely impaired.[90,91] Table 2 shows stroke risk by systolic blood pressure in patients with and without flow-limiting (≥70%) carotid stenosis who were randomly assigned to medical treatment in RCTs of endarterectomy.[92] Major increases in stroke risk were noted in patients with flow-limiting stenosis, but only if systolic blood pressure <150 mm Hg: 5-year risk in patients with bilateral (≥70%) stenosis was 64·3% versus 24·2% (p=0·002) at higher blood pressures. This difference in risk was absent in patients who had been randomly assigned to endarterectomy (13·4% *vs* 18·3%, p=0·6), suggesting a causal effect and indicating that aggressive blood pressure-lowering would very probably be harmful in patients with bilateral severe carotid disease in whom endarterectomy was not possible.

	Systolic blood pressure (mm Hg)			
	<130	130–149	150–169	>170
Stenosis group				
Bilateral <70%	1	1	1	1
Unilateral ≥70%	1·90 (1·24–2·89) p=0·02	1·18 (0·92–1·51) p=0·30	1·27 (0·99–1·64) p=0·13	1·64 (1·15–2·33) p=0·03
Bilateral ≥70%	5·97 (2·43–14·68) p<0·001	2·54 (1·47–4·39) p=0·001	0·97 (0·4–2·35) p=0·95	1·13 (0·50–2·54) p=0·77

The hazard ratios are derived from a Cox proportional hazards model stratified by trial and adjusted for age, sex and previous coronary heart disease. Patients with bilateral <70% stenosis are allocated a hazard of 1. ≥70% stenosis is only consistently associated with an increase in the risk of stroke at lower levels of systolic blood pressure.

Table 2: **Hazard ratios (95% CI) for risk of stroke in patients categorised according to severity of carotid disease within pre-defined blood pressure groups[92]**

Biological heterogeneity

Subgroup analyses can also be useful when there are predictable differences in the biological response to the underlying disease. For example, perioperative administration of antilymphocyte antibodies reduces rejection in cadaveric renal transplantation by 30%,[93,94] but is expensive and has serious adverse effects. Clinical concern that benefit might depend on pre-existing immune sensitisation prompted a meta-analysis of individual patient data from five RCTs. As predicted, treatment was highly effective in sensitised patients (hazard ratio for allograft failure at 5 years=0·20, 95% CI=0·09–0·47) but was ineffective in the remaining 85% (0·97, 0·71–1·32).[94] The subgroup-treatment effect interaction was significant (p=0·009)—ie, the effect of treatment was significantly different between the subgroups. A similar pre-specified immunological subgroup analysis in a large trial of

roxithromycin versus placebo after coronary angioplasty showed that treatment reduced restenosis and the need for revascularisation if the titre of *Chlamydia pneumoniae* antibody was high but was ineffective or harmful if the titre was low (interaction p=0·006).[95]

Genetic heterogeneity

Individuals respond differently to some drugs and this tendency can be inherited.[96,97] Genotype is an important determinant of both the response to treatment and the susceptibility to adverse reactions for a wide range of drugs.[98,99] For example, response to chemotherapy is dependent on gene expression in both colon cancer[100] and breast cancer,[101] and HDL cholesterol response to oestrogen replacement therapy is highly dependent on sequence variants in the gene encoding oestrogen receptor α.[102] In each of these cases, significant subgroup-treatment effect interactions have been reported. There is also great interest in the effects of genetics on the response to treatment in patients with HIV-1.[103] Subgroup analyses based on genotype have particular methodological problems since many genotypes may be studied and analyses will often be post hoc.

Heterogeneity related to practical application

Many of the arguments used against subgroup analyses misinterpret their main function. The main potential of subgroup analysis is not in the identification of groups that differ in their response to treatment for reasons of pathophysiology, but is in answering practical questions about how treatments should be used most effectively, such as at what stage of the disease is treatment most effective, how soon after a clinical event is treatment sufficiently safe or most effective, or how are the risks and benefits related to comorbidity? Subgroup analyses related to questions of the practical application of interventions can be vital to effective clinical practice.

Severity or stage of disease

Treatment effects often depend on severity of disease. In primary prevention of vascular disease, a pooled analysis of RCTs of pravastatin showed that the relative risk reduction with treatment increased with baseline LDL cholesterol (interaction p=0·01): relative risk reduction=3% in the lowest quintile and 29% in the two highest quintiles.[104] In stroke medicine, carotid endarterectomy is highly effective for ⩾70% recently symptomatic stenosis, modestly effective for 50–69% stenosis, but harmful for <50% stenosis (interaction p<0·0001).[105] In cardiology, thrombolysis for acute myocardial infarction is ineffective or harmful in patients with ST segment depression, but highly beneficial in patients with ST elevation (interaction p<0·01),[106] and early invasive treatment of unstable angina is of no benefit in patients with only minor ST segment change but of major

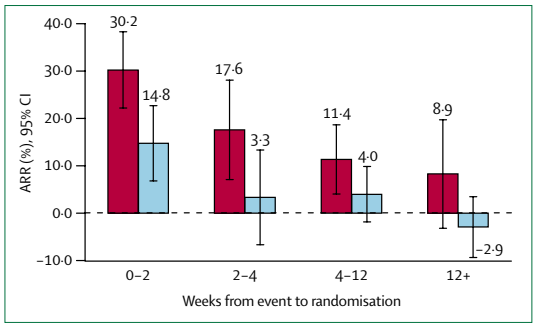

Figure 1: **Effect of carotid endarterectomy in patients with 50–69% and ⩾70% symptomatic stenosis in relation to time from last symptomatic ischaemic event to randomisation[70]**

Numbers above bars indicate actual absolute risk reduction. Vertical bars are 95% CIs. ARR=absolute risk reduction.

benefit in patients with more marked changes (interaction p=0·006).[107] The stage of disease can also determine the effect of treatment of non-vascular disease, as is seen in people with cancer,[108,109] or HIV/AIDS.[110–112]

Timing of treatment and comorbidity

Effect of treatment is often critically dependent on timing, as shown in figure 1, for benefit from endarterectomy for recently symptomatic carotid stenosis. The risk of a stroke is very high during the first few days and weeks after a transient ischaemic attack,[113] especially in patients with carotid stenosis,[114] but falls rapidly with time, as therefore does benefit from endarterectomy.[70] Similar time-dependence has been shown for benefit from thrombolysis for both acute myocardial infarction[106] and acute ischaemic stroke.[115]

Treatment effects may also depend on comorbidity. For example, angiotensin-converting enzyme inhibitors and angiotensin II receptor blocking drugs are harmful in patients with renovascular disease but highly beneficial in other hypertensive patients.[116] Benefit from diltiazem after myocardial infarction may depend on the presence of heart failure because of the negative chronotropic and inotropic effects of the drug.[117]

Underuse of treatment in specific groups

Treatments that are effective in trials are often underused in specific groups of patients in routine practice. For example, statins were not used in elderly people for many years until the drugs were proved highly effective by subgroup analysis in the Heart Protection Study.[53] Proof of some benefit by subgroup analysis was also needed to counter underuse in elderly patients of thrombolysis for acute myocardial infarction in elderly people,[106] and similar underuse of endarterectomy for symptomatic carotid stenosis.[70] In each case, treatment had already been shown to be highly effective overall. Use of treatment in routine clinical practice is also often inappropriately limited to patients with measurements of

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.