JOHNS HOPKINS
UNIVERSITY

Center for Clinical Trials

Department of Biostatistics
Department of Epidemiology
Department of International Health

Department of Medicine
Department of Ophthalmology
Oncology Center

(Sunday) 11 March 2001

Memorandum

To: Center for Clinical Trials faculty and staff

Fr: Curt Meinert

Re: On subgroup analysis in trials

Definitions

ad hoc subgroup *n* - A **subgroup** (defn 2) identified by **data analysis**. ant: **specified subgroup**

baseline subgroup *n* - [**trials**] A collection of **treatment units** (usually persons) having a particular **baseline** feature, characteristic, or measure. rt: **baseline**, **subgroup**

baseline subgrouping variable *n* - [**trials**] A **subgrouping variable** based on a **baseline** measure or observation. rt: **subgrouping variable**

data dredging *v* - **Data analyses** done on an **ad hoc** basis, without benefit of prior stated **hypotheses**, especially those done with the aim or intent of trying to find noteworthy differences within or among different **subgroups**; **exploratory data analysis**; see **dredge**. ***Usage note***: Often used in a pejorative sense, especially in reference to analyses in which it appears that only large differences are presented and where the number of **comparisons** made is not specified. Not to be confused with **subgroup analysis**; see usage note for that term.

heterogeneous treatment effect *n* - [**trials**] A **treatment effect** that is different across **subgroups** represented in the trial; a treatment effect that is not **homogeneous**. rt: **treatment interaction**

homogeneous treatment effect *n* - [**trials**] A **treatment effect** that is the same across **subgroups** represented in a trial; a treatment effect that is not **heterogeneous**. rt: **treatment interaction**

specified subgroup *n* - A **subgroup** (defn 2) specified prior to the start of **data analysis**. syn: prespecified subgroup ant: **ad hoc subgroup** rt: **designed subgroup comparison**

split-half reliability *n* - The extent to which **results** obtained from a defined part of a **dataset** (eg, the first half of the dataset as divided on the basis of time) correspond to those obtained from the remaining part. rt: **reliability**

subgroup *n* - 1. [general] A subordinate **group** whose members share some differentiating or distinguishing trait or feature. 2. [**research**] A subpart or subset of a **study population** distinguished by a particular characteristic or set of characteristics (eg, males under age 45 at entry). rt: **cluster**

subgroup analysis *n* **-** 1. Any **data analysis** focused on a selected **subgroup** (defn 2). 2. Analysis aimed at characterizing **observed differences** among different subgroups, eg, **comparison** of **treatment differences** in a **trial** for different subgroups of **patients** defined by sex, age at entry, and other **baseline characteristics**. 3. A form of **exploratory data analysis** aimed at trying to identify a subgroup of persons that account for an observed difference, eg, such an analysis in a trial to determine whether or not an observed **treatment difference** can be accounted for by some **subgroup**. See also **data dredging**. ***Usage note***: Not to be used interchangeably with **data dredging**. Data dredging is **value-laden** and pejorative. **Subgroup analysis** is neutral in connotation and is descriptive of a process. Analysis involving subgroups formed using entry **demographic** and other **baseline characteristics** is an essential part of the analysis process for a trial. The analyses are done to determine whether or not it is reasonable to regard the **treatment effect** observed as being **homogeneous** (ie, **independent** of entry and other important baseline characteristics). The analysis has bearing on conclusions reached from the trial. Evidence of **qualitative** or **quantitative** treatment by baseline characteristic **interaction** obligates the trialist to temper or qualify the conclusion accordingly. A treatment effect cannot be assumed to be homogeneous across subgroups absent analyses aimed at addressing the question of homogeneity of treatment effect. Subgroup analyses become forms of data dredging if results of such analyses are used to identify "**significant**" differences without regard to the number of subgroups studied or when the results are presented so as to suggest that the difference is the result of clinical insight regarding an underlying **disease** process.

subgroup difference *n* **-** [**trials**] 1. A **difference** in the **treatment assignment ratio** across **subgroups** defined by different levels of a particular **subgrouping variables**; especially such a difference considered to be **statistically significant**. 2. **subgroup treatment difference** rt: **subgroup**

subgroup treatment difference *n* **-** [**trials**] A **difference** in **treatment effect** as measured across **subgroups** for different levels of a particular **subgrouping variable**; especially such a difference considered to be **statistically significant**; a treatment by subgroup **interaction**. rt: **treatment interaction**

Introduction

Broadly, a subgroup, in the context of trials, is a subset of persons defined by some characteristic (eg, gender) or baseline measure (eg, cholesterol level). The purpose of the subgrouping is see whether the treatment effect differs by subgroup. If it does, the subgrouping variable is said to "explain" the treatment effect.

The subgrouping variable must be independent of treatment assignment. Independence, is achieved by limiting the choice of subgrouping variables to those that are temporally invariant, eg, one's DNA, or to those observed at or prior to randomization. Subgrouping variables that have the potential of being influenced by treatment (largely any variable observed after the moment of randomization) are not suitable for subgroup analyses.

The realities regarding subgroup analyses are:

1. Trials are not powered to detect subgroup effects
2. The majority of subgroup differences that are reported are not reproducible; for evidence see

Yusuf S, Wittes J, Probstfield J, Tyroler HA; Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials; JAMA 266:93-98, 1991.

3. Most subgroup differences that are reported relate to disease state or medical history; few relate to gender or ethnic origin
4. Most subgroup differences reported are the result of ad hoc analyses

On the purpose of subgroup analyses

Every trial, save for the pitifully small one, is amenable to subgroup analysis. However, the extent to which such analyses can be used to assess treatment effects depends on the size of the trial and on the diversity of the study population allowed with the selection and exclusion criteria used in the trial.

Largely, in trials, subgroup analyses are undertaken as a precautionary measure to guard against reporting an effect as being homogeneous when it is not. Therefore, one can argue that the trialist is obligated to look for subgroup differences before publishing results of the trial. Hence, the trialist undertakes subgroup analyses without any expectation of finding differences and tends to be skeptical of any differences that are found.

By and large, treatments that work, work across the subgroups represented in a trial. One reason is because of the homogenizing effect of inclusion and exclusion criteria. People for whom the treatment is not indicated or unlikely to work are excluded from enrollment. Another reason in regard to treatment trials is because disease tends to be the "equalizer". Even though the route to disease may differ depending on personal characteristics, once a person is diseased, the course of disease and treatment, as a rule, is more heavily influenced by the state of disease than by the personal demographic or baseline characteristics of the diseased persons.

On the difference between stratification and subgrouping

Stratification in trials is done at "randomization time" to ensure the comparability of the treatment groups with regard to the stratification variable. The comparability is achieved by ensuring that the treatment groups have the same mix of persons with regard to the stratification variable. For example, if one stratifies on gender each treatment group will have the same proportionate mix of males to females.

Subgrouping is done at "analysis time". It is done to determine whether the treatment effect is constant across the different subgroups. A stratification variable is also a subgrouping variable if, at analysis time, the variable is used for subgrouping – generally the case, but there is no requirement that it be so used.

Typically, trialists do not present results by subgroup in finished manuscripts unless there are differences to report. Exceptions may be for stratification subgroups, subgroups were differences are speculated to exist, and subgroups reported in other publications.

On the difference between specified and ad hoc subgroups

A specified subgroup is one defined in design documents of the trial, hence, identified prior to the start of the trial. An ad hoc subgroup is one identified during analysis. Generally, results for specified subgroups are presented in study publications, whereas results for ad hoc subgroups are not presented unless the subgrouping reveals differences considered to be large and important.

On ad hoc subgroup analysis vs data dredging

Ad hoc subgroup analysis and data dredging involve a common activity – casting about in the finished data of a trial. The difference is in motivation. The data dredger sets about the activity in the hope of finding a "significant" difference and is content to regard differences as "real" if they are statistically significant by any measure. The subgroup analyst sets about the activity expecting to find nothing, is skeptical of differences that are found, and prone to disbelief even if the difference is "significant" by conventional measures.

The data dredger is interested in advancing some end. The end may be to elevate the importance of a dataset by being able to report a difference in an otherwise ho hum set of results or it may be to debunk a set of results (eg, as with Kilo C, Miller JP, Williamson JR in regard to results from the UGDP in an article entitled *The crux of the UGDP. Spurious results and biologically inappropriate data analysis*; Diabetologia 18:179-85, 1980).

General rules for subgroup analysis

1. Limit the choice of subgrouping variables to those that are operationally independent of treatment assignment and course of treatment; typically achieved by limiting to variable measured and observed prior to randomization
2. Count all events regardless of course of treatment or length of followup and regardless of time of occurrence after randomization
3. Perform the analysis within subgroups by counting persons to the treatment assigned; count to that group regardless of course of treatment or length of followup
4. Present results for all of the subgroups defined by the subgrouping variable, eg, if the variable is age and used to define 3 subgroups (< 35, 36 - 55, and > 55), present results for all 3 age subgroups
5. For continuous variables, choose the cutpoints to define subgroups independently of subgroup treatment differences (generally best achieved by specifying rules for determining cutpoints based on the distribution for all treatment groups combined)
6. Exercise skepticism in use of the subgrouping variable as an explanatory variable for treatment differences

Desirable conditions of subgroups

To be useful in explaining differences, the subgrouping variable and member subgroups should satisfy conditions 1 - 5 below and ideally conditions 6 - 8 as well:

Condition 1: Chosen independently of treatment assignment

Comment

This condition is ensured if the variable was observed and recorded at or prior to randomization. It is open to question if the variable is observed at any point after randomization, even if only moments after randomization.

In the case of classifications based on readings of records: The condition can be satisfied even if the classification is made after randomization if the records were obtained at or prior to randomization and if the classification is made by persons masked to treatment assignment.

Condition 2: Chosen independently of subgroup treatment differences

Comment

The operational requirements for this condition are satisfied when subgroupings are made without regard to observed treatment differences, eg, in the case of a continuous variable, by choosing cut points for subgroups based on the distribution of that variable for all treatment groups combined.

Condition 3: Unconfounded
Comment

It has to possible to estimate treatment by subgroup. The ability to do that is lost if the treatment groups are confounded across subgroups, ie, a chance occurrence where all or mostly all persons in subgroup 1 were assigned to treatment A and all or mostly all persons in subgroup 2 were assigned to treatment B.

Condition 4: Adequate size
Comment

There is also no point to subgrouping if virtually all persons fall into one subgroups. The estimates of treatment effect in the other subgroup(s) will not be reliable.

Condition 5: Reliable, ie, capable of being reproduced by others using the same dataset
Comment

The condition requires that the persons doing the subgrouping provide adequate detail and documentation to enable others to reproduce the result presented.

Condition 6: Biological or medical plausibility
Comment

This condition, while not essential, is highly desirable. Subgroup treatment differences for a variable having no obvious biological or medical plausibility are not likely to be taken seriously, regardless of size.

Condition 7: Internal consistency
Comment

The condition of internal consistency is related to condition 6. Generally, subgroup differences satisfying condition 6 are internally consistent, ie, the results are consistent with patterns seen in other subgroup analyses. The lack of internal consistency does not mean that the subgroup difference is spurious, but the lack of consistency should give the subgroup analyzer pause.

Condition 8: Independently reproducible in another study
Comment

This condition, if satisfied at all, is unlikely to be satisfied by the time the trial in question is published. Evidence as to the reproducibility of a subgroup difference is not likely until other studies are designed and completed.

When is a subgroup difference worthy of action during a trial?

The answer depends on the size and direction of the difference and where things stand in the trial when the difference is detected.

Prudence may call for action even if the difference is questionable. For example, that was the case with a subgroup treatment difference against dextrothyroxine vs placebo in the CDP in a small subgroup of men with a baseline ECG indicative of frequent ectopic ventricular beats (FEVB). Investigators decided to stop treatment for such persons and to change the eligibility criteria to exclude such persons from enrollment because of a higher mortality in the subgroup compared to men assigned to placebo (JAMA 220: 996-1,008; 1972).

As a rule, subgroup treatment differences do not appear out of the blue, assuming a regular monitoring process. Hence, there is opportunity, within limits, to assess "reproducibility" of the difference. One approach is to do "split half" analyses using calendar time. One would expect, if the effect is real, to see evidence of the effect in both halves of the dataset.

A variation on the approach is to "reset" the clock when a difference is noted and to consider events observed from that time forward as occurring in a "replication" of the trial. Again the expectation is for the difference to reproduce if the effect is real. (See Canner P: *Monitoring Clinical Trial Data for Evidence of Adverse or Beneficial Treatment Effects* for discussion of this technic and others in subgroup analyses done in the CDP when deciding whether to stop D-T4 treatment in a subgroup of men or in everyone; in Essais Controles Multicentres: Principes et Problemes INSERM 76:131-149, 1977).

The ultimate test of plausibility is to see whether the effect reproduces in an independent follow-on study. Obviously this approach has strengths, but rarely feasible. Perhaps the best example of attempts to replicate via a second trial comes from PARIS II (J Am Coll Cardiol 7:251-69, 1986). The trial was spawned by a subgroup difference noted in PARIS (Circulation 62:449-461, 1980) (the subgroup of patients enrolled within 6 months of their last MI appeared to benefit from treatment). PARIS II was limited to patients who had a recent MI (4 weeks to 4 month previous to enrollment). The follow-on study failed to show an effect.

On reporting subgroup treatment differences in manuscripts

Rule 1: Be cautious

Comment

The old adage, "look before you leap", most assuredly applies to reporting subgroup differences. Once something is published it is too late. The time for circumspection is before publication.

"Age" the subgroup analysis and intended conclusion. "Aging" is automatic if the subgroup is identified during treatment effects monitoring. Some form of enforced "aging" for reflection and circumspection is prudent if the subgroup is identified when the paper is being readied for publication.

Rule 2: Present results for all subgroups represented for the subgrouping variable

Comment

Operationally, this means for a variable such as gender that results are presented for both subgroups. For a variable yielding 3 or more subgroups, it means that results are presented for each of the subgroups.

Rule 3: Indicate the means of identification; if specified in design documents, reference the document containing the specification; if an ad hoc subgroup, indicate when in the course of the trial identified and how

Rule 4: Discuss the observed subgroup treatment difference in regard to presumed biological and medical relevance

Comment

Provide supporting evidence from other studies (if any); indicate degree of conviction that the effect is real.

Rule 5: Reflect proper circumspection and skepticism in regard to conclusions based on subgroup differences

Comment

Keep in mind that the majority of subgroups treatment differences do not reproduce when submitted to replication (eg, as seen by Yusuf et al; JAMA 266:93-98, 1991).

Myths regarding subgroup analyses

Myth 1: That subgroup analyses, save for those specified prior to the start of the trial, should not be done

Comment

Wrong. It can be argued that the trialist has a duty to probe for heterogeneity of treatment effects. The pooled overall results by treatment group should not be presented as an estimate of treatment effect if there is evidence to indicate that the treatment effect is not homogeneous. The only way results can be probed for heterogeneity of effect is by subgroup analyses.

Myth 2: That only subgroups specified prior to the start of the trial should be analyzed

Comment

Wrong, for reasons stated above.

Myth 3: The conditions for pre-specification of subgroups are satisfied if the specification is made before data are analyzed

Comment

Wrong, strict adherence to the pre-specification notion requires specification in the study protocol before any data are collected.

Myth 4: That one is obliged to report results for the subgroups represented by stratification variables

Comment

Wrong. There is no such obligation nor is it practical when the number of strata is large. For example, there were 106 strata in the CDP (53 clinics and 2 risk groups per clinic). The only time there is an obligation is if analyses indicate that the treatment effect is different by strata.

Myth 5: That stratification variables are pre-specified subgrouping variables and, hence, that results must be presented by strata

Comment

No. Use of a variable for stratification does not imply the existence of treatment differences for the variable. See also Myth 4.

Myth 6: That one is obliged to report results for subgroup analyses preformed

Comment
Wrong. The only obligation is to report differences suggestive of a differential treatment effect.

Myth 7: That the absence of subgroup analyses in published papers is prima facia evidence of failure to have performed subgroup analyses
Comment
Wrong. One does not typically report subgroup analyses unless they are important.

Critics have taken the absence of subgroup analyses by gender as evidence of lack of interest in gender as an explanatory variable in trials. Generally, the more plausible explanation is that gender did not produce any difference worth publishing.

Myth 8: That investigators have a responsibility to report results by gender and ethnic origin
Comment
Wrong. The responsibility is to report things that matter. If the results do not differ by gender or ethic origin there is no obligation to report.

Myth 9: That the size of the p-values indicates likelihood of reproducibility of the difference
Comment
Wrong. Evidence of reproducibility is hard to come by, save for replication of the trial. P-values are, at best, only crude gauges of "statistical likelihood". In any case, p-values are sample size dependent. Hence, the larger the sample size the smaller the p-value, all other things equal.

Myth 10: That stratification variables must be used for subgroup analyses
Comment
Wrong. See Myth 4.

Myth 11: That a difference in the distribution of a baseline variable by treatment group is indicative of a breakdown in the randomization process
Comment
Randomization does not ensure baseline comparability. Hence, lacking other evidence, the most plausible explanation is that the investigators were just unlucky.

Myth 12: That an imbalance in distribution of a subgrouping variable can be "corrected" by a change of the randomization procedure to offset the imbalance
Comment
Wrong. A difference is a difference. One can, of course, alter the design to make the wayward variable a stratification variable but such alterations are not advised – not advised because changes in the design have to be explained when results are published. What would the explanation be?

In any case, even if a change is made, it is likely to be merely cosmetic in nature. A difference early in the course of enrollment, even if balanced by the end of enrollment, will still be evident for "event" variables because of differences in exposure to followup. The group with the larger number when the change was made will contribute more events simply because they are treated and followed longer than those coming later in the enrollment process.

Myth 13: That differences in the distribution of a baseline variables requires that one present treatment results by subgroups defined by the variable
Comment
Only if the treatment effect differs by subgroup. If it does not, the pooled effect is the best estimate of the treatment effect.

Myth 14: That a difference in the distribution of a baseline variable confounds treatment comparisons
Comment
Only if there is a treatment difference by different levels of the variable; generally not the case.

Myth 15: That ad hoc subgroup analysis is tantamount to data dredging
Comment
Wrong, as discussed above.

Distribution
via local e-mail
(See names8.wpd list)
CLM Musing file
Chronologic file