Subgroups, significance, and circumspection: investigating interactions in clinical trials

Stephen Senn

Department of Statistics, University of Glasgow, Glasgow, UK

Corresponding Author:
Stephen Senn, Department of Statistics
15 University Gardens, University of Glasgow, G12 8QQ, Glasgow, UK
E-mail: stephen@stats.gla.ac.uk

Summary
All statistical analysis of clinical trials involves pooling results from patients who will differ in terms of demographic characteristics or prognostic factors. It seems natural to want to try to establish the extent to which a treatment effect varies by subgroup. However, the number of potential subgroups can be large and the number of patients in a trial is often such that only pooling all of them will make it possible to reach a useful conclusion. Hence many problems arise when looking at subgroup effects, especially if the investigation is not guided by pre-specified approaches. In this paper, some pitfalls in examining subgroups are discussed and some possible approaches to investigating interaction illustrated using two different examples: the BHAT and the ATAC studies.

KEY WORDS: *multi-centre trials, subgroups, interactions*.

Introduction

As Artemus Ward said, “It ain’t so much the things we don’t know that get us in trouble. It’s the things we know ain’t so.” Statisticians are cynics motivated by the desire to debunk. Their catchphrase could be, “This looks plausible – how can I prove it wrong?” The study and practice of statistics teaches conservatism in inference and, of course, a faith in calculation, which the average (what else!?) statistician sees as necessary to correct the gullibility of mankind.

Most non-statisticians, on the other hand, underestimate the role of chance in what they observe. If you ask them the *birthday question* (“how many people must there be in a room before it is odds on that at least two share a birthday?”), they guess far higher than the correct answer of 23.

Depending on one’s philosophy of statistics, Bayesian or frequentist, there are two devices by which to implement statistical conservatism. The Bayesian injects prior beliefs into any calculation. These prior beliefs have to suggest that small differential effects are more plausible than large ones, otherwise the slightest change of circumstance would forbid the use of any previous experience to guide prediction. We would have to believe that a trial of treatment for breast cancer in Lutheran Sweden will tell us nothing about the treatment of sufferers in Catholic Belgium. Thus, when the data mean is averaged with the prior mean, as the Bayesian recipe dictates, a less extreme posterior mean results (1). The frequentist believes, on the other hand, that statistics is a matter of calling the shots. You can’t proceed to the snooker or pool table, try 20 pots and credibly claim that the one that came off was the one that you meant to come off. Similarly if you are looking for subgroup-by-treatment interactions in a clinical trial, and do not want to be misled by chance patterns, you had better pre-specify what you are looking for (2), have good reasons for doing so and take the rough with the smooth – if most of your subgroup investigations are negative perhaps you should be rather sceptical about those few that are not.

Yet, despite the warnings of statisticians, medical researchers continue to investigate subgroup-by-treatment interactions optimistically. A notorious example, to be discussed below, is the re-analysis ten years ago (3) of the beta-Blockers in Heart Attack Trial (BHAT) (4-7). Perhaps, researchers might reasonably argue that if one is not prepared to 'push the boat out' from time to time no new discoveries will be made. They are favouring William James's (8) "Believe truth!" over his "Shun error!". Nevertheless, if we take Artemus Ward's warning seriously, the case against "fishing trips" deserves a hearing. This case has, of course, often been made (9-12), but the excuse for making it again here is that it is regularly ignored. Indeed, for example, the inappropriate re-analysis of the BHAT (3) continues to be cited with approval (13) more often than the necessary corrections (14, 15). If anything, since, with the rise of pharmacogenomics, the number of ways we can classify patients is increasing, so too are the risks of over-interpretation and of irreproducible results (16, 17). Ioannidis recently came to the pessimistic conclusion that most research findings are false (18) and a survey of genetic association studies by Hirschhorn et al. found extremely poor reproducibility of results (19). Nor is understanding of these issues what one would wish. A recent survey of physicians in Ontario, Canada, found that among 435 respondents 44% would not prescribe a treatment that was effective on average to patients in a subgroup in which efficacy had not been demonstrated (20).

The purpose of this article is to make some points about approaches to analysing subgroups in clinical trials and in this regard two cases of rather different sorts are considered. The first is where the subgroups are formed by many clusters of patients, the labelling of the clusters being largely irrelevant and the individual clusters (from a global point of view) being of no particular interest in themselves. The multi-centre trial, in which the centre is the cluster, is the perfect paradigm of this case. The second is where there are some relevant groupings of patients based on prognostic classifications.

An example from the literature of each of these two cases will be discussed below in due course. But because it is hoped that this article may also be of use to the non-statistician, a brief explanation of some statistical terminology is given first.

Characteristics to be used in statistical models that can be divided into various categories are referred to as *factors* and the categories themselves are *levels* of the factors. Thus, in a breast cancer trial, oestrogen receptor status might be a factor and *positive* or *negative* could be the levels. The average extent to which a factor affects a relevant outcome is called a *main effect*. If this effect varies with a second factor then one speaks of the *interactive effect* or simply *interaction* of the two factors concerned. (Epidemiologists sometimes refer to this as *effect modification*). Thus, for example the average difference between anastrozole and tamoxifen in a breast cancer trial would be the main effect and if this difference, in turn, differed according to oestrogen receptor status then this would be an oestrogen receptor-by-treatment interaction. Finally the complexity of particular terms in a model is indexed by their *order*. Thus a main effect is of order one, an interaction of two factors is of order two, and a three-way interaction is of order three etc.

Let us now consider our two examples.

First example: a multi-centre heart disease trial

The BHAT was a randomised placebo-controlled study of the effect on mortality of propranolol in 3837 survivors of myocardial infarctions treated in 31 centres (6). Observation was planned for two to four years but the study was stopped nine months early with an average follow up of two years at which point total mortality under active treatment was 7.2% as opposed to 9.8% under placebo. (The fact that the trial was stopped early will be ignored here).

In 1996, Horwitz et al. presented a re-analysis of the BHAT in which they compared a group of 21 centres, which they labelled *dominant*, with a group of 10 centres labelled *divergent* as regards the treatment effect (3). They found that the effect of propranolol compared to placebo differed significantly between the divergent and the dominant centres. Propranolol was effective in the dominant centres and harmful in the divergent centres.

In fact, this finding was hardly a surprise (14, 15). The centres had been divided on the basis of the observed results: they had been labelled dominant if the propranolol mortality observed was lower than that under placebo and divergent otherwise. This is a clearly illegitimate way to form groups for a significance test. If this is not obvious, then one must consider that any test we use will be correlated with the Wilcoxon rank sum test; it may be more or less powerful than the rank test but it must often agree with it in judging significance. However, by grouping the centres in the way described above it is clear that we are putting all the highest ranked centres in one group and all the lowest ranked ones in the other. Hence, given that there were 10 divergent and 21 dominant centres it is hardly surprising that a significant difference emerged.

One might counter, of course, that there is nothing inherent in multi-centre trials that dictates that one *should* have divergent centres, or what, in fact, are sometimes referred to as *effect reversals* (21). However, although effect reversals are not necessary, they are in fact highly probable in any trial that includes a reasonable number of centres, even when the true treatment effect does not vary.

Suppose, for example, that a trial has been designed to have a power of 80% for a two-sided 5% type I error rate. This implies that the so-called *non-centrality parameter* δ, that is to say the ratio of treatment effect to standard error, will be about 2.8. However, if there are k centres of the same size, then the average number of patients per centre will be N/k where N is the total number in the trial and the standard error in a typical centre will be $\sqrt{k}$ times what it is in the trial as a whole. This means that the non-centrality parameter for the *centre* is now $\delta_c = \delta/\sqrt{k}$. For example, with $k = 16$, the non-centrality parameter drops to $2.8/\sqrt{16} = 0.7$. But about 24% of a standard normal distribution lies beyond 0.7 so that the probability of an effect reversal becomes 0.24. Now, of course, however many centres there are and hence, however small the typical centre, this probability of an effect reversal, where the treatment effect is genuine, cannot exceed 50%. However, in the example just mentioned above there are many centres and just as it is unwise to suppose that because the risk of death is small for a single trial of Russian roulette it is *therefore* small if played regularly, so it is unwise to suppose that the phenomenon of effect reversal is unlikely. It is *probable*. In the case of the 16 centres the probability that none will show an effect reversal is $(1 - 0.24)^{16} = 0.013$ and so the probability of at least one reversal is $1 - 0.013 = 0.987$ and, in fact, the expected number of effect reversals is $16 \times 0.24 \approx 4$.

As the number of centres increases then, other things being equal, the probability of at least one effect reversal increases. Indeed it does so as a result of two phenomena: not only does the probability of an effect reversal increase in any given centre but also the chances of escaping it reduce because there are more centres in which it may occur (21, 22). The position is illustrated in figure 1.

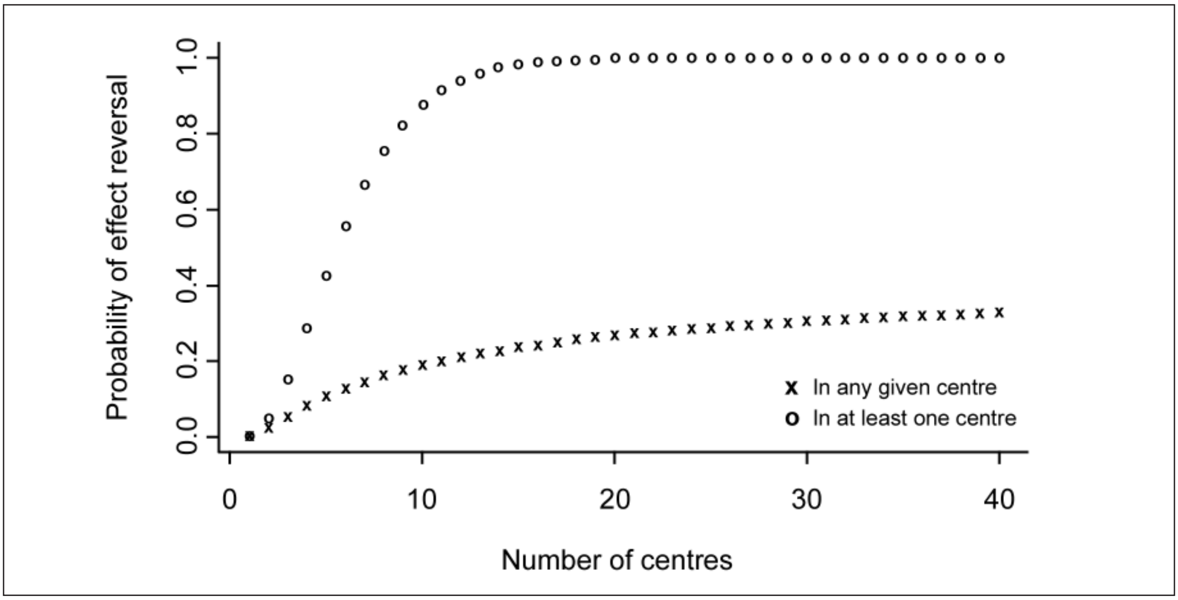

Figure 1. Probability of effect reversal in any given centre and in at least one centre as a function of the numbers of centres in a study with 80% power overall to detect a treatment effect for a two-sided 5% type I error rate.

In fact, an ensemble of centres, such as that provided by the BHAT, makes it possible to analyse the data using a so-called *random effects* model. Horwitz et al. (3) provided only total numbers (over both arms) by centre and the percentage death rate per arm to one decimal place. Table 1 shows an attempted reconstruction of the original data. The total numbers per centre agree with Horwitz et al. and the percentage death rates agree, to one decimal place, for both treatments and all centres apart from two cases: propranolol for centres 10 and 16, the percentage differing by 0.1 in the first case and 0.3 in the second. However, the data as thus reconstructed will be used in the analyses that follow.

There are various possible approaches for investigating interaction in such an ensemble. For example if a Mantel-Haenszel test is carried out using PROC FREQ® in SAS®, the Breslow-Day (23) chi-square for homogeneity of the common odds ratio will be 31.2 which, since there are 30 degrees of freedom, is almost exactly what is expected in the case of homogeneity. The odds ratio itself is estimated at 0.71 with 95% confidence limits of 0.57 and 0.90.

Another approach is to analyse the data using a model with a normal random effect for the centre and the treatment-by-centre interaction on the log-odds scale, treating treatment effect as fixed with cases having conditionally a binomial distribution. This can be implemented in SAS® using the GLIMMIX® macro. Again, the estimated odds ratio is 0.71 and the 95% confidence limits are 0.57 and 0.90. More importantly, however, the estimated

Table 1. Comparison of formal and narrative statements.

	Treatment					
		Placebo			Propranolol	
Centre	Patients	Survived	Died	Patients	Survived	Died
1	48	45	3	49	49	0
2	58	51	7	57	56	1
3	56	51	5	57	56	1
4	42	38	4	42	41	1
5	66	56	10	64	61	3
6	71	63	8	70	67	3
7	65	60	5	66	64	2
8	55	48	7	55	52	3
9	56	49	7	55	52	3
10	43	39	4	45	43	2
11	78	70	8	77	73	4
12	59	55	4	58	56	2
13	70	63	7	70	66	4
14	99	87	12	97	90	7
15	48	43	5	48	45	3
16	52	41	11	53	46	7
17	63	56	7	66	61	5
18	41	37	4	43	40	3
19	46	40	6	48	43	5
20	63	58	5	62	58	4
21	59	53	6	60	55	5
22	59	58	1	59	54	5
23	63	60	3	64	56	8
24	55	51	4	55	48	7
25	75	70	5	75	67	8
26	98	92	6	95	88	7
27	70	63	7	70	62	8
28	45	42	3	43	40	3
29	33	29	4	32	28	4
30	58	48	10	57	47	10
31	126	116	10	125	115	10

random-effects variance for the interaction is effectively zero.

A third approach, that of Lee and Nelder's Hierarchical Generalised Linear Model system (24) can be implemented in GenStat®. The main effect of centre can be either fixed or random. The results are very similar with the two approaches. Treating centre as fixed with treatment-by-centre alone as random, the estimate of the odds ratio is again 0.71 and the estimate of the random effect variance is negligible at 0.00009.

Finally, a meta-analytic approach can also be revealing. We proceed as follows: first we take the fourfold tables formed by the classifications of treatment (placebo/propranolol) and outcome (alive/dead) for each centre, adding 0.5 to each of the four frequencies to produce new values, *a, b, c, d*. Then we calculate the estimated log-odds ratio for centre *i* as

$$\hat{\theta}_i = \log\left(\frac{a_i c_i}{b_i d_i}\right)$$

and estimate its standard error as

$$SE_i = \sqrt{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}} \ .$$

We then proceed to analyse these estimates. For example the META procedure of GenStat® produces a Galbraith or radial plot (25) as shown in figure 2 where we see the standardised values $z_i = \frac{\hat{\theta}_i}{SE_i}$ plotted against the precision $\frac{1}{SE_i}$. The fixed-effects estimator is then given by the slope of the least squares regression line through the origin and is equal to –0.295 on the log-odds scale or 0.74 on the odds ratio scale, for which the 95% confidence limits are 0.59 and 0.94 respectively. The regression line through the origin is the thick line on the plot. Also shown are two control lines (dashed) at ±1.96. If there is no over-dispersion then we should expect that 95% of the points should lie between these lines. In fact they *all* do.

The graphical presentation suggests, therefore, that there is no variation in the treatment effects from centre to centre and this is backed up by the meta-analysis itself in the META procedure. The Q statistic for heterogeneity is 25.2 on 30 degrees of freedom and therefore less than its expected value under a hypothesis of heterogeneity; a random-effects analysis using the method of Hardy and Thompson (26) produces an estimated odds ratio of 0.75 with 95% confidence limits of 0.62 and 0.92 and a negative estimated random-effects variance of -0.07.

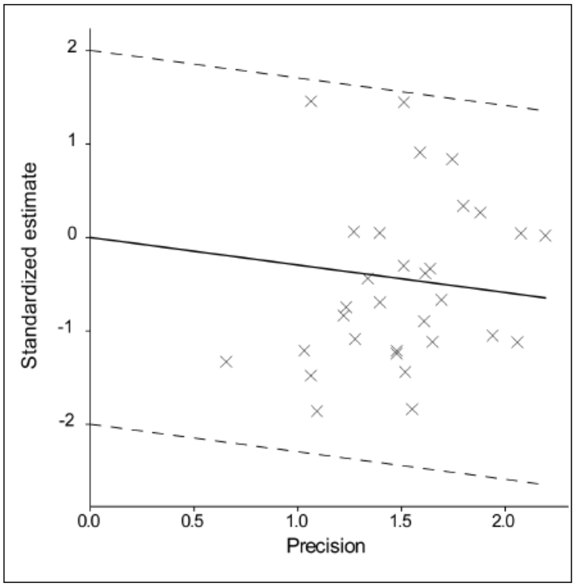

Figure 2. Galbraith Plot of the data from the BHAT. Standardised estimates (ratios of estimates to standard errors) for the 31 centres plotted against precisions (reciprocals of the standard errors).

To sum up, there is actually slightly less variation than one would expect by chance. On the log-odds scale, at least, it seems that the treatment effect is plausibly constant (14, 15).

Interactions more generally

Here, of course, *centre* is a factor comprised of a large number of groups that may be treated similarly. (See discussion above). It lends itself to treatment in terms of a single variance of effect using what is sometimes called a *hierarchical* model (the hierarchy being patients within centres). Other classifications do not so readily lend themselves to being handled in this manner.

For example in a heart disease trial we might have patients classified by sex, smoking status, drinking habits, age, social class, body mass index and family history of disease. Suppose these are treated as factors with few levels: the two sexes; never, ex- and current smokers, abstainers, moderate and heavy drinkers, etc. Each of these will have too few levels to be useful in estimating variation of effects and only the most enthusiastic Bayesian will wish to model them together. [Although techniques to do this *do* exist (27)].

The common statistical approach, then, is to compare treatment effects by the levels of a factor (28, 29). For example, the sex-by-treatment interaction is calculated as a *double* difference as follows: active minus placebo amongst males *minus* active minus placebo amongst females. The fact that each of these quantities being compared has on average half as many patients as for the simple treatment difference averaged over the sexes means that it is estimated with less precision (30). This fact is simply demonstrated in two lines of algebra and has been known at least since Jerzy Neyman pointed it out in 1934 (31). The consequence is that clinical trials and other studies generally guarantee less precision in measuring interactive effects than main effects. It also seems reasonable to suppose that interactive effects will be less marked than main effects and that one should be suspicious about apparently impressive effects. Hence interactions in general are an area where statisticians are wary because they know that 'chance rules'.

Second example: prognostic factors in a breast cancer trial

A retrospective analysis of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial (32) by Dowsett et al. (33) looked at treatment effects in 9366 patients with primary breast cancer classified into subgroups by receptor status with time to recurrence as the outcome. (Here, Arimidex © is the proprietary name for anastrozole.) With ER standing for oestrogen receptor, PgR for progesterone receptor, + for positive, - for negative and nk for not known, there were nine possible groups that could be defined by cross-classification according to the two receptors, although Doswett et al. looked at response in only six of these: ER+/PgR+, ER+/PgR-, ER-/PgR+, ER-/PgR-, ER+/PgRnk and ERnk/PgRnk. The data giving the status of the patients (recurrence or not) after a median follow up of 68 months are summarised in Table 2, which is based on Dowsett et al.'s Table 1 (See acknowledgements).

The data were analysed by Doswett et al. using the Cox regression (34) both unadjusted and adjusted for various covariates. If A stands for anastrozole, T for tamoxifen and for C the combination of both, then the contrast A versus T came out significantly in favour of A with a hazard ratio (HR) of 0.79, a 95% confidence interval of 0.70 to 0.90 and a *p* value of 0.0005. The contrast C versus T was not significant: HR = 0.97(0.86,1.10) *p* =0.6.

However, looked at subgroup by subgroup, the effect of A versus T was found to be significant only in the ER+/PgR- group: HR = 0.43(0.31,0.61) *p* <.0001. The authors very wisely wrote of this finding "… it should be considered as hypothesis generating and requires confirmation before it influences clinical decision making." (33, p. 7515).

Note that the above finding implies a very high level

Table 2. Patients and events in the Arimidex, Tamoxifen, Alone or in Combination study cross-classified by treatment and receptor status (ref. 29).

Treatment		Anastrozole		Combination		Tamoxifen	
ER	PgR	Patients	Events	Patients	Events	Patients	Events
+	+	1930	191	1875	205	1904	222
	–	451	50	492	102	429	102
	NK	167	22	170	24	181	20
–	+	63	17	81	22	76	25
	–	233	66	220	71	250	79
	NK	25	7	19	5	23	2
NK	+	7	2	6	1	8	1
	–	5	1	5	0	3	0
	NK	244	46	257	49	242	47

ER = oestrogen receptor; NK = not known; PgR = progesterone receptor.

of interaction. Cross-classifying patients by ER and PgR receptor status already involves an interaction of receptors. The finding that A is effective versus T but that C (which equals A+T) is not, is arguably another interaction, so a differential effect of A depending on whether or not it is given with T, and also depending on ER status but further modified by PgR is, from one point of view, a four-way interaction. The fact that such a high degree of interaction is involved and that the effect, however plausibly explained afterwards, was not expected *a priori* are grounds for caution.

Of course, further analysis of trials is a perfectly acceptable activity provided that, as was the case with Dowsett, its exploratory nature is accepted. In this spirit I now wish to illustrate a common statistical approach to looking at such data. To simplify discussion and analysis, only patients with known receptor status are included. At the risk of some loss of information, the data are re-analysed using logistic regression (because this is what data in the form of Table 2 permit) rather than proportional hazards (Cox) regression. For this analysis *treatment* is a factor with three levels (A,C,T) and there are two further factors, ER and PgR, each with two levels: - and +. The statistical approach to fitting would observe the following general principles:

1. Higher order interactions are not to be included in a model unless *relevant* lower order effects are also in the model (35, 36). For example, if the interaction of ER and PgR is in the model then the main effects of ER and PgR must be in the model. If the interaction of treatment and PgR status is in the model both the main effects of treatment and PgR status must be in the model but there is no requirement for ER to be in the model.
2. The significance of a factor or interaction is to be judged by considering the difference it makes to the model once other factors have been included in the model. Thus the effect of PgR is the effect of adding it to a model that already has ER in it and vice versa.
3. In a designed experiment, such as a clinical trial, where the main focus is on treatments, the covariate (prognostic) model should be established first [ideally, and certainly always in a drug regulatory context, by pre-specifying it in the protocol as required by the International Conference on Harmonisation guidelines (37)]. The effect of treatment is then judged according to the difference it makes when added to the covariate model.
4. Other things being equal one should prefer simpler models to more complex ones.
5. Treatment-by-prognostic factor interactions should not be judged confirmed unless the particular interactions found have been pre-specified in the protocol and an appropriate strategy for dealing with multiple hypotheses has been identified.

Illustration of this general approach with reference to the ATAC trial

This approach is illustrated in the GenStat® analysis shown in Table 3. The first and second columns show the steps of the fitting process (there were 11 in total). The third column, headed *df* for *degrees of freedom,* indicates the number of parameters added to or taken away from the model at each step. Frequently these values are 1 or -1 but note that since treatment is a factor with three levels it requires two parameters to describe it so that for step 6 for example, where treatment was added to the model, two parameters were added. The column headed *Deviance* is a measure of the increase in fit provided by the term added and, under the null hypothesis that this term is irrelevant, it should have an approximately Chi-square distribution with degrees of freedom given by df. The final column gives the p-value associated with the test of the null hypothesis that the term in question is not needed. It should also be noted that GenStat® indicates interactions of factors by joining them with a dot. Thus ER.Treatment is the interaction of ER and treatment.

Before proceeding to illustrate the fitting process, the first point to note is that since this is a post-hoc analysis I clearly cannot have specified any model in advance and therefore in the light of point 5 above all results must clearly be deemed to be of exploratory value.

In step 1 the term ER is added and is clearly significant ($p < 0.001$) and in step 2 PgR is added and this too is significant ($p < 0.001$). It now becomes necessary, however, to subtract ER from the model to check that its significance is not caused by association with PgR (see point 2 above). This is done in step 3 where its significance is confirmed and so it is

Table 3. Analysis of deviance for logistic regression for the Arimidex, Tamoxifen, Alone or in Combination data illustrating various modelling principles.

Step	Change	df	Deviance	Approximate Chi pr
1	+ ER	1	178.408	<0.001
2	+ PgR	1	47.313	<0.001
3	– ER	-1	-82.388	<0.001
4	+ ER	1	82.388	<0.001
5	+ ER.PgR	1	8.245	0.004
6	+ Treatment	2	17.705	<0.001
7	+ ER.Treatment	2	0.854	0.653
8	+ PgR.Treatment	2	10.669	0.005
9	– ER.Treatment	-2	-5.032	0.081
10	+ ER.Treatment	2	5.032	0.081
11	+ ER.PgR.Treatment	2	3.301	0.192
	Total	11	266.494	

df = degrees of freedom; ER = oestrogen receptor; PgR = progesterone receptor.

added back in step 4. In step 5 the significance of the interaction of ER and PgR is established ($p = 0.004$). At this stage the covariate structure of the model is established. This becomes the standard against which the effect of judging treatment is examined (point 3 above). Ideally this covariate model would have been pre-specified based on knowledge gained from previous trials and one could have proceeded right away with step 6, which establishes that the effect of treatment is significant ($p < 0.001$). In steps 7 and 8 the effects of adding the interactions with ER and PgR are examined. Note that because PgR.Treatment is added after ER.Treatment its p value of 0.005 is relevant but that the effect of ER.Treatment should be judged when it is added to PgR.Treatment and this has to be examined by subtracting it again as in step 9. The result is not significant but at 0.081 the p-value is much lower than that in step 7, where it was 0.653. Finally if one wishes to look at the three-way interaction, principle 1 above requires that ER.Treatment be put back in the model (as it is in step 10) and then in step 11 the three-way interaction is examined and found to be not significant ($p = 0.192$).

Thus, on the basis of the principle stated in point 4 above, the final model arrived at is that of step 9 and it includes both receptor factors and their interaction with each other as well as treatment and its interaction with PgR but not with ER. One should be cautious, however, in interpreting this finding. As regards PgR status, the number of patients are split 5929/ 2075 +/- with a 682/470 split of events, whereas the corresponding splits for ER status are 7081/923 and 872/280. In short, there are far fewer ER- patients and events than for PgR- and this affects the power of any tests of interactions involving ER status.

However, if we accept the data as we find them then the final model does not include the interaction of treatment with ER status. Table 4 shows the number of expected events this model would predict for these data together with the observed number of events. Category by category, the fit is remarkably good. Of course one now needs to look more carefully at the final model, in particular because treatment is a factor with three levels, to see what the implications are. (This process will not be illustrated here.) Nevertheless, the conclusion is somewhat different to that of Dowsett at al. who concentrated on differential effects of treatment according to the *combination* of ER and PgR. The analysis here suggests that only PgR is needed.

The data used by Dowsett et al., being survival times rather than events, are somewhat different and, furthermore, these authors have an expert knowledge of the trial (33). In any case, the purpose of this re-analysis here has been to illustrate some general points about fitting models rather than to challenge previous findings and there was never any intention to say comment on the results of the ATAC trial. In short, the purpose was didactic only: to illustrate some general approaches to modelling. Furthermore, both of the analyses, the one presented by Dowsett and, especial-

Table 4. Expected events according to the final model and observed events, cross-classified by receptor status and treatment.

ER		+		–	
Treatment	PgR	Expected	Observed	Expected	Observed
A	+	191	191	17	17
	–	61	50	55	66
C	+	203	205	24	22
	–	100	102	73	71
T	+	224	222	23	25
	–	93	102	88	79

A = anastrozole; C = combination; ER = oestrogen receptor; PgR = progesterone receptor; T = tamoxifen.

ly, the one presented here, are exploratory rather than confirmatory. In general, such analyses need to be compared with results from other studies, ideally involving the same treatment, but, if not, using treatments from the same class of drugs. For example, the BIG 1-98 study comparing letrozol to tamoxifen found a remarkably similar overall treatment effect to that found comparing anastrozole to tamoxifen in the ATAC study but did not find a PgR interaction (38).

Advice

Some general rules of procedure may be helpful in dealing with subgroup effects and interactions.

The first is to tread warily. It is necessary to have some way of ensuring a conservative approach. Adjusting p-values by using, for example, Bonferroni corrections is one of the least useful but it is better than ignoring the problem altogether. Some way of taking account of the totality of results is needed. Hierarchical approaches, where appropriate, will help shrink extreme results towards some common mean and this general philosophy is at the heart of the Bayesian approach.

The second is to pre-specify. Observed interactions are more plausibly genuine if they are expected. The frequentist may regard the discipline of pre-specifying in itself as valuable. The Bayesian may regard this as mere superstition but will certainly give more credence to effects that are believed on the basis of prior knowledge to be more likely. It will be useful and it increases credibility to have this prior belief explicitly documented. Of course, a confirmation of a result by a subsequent study is an even more impressive validation of a subgroup finding (39), since the first trial has the status of a evidence-based pre-specification for the second (2).

The third is to find a suitable scale for analysis. Interactions that are present on one scale will often disappear on another and some scales are more plausibly *additive* (that is to say do not show interaction) than others. For binary data, for example, the log-odds scale is much more plausibly additive than the risk difference scale and this is one reason why a meta-analysis is not usefully reported in terms of numbers needed to treat (40). An analysis on an additive scale without interactions will usually provide a much more useful and reliable summary of the treatment effect. If one then wants to make a prediction on a scale other than that used for analysis, this can often be achieved with a little ingenuity and some further background information. A fine example is given by Glasziou and Irwig (41).

The fourth is to proceed formally by building models for random effects as illustrated above for the first example or for interactions as shown for the second. Do not compare significance across strata to judge whether interactions have occurred (28). Remember that an effect with 50% power will only be significant half of the time it is tested and that in consequence the result ‘significant’ is likely to have poor inherent reproducibility. [In fact, this is a point perhaps overlooked by Hirschhorn et al. (19) in their otherwise interesting study of the reproducibility of genetic effects].

Finally, it cannot be stressed too strongly that the need for careful thought, caution and appropriate analysis, already apparent in the analysis of main effects, applies *a fortiori* to subgroups and interactions.

Acknowledgements
I am grateful to Novartis for funding and to the ATAC steering committee and in particular Jack Cuzick and Mitch Dowsett for providing data. I also thank two referees for their helpful comments.

Conflict of Interest Statement

I have acted as a consultant to the pharmaceutical industry for many years and my clients have included Novartis and AstraZeneca. (Products from both of these companies are mentioned in the paper.) I own stock in Novartis. As an academic my career is furthered by the publication of scientific papers.

References

1. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. Stat Med 2002; 21: 2909-2916.
2. Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. Med J Aust 2004; 180: 289-291.
3. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. J Clin Epidemiol 1996; 49: 395-400 [see comments].
4. beta-Blocker Heart Attack Study Group. A randomized trial of propranolol in patients with acute myocardial infarction. II. Morbidity results. JAMA 1983; 250: 2814-2819.
5. beta-Blocker Heart Attack Study Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. JAMA 1982; 247: 1707-1714.
6. beta-Blocker Heart Attack Study Group. Beta Blocker Heart Attack Trial: design features. Control Clin Trials 1981; 2: 275-285.
7. beta-Blocker Heart Attack Study Group. The beta-blocker heart attack trial. beta-Blocker Heart Attack Study Group. JAMA 1981; 246: 2073-2074.
8. James W. The will to believe: an address to the Philosophical Clubs of Yale and Brown Universities. In: James W ed The Will to Believe and other Essays in Popular Philosophy. New York; Longman 1897.
9. Altman DG, Matthews JN. Statistics notes. Interaction 1: Heterogeneity of effects. BMJ 1996; 313: 486.
10. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-1069 [see comments].
11. Peto R. Misleading subgroup analyses in GISSI. Am J Cardiol 1990; 66: 771-772.
12. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. Statistics in Medicine 1990; 9 (6): 657-671.
13. Kravitz RL, Duan NH, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004; 82: 661-687.
14. Senn SJ, Harrell FE, Jr. On subgroups and groping for significance. J Clin Epidemiol 1998; 51: 1367-1368.
15. Senn SJ, Harrell F. On wisdom after the event. J Clin Epidemiol 1997;50:749-751 [see comments].
16. Senn S. Individual response to treatment: is it a valid assumption? BMJ 2004; 329: 966-968.
17. Senn SJ. Individual Therapy: New Dawn or False Dawn. Drug Information Journal 2001; 35 (4): 1479-1494.
18. Ioannidis JP. Why most published research findings are false. PLoS Med 2005; 2: e124.
19. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med 2002; 4: 45-61.
20. Parker AB, Naylor CD. Interpretation of subgroup results in clinical trial publications: Insights from a survey of medical specialists in Ontario, Canada. Am Heart J 2006; 151: 580-588.
21. Senn SJ. Statistical Issues in Drug Development. Chichester; John Wiley 1997.
22. Li ZQ, Chuang-Stein C, Hoseyni C. The probability of observing negative subgroup results when the treatment effect is positive and homogeneous across all subgroups. Drug Inf J 2007; 41: 47-56.
23. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - The analysis of case-control studies. IARC Sci Publ 1980: 5-338.
24. Lee Y, Nelder JA. Hierarchical generalized linear models. Journal of the Royal Statistical Society Series B- Methodological 1996; 58 (4): 619-656.
25. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. Stat Med 1988; 7: 889-894.
26. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med 1996; 15: 619-629.
27. Lindley DV, Smith AFM. Bayes estimates for the linear model (with discussion). Journal of the Royal Statistical Society B 1972; 55: 1-41.
28. Matthews JN, Altman DG. Interaction 3: How to examine heterogeneity. BMJ 1996; 313: 862.
29. Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not P values. BMJ 1996; 313: 808.
30. Lachenbruch PA. A note on sample size computation for testing interactions. Stat Med 1988; 7: 467-469.
31. Neyman J. Comment on Mr Yates's paper. Journal of the Royal Statistical Society (supplement) 1934; 2: 235-241.
32. Howell A, Cuzick J, Baum M, Buzdar A, Dowsett M, Forbes JF, Hoctin-Boes G, Houghton J, Locker GY, Tobias JS; ATAC Trialists' Group. Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer. Lancet 2005; 365: 60-62.
33. Dowsett M, Cuzick J, Wale C, Howell T, Houghton J,

Baum M. Retrospective analysis of time to recurrence in the ATAC trial according to hormone receptor status: an hypothesis-generating study. J Clin Oncol 2005; 23: 7512-7517.

34. Cox DR. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society Series B 1972; 34: 187-220.
35. Nelder JA. A reformulation of linear models. Journal of the Royal Statistical Society A 1977; 140: 48-77.
36. Nelder JA. The importance of marginality rules. Journal of the Royal Statistical Society Series C-Applied Statistics 1998; 47: 447-448.
37. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. Stat Med 1999; 18: 1905-1942.
38. Coates AS, Keshaviah A, Thurlimann B, Mouridsen H, Mauriac L, Forbes JF, Paridaens R, Castiglione-Gertsch M, Gelber RD, Colleoni M, Láng I, Del Mastro L, Smith I, Chirgwin J, Nogaret JM, Pienkowski T, Wardley A, Jakobsen EH, Price KN, Goldhirsch A. Five years of letrozole compared with tamoxifen as initial adjuvant therapy for postmenopausal women with endocrine-responsive early breast cancer: update of study BIG 1-98. J Clin Oncol 2007; 25: 486-492.
39. Simes RJ, Gebski VJ, Keech AC. Subgroup analysis: application to individual patient decisions. Med J Aust 2004; 180: 467-469.
40. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses-sometimes informative, usually misleading. BMJ 1999; 318: 1548-1551 [see comments].
41. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. BMJ 1995; 311: 1356-1359.