Wiestner, M., Rohde, H., Helle. O., Krieg, T., Timpl, R., and Müller, P. K.: Low rate of procollagen conversion in dermatosparactic sheep fibroblasts is paralleled by increased synthesis of type I and III collagens. EMBO J. 1: 513-516, 1982.

Wilder-Smith, C. H., Raue, F., Holz-Gottswinter, G., and Ziegler, R.: Procollagen-III-peptide serum levels in Paget's disease of the bone. Klin. Wochenschr. 65: 174-178, 1987.

Witter, J., Roughley, P. J., Webber, C., Roberts, N., Keystone, E., and Poole, A. R.: The immunologic detection and characterization of cartilage proteoglycan degradation products in synovial fluid of patients with arthritis. Arthritis Rheum. 30: 519-529, 1987.

Wolmann, M., and Kasten, F. H.: Polarized light microscopy in the study of the molecular structure of collagen and reticulin. Histochemistry 85: 41-49, 1986.

Woolley, D. E.: Mammalian collagenases. In: Piez, K. A., and Reddi, A. H. eds.: Extracellular matrix biochemistry. Elsevier, New York, 1984: 119-157.

Wright, V., and Johns, R. J.: Observations on the measurement of joint stiffness. Arthritis Rheum. 3: 328-340, 1960.

Wu, C. H., Donovan, C. B., and Wu, G. Y.: Evidence for pretranslational regulation of collagen synthesis by procollagen propeptides. J. Biol. Chem. 261: 10482-10484, 1986.

Würz, H., and Crombach, G.: Radioimmunoassay of laminin P1 in body fluids of pregnant women, patients with gynaecological cancer and controls. Tumor Biol. 9: 37-46, 1988.

Zettner, A., and Duly, P. E.: Principles of competitive binding assays (Saturation analyses). II: Sequential saturation. Clin. Chem. 20: 5-14, 1974.

Yamada, K. M.: Cell surface interactions with extracellular materials. Annu. Rev. Biochem. 52: 761-799, 1983.

Yurchenco, P. D., and Furthmayr, H.: Self-assembly of basement membrane collagen. Biochemistry 23: 1839-1850, 1984.

Yurchenco, P. D., and Ruben, G. C.: Basement membrane structure in situ: Evidence for lateral associations in the type IV collagen network. J. Cell Biol. 105: 2559-2568, 1987.

Bias in double-blind trials

Peter C. Gøtzsche

This review has been accepted as a thesis together with 6 previously published papers, by the University of Copenhagen, February 15, 1990, and defended on May 10, 1990.

Official opponents: Ib Lorenzen, Professor, M.D., and Henrik R. Wulff, M.D.

Reprints: Peter C. Gøtzsche, Ellehegnet 18, DK-2950 Vedbæk.

INTRODUCTION

Bias in research may be defined as anything which tends to make a conclusion differ systematically from the truth (1).

The gold standard for unbiased evaluation of therapies is the randomized clinical trial (RCT), which represents a major advance in medicine. In 1898, Fibiger reported a study in which patients were treated with or without anti-diphteria serum according to date of admission (2). The first time random numbers were used in treatment allocation appears to be in the trial of streptomycin in tuberculosis, reported in 1948 by the Medical Research Council in Britain (3).

However, neither randomization nor double-blinding of therapies guarantees freedom from bias. The reading of published reports of RCTs may lead to a biased evaluation of therapies if 1) published reports represent a biased sample of all trials carried out, 2) the collected reports represent a biased sample of all reports, 3) the design, analysis, or interpretation of the trials are biased, 4) the reader is biased.

It is an empirical question to what degree these factors influence the truth about therapies. Bias that may escape detection in a narrative literature review may become apparent in a *meta-analysis* (ie a quantitative integration of research findings from several studies with statistical methods (4)). Accordingly, one may evaluate the credibility of a research area by collecting as complete a sample of papers as possible, irrespective of language and place of publication, and by judging whether the collection of results is both likely and compatible with other knowledge.

To elucidate sources, amount and importance of bias in drug trials, this review focuses particularly on reports of double-blind RCTs of nonsteroidal anti-inflammatory drugs (NSAIDs) in rheumatoid arthritis, which I have analyzed in depth (5-10). Apart from a few dose-response studies, all the trials involved a comparison of two or more active drugs; in addition, some of them had a randomized placebo.

PUBLICATION BIAS

Publication bias, defined as preferential publication of "positive" (statistically significant) results over "negative" ones, is an important bias to which companies, authors, and editors contribute.

I know of one trial that was never published for the very reason that the company's drug was less effective than the control drug. Others have also described company pressures on investigators or on journals to prevent publication of unfavourable results (11).

In psychology, the probability that potential authors would submit a manuscript before further data collection was 0.49 if the result was positive, but only 0.06 if it was negative (12). The effect size in cancer trials was smaller in unpublished studies (13), whereas in studies of sex bias in counselling and psychotherapy, the effect size was of the same magnitude in unpublished and published studies, but with the opposite sign! (14). In a review with several deficiencies, pointed out by the authors themselves, 14% of unpublished RCTs favoured the new therapy compared with 55% of the published ones (15).

A manuscript guideline in *Diabetologia* stated simply that reports with negative outcomes were not desired (16), and a journal replied: "The editorial board regrets to inform you that the abundance of paper submitted for publication makes it impossible for us to use space to publish "negative trials"" (1).

The null hypothesis was rejected in 97% of 294 articles in psychological journals (17), and in all medical experiments in a medical journal (18). In 62% of reports of RCTs in gastroenterology, a significant difference between the treatments was found which favoured the control (established treatment or placebo) in only 5 of 191 positive trials (19). A significant difference was reported in 33% of 218 cancer trials, half of which were randomized; the chance of a favourable conclusion was not related to endpoint or sample size (20). A significant difference in effect appeared in 34% of 68 reports of newly introduced analgesics and NSAIDs (21), and in 38% of 196 NSAID trial reports (6), although all of these trials compared active drugs.

However, medical progress is rarely that successful (22), and this preponderance of positive results would definitely not be expected in NSAID trials with a median sample size per group of 27 (6), that compared active drugs in a highly fluctuating disease (23, 24) with a large interobserver variation (25-30), and with diurnal variation (27-34). In fact, differences between NSAIDs seem to be negligible compared with other sources of variation (35, 36).

In a meta-analysis of grip strength, I initially found no evidence of important publication bias favouring the new NSAID over the control drug (9). The differences between the drugs showed the expected symmetrical funnel pattern (37) and there was no overrepresentation of small trials with large differences that favoured the new drug (Fig. 1). However, although there was no difference between the drugs, the authors reported significant differences in favour of the new drug in 12 of 175 trials, or more often than expected ($p < 0.01$), but only 4 significant differences in favour of the control drug, as expected (2.5% of 175). This suggests biased data analysis (6, 9).

Assuming negligible differences between NSAIDs (6, 8-10, 35, 36), the number of trials in which all significant differences in effect favoured the new drug should be similar to the number in which all

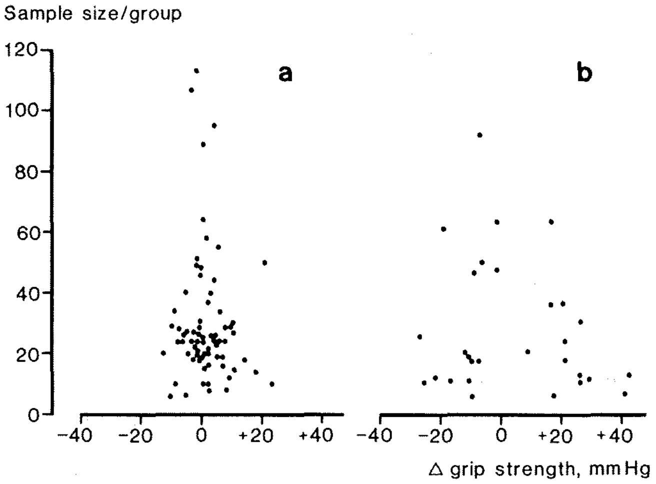

Fig. 1 *Sample size of NSAID trials in rheumatoid arthritis related to difference in grip strength between new drug and control drug. a) 70 crossover trials, b) 29 trials with parallel groups. Reproduced with permission (9).*

differences favoured the control. This was not so, since 53 and 12 trials were identified, respectively ($p < 0.0001$, sign test). However, if the newer drugs had been more effective, one would expect more significant trials, the larger the sample size. Nor was this the case; in 109 crossover trials which compared only two active drugs, the proportion of positive trials was unrelated to sample size (Table 1, $p = 0.33$, Mann-Whitney test).

Out of a total of 1,545 effect variables in 196 NSAID trials (6), significant results in favour of the new drug were claimed for 184 (11.9%) and in favour of the control drug for 29 (1.9%). The latter result is close to the expected 2.5%, assuming equally effective drugs, especially if allowing for omission of some significant results in favour of the control drug (6). Thus, bias seemed to have caused approx. 80% ((11.9-2.5)/11.9) of the significant differences in favour of the new drug. This means that publication bias is not simply a matter of selective submission and acceptance of correct positive results, but may predominantly be caused by biased data analysis (6). Most of the significant results in NSAID trials seemed to have been caused by bias, and the majority of the remaining by type I errors.

Suggestions for avoiding publication bias

Trials planned to be large would be expected to become published irrespective of the result because of the effort and number of people involved. Unfortunately, most trials are very small (19-21, 38-45), which may be defensible for phase 2 trials, performed when little is known of a drug. However, the median number of patients per treatment group was only 27 in the NSAID trials (6), although all were phase 3 or 4 trials.

When interim analyses are made without being reported (46), and without adjustment of the significance level, the likelihood of publication of false significant results increases. *Peto et al* stated simply that "significant results from small trials are often wrong" (47). Similarly, single investigators in a multi-centre trial should not be allowed to publish their own results (6, 11, 48).

It is unethical to publish misleading results (49). Hence, it is not defensible to do phase 3 or 4 trials that are too small to give a reliable answer about the value of a drug, since the population of published results would be expected to become misleading because of publication bias. To reduce the risk of introduction of unnecessary or even harmful therapies (20), the ethical committees and the drug agencies should be concerned with statistical aspects of proposed trials. Sample size calculations are often missing (21, 43, 44, 50-53), but it would be simple to ask for them.

Registers of clinical trials could be made obligatory, and comparison of published and unpublished reports from such registers may be valuable (13).

Bias in data analysis may be limited by blinding (6).

Inflation of the sample of reports

A sample of reports is inflated if it contains multiple publications of the same data. Among 244 reports on NSAID trials, I identified 44 (18%) as multiple publications (5, 7). The fact that they had been published elsewhere was not noted in 32 of the 44 reports. In six of the 31 trials involved (some trials were published more than twice), multiple reports were only revealed after thorough cross-checking by authors, drugs, and results, and they might therefore be mistaken for separate trials in a meta-analysis (7). In fact, before performing two subsequent meta-analyses (8, 9), I found one further case of multiple publication which was overlooked previously since both reports had only one, but a different, author.

Bias caused by inflation would be avoided if authors adhere to the Vancouver code (54), with proper cross-references.

REFERENCE BIAS

If reference lists in reports particularly reflect the authors' prejudices, collecting reports by means of reference lists may lead to a biased sample.

A bias towards a selection of positive references on the new drug was demonstrated for the NSAID trials. There was an overrepresentation of reports which in their reference lists contained a higher proportion of references with a positive outcome for the new drug than among all articles assumed to have been available to the authors, both when the language was disregarded ($p < 0.01$) and when only references in English were considered ($p < 0.05$) (5).

The reference bias was not caused by overrepresentation of highly cited journals among reports with a positive selection of references, and it was at its highest when the authors had had many articles to choose from (Table 2). The reference bias was caused mainly by a biased selection of references on indomethacin, the most common control drug. Thus, the more an area has been researched, the more self-confirmatory the conclusions may become.

Since there was no trend towards a positive selection of reports in a MEDLINE search or in the lists provided by the companies on their own drug (5), an extensive literature search along these lines may diminish or eliminate the influence of reference bias on the reports sampled.

In an analysis of the references of 53 articles on adverse reactions to phenylpropanolamine, some evidence of bias against the drug was found with relatively fewer citations to human safety studies (counter-hypothesis literature) than to similar reports of adverse drug reactions (55). Case reports received an average of 4.05 citations/reference, human experimental studies 2.94 citations/reference. However, no analytical statistics were presented in support of the supposed bias, and even assuming an overrepre-

Table 1. *Number of crossover trials in which all significant differences in effect favoured the new drug, related to sample size (p=0.33, Mann-Whitney test).*

No. of patients	No. of positive trials	Total no. of trials
1-19	3	12
20-29	6	34
30-39	10	30
40-59	3	17
60+	6	16
Total	28	109

Table 2. *Number of articles with positive, neutral, and negative selection of references to the new drug in relation to number of possible references. Reproduced with permission (5).*

	Positive selection	Neutral selection	Negative selection	Bias not possible	Total
1-3 possible references	5	5	4	26	40
4-7 possible references	16	3	11	8	38
≥ 8 possible references	23	2	7	1	33

sentation of similar reports, this does not necessarily reflect bias, since the two types of articles are quite different, and the authors may simply have felt that case reports were slightly more relevant.

In future work on reference bias, meta-analyses of reports obtained mainly by reference lists could be compared with those made after extensive searches. Review articles and book chapters might also be examined for reference bias.

QUALITY OF DESIGN

Randomization

The randomization method is rarely reported (6, 43, 44, 50-53). Requests for clarification may produce laconic replies like "the company did it", "in blocks of ten", or "source material no longer available" (8). Everybody involved seems to assume that the method was correct and too unimportant to bother about. However, if the blinding of the randomization process is lost, eg if packages are labelled A and B (6), selection bias may result (56).

Blinding

The term "double-blind", typically viewed as a guarantee of immunity to bias, is used too freely. Trials called double-blind are not always so (6, 8, 57) or may become unblinded (58-61); 8% of the NSAID trials were probably not truly double-blind (6).

Information on the method is often lacking (6, 43, 51-53) and the efficiency of the blinding is practically never tested (6, 42, 50). However, in 34% of the NSAID trials, matching capsules were used (6) which may easily be opened by the patients. Further, it was stated in only 1% of the reports that none of the patients had had any experience with the drugs before, although the commonest controls, aspirin and indomethacin, have characteristic side-effects such as tinnitus and headache, respectively (62), which may easily be recalled. Hence, a trial may be biased if patients exposed to the control drug previously tend to be dissatisfied with and recognize this drug.

The efficiency of vaccinating a trial against bias should be tested, as recommended by the International Committee of Medical Journal Editors (63).

Recruitment

Reject logs are rarely reported (50) and none of the NSAID reports described eligible patients who were not entered in the trial (6). Thereby it may become more difficult to generalize the results, since patients volunteering for a trial are often quite different from those who don't (64).

Dose-response bias

If we compare a drug with itself, in two different doses, the situation depicted in Fig. 2 may arise (65). Dose A is preferable to B since it gives a similar effect with fewer side-effects. This would not lead us to conclude that the drug is better than itself. However, when A and B are two different drugs, we tend to conclude that A is better than B, although we only know that *given the chosen doses* A performed better than B.

Medical companies probably focus most on the effect when the very first drug within a therapeutic area is developed. Later, they would try to reduce the side-effects; in fact, new molecular analogues are often claimed to be as effective as the standard drug but with fewer side-effects (8, 42). Such claims may be caused by *dose-response bias,* ie a bias due to comparison of nonequipotent doses (65, 66). This bias may also occur when the true difference between the drugs is of the same magnitude for both the effect and the side-effect measure. For instance, if the side-effect measure is more sensitive (67) than the effect measure, the conclusion could be that the new drug is "equally effective", with significantly fewer side-effects.

It is therefore unsatisfactory when the method of side-effect registration is not reported, as in more than half the NSAID trials (6). All 39 trials with a significant difference in side-effects favoured the new drug (6). As discussed later, this is probably to some extent caused by biased data analysis. Further, since side-effects of both aspirin and indomethacin are clearly dose-related within the dose range used in NSAID trials (62, 68, 69), the control drugs might have been relatively overdosed. This hypothesis is supported by analyses of dose-response relationships in NSAID trials, which have shown little or no increase in effect at the higher doses (8-10), and by the fact that aspirin seems to be effective in a relatively low dose, 3.6 g daily (70). Most important, however, these false or true differences in side-effects did not seem to matter, since the patients preferred indomethacin similarly often as the new drugs in a meta-analysis of 1,262 patients (8).

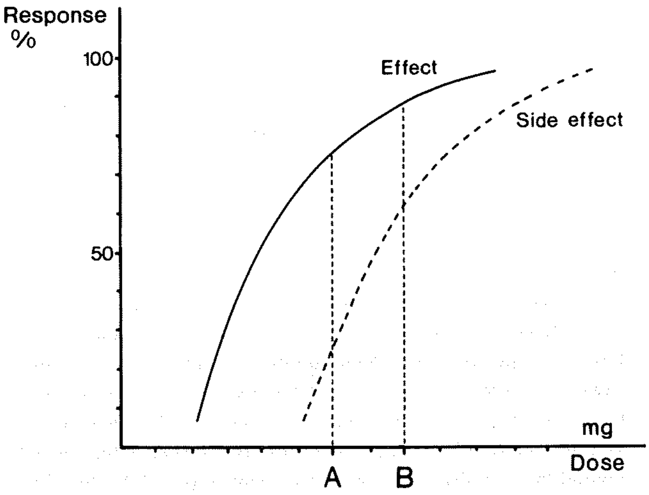

Fig. 2. *Theoretic dose-response curves for the effect and a side-effect, respectively, of a drug. A and B indicate two different dose levels. Double-arithmetic plots.*

Choice of dose was motivated in only 14% of the NSAID trials, and in only 2% for both drugs (6). Reference to randomized dose-response studies was only made in 1%. The dose ratio often varied with a factor of two or more in repeated comparisons of the same two drugs, again suggesting the existence of a dose-response bias in some trials.

The use of more than one dose level in a drug trial is recommended (65, 71, 72), preferably for both drugs. Contrary to popular belief, the sample size does not increase by 50% with an extra dose level, since more powerful statistical methods can be used. Should there be no difference between the two doses (which is an important information!), then, with the same total number of patients, the power of the study would decrease only little compared with the usual randomization, with half the patients on each drug (47, 72).

Multiple outcomes

"Investigators seem to have settled for what is measurable instead of measuring what they would really like to know." (*Edmund D. Pellegrino,* cited in (73)). When many effect variables are used and no main variable is chosen, the risk of bias and circular evidence is obvious. For instance, a variable may be chosen as main variable after the trial because a significant difference was found, reasoning that the variable is useful and relevant because of its discriminatory ability. In articles from major journals, a median of 6 endpoints was noted; a main variable was indicated in only a quarter of the reports (45), and in less than half in another survey (44). More than 70 different variables were reported in the NSAID trials; the results sections had a median of 8 (6). A main variable was selected beforehand in only 6% of the trials; in some other reports, it did not appear before the discussion section.

There is no agreement as to which variables should be used when evaluating the effect of NSAIDs (23, 71, 74-79). In reports on drug trials, and in methods articles (25, 80, 81), tautologies and circular reasoning are not uncommon, mainly because of lack of testable hypotheses (82, 83). The most common variable, found in 81% of the NSAID trials (6), was grip strength. In a meta-analysis (9), the median grip strength was 133 mm Hg, but the difference between new drugs and control drugs was only 1 mm Hg, and the full dose of a drug was not significantly superior to half the dose (difference of 3.5 mm Hg). The drugs were significantly better than placebo, but the difference of 12 mm Hg was less than what most rheumatologists considered to be of relevance. Thus, the most common variable, although easily measurable, appeared to be superfluous.

The reporting of the erythrocyte sedimentation rate (ESR) varied (6). It appeared as an effect variable in some reports, and as a laboratory measure, like haemoglobin, in others. In yet others, in which the ESR appeared as a laboratory measure in the methods section, but as an effect variable under results, there was an overrepresentation of significant results ($p = 0.006$) which usually favoured the new drug over the active control ($p = 0.06$) (6). Thus, the interpretation of the ESR seemed to be related to the trial's outcome for the new drug, despite convincing data showing that NSAIDs have no effect over *placebo* on the ESR (80, 84-88).

It is noteworthy that the effect variables presented in the methods and the results sections were all the same in only 51% of the NSAID trials (6). If variables are deleted from both sections, bias caused by interpretation of variables in the light of the results obtained may escape detection. In fact, inconsistency in the total number of variables reported was found in no less than five of the 31 trials subjected to multiple publication (7), which shows that such a bias is a real possibility.

Utility measures

Effect and side-effects are inseparable when symptomatic treatments are compared. A patient with severe pain may prefer a drug with side-effects if it helps. Since the aim of pragmatic trials is to decide which drug should be preferred (89), an overall utility measure seems most relevant as the main outcome variable. If not used, access to the results might bias the relative weighting of effect and side-effects (6). Further, when all that happens after the randomization is expressed in a utility or preference (8, 90, 91), it becomes easier to include withdrawals in the analysis. Also, the decision about whether a patient stopped due to lack of effect or because of side-effects, which may be difficult (92), would become irrelevant.

Patient preference was used in two-thirds of the crossover trials of NSAIDs (6). In a meta-analysis of patient preference in indomethacin trials, withdrawals were included under different assumptions (8). There was little evidence that the newer NSAIDs were better than indomethacin (Fig. 3); the best estimate of the difference in the proportion of patients who preferred the new drug and the proportion who preferred indomethacin was 5%. Interestingly, the two outlying studies were the only ones which I had not believed, due to other factors than their recorded patient preferences, in my previous analysis of NSAID trials (6).

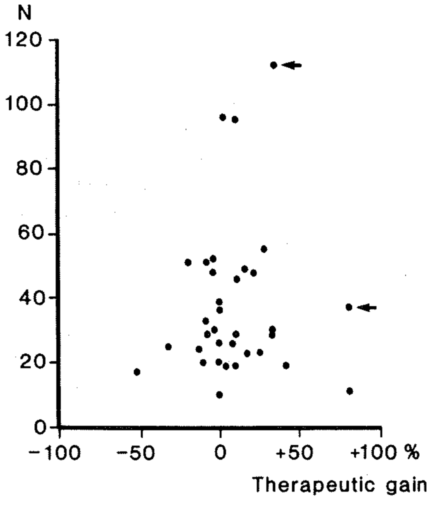

Fig. 3. *Difference in the proportion of patients preferring a new NSAID and the proportion preferring indomethacin (therapeutic gain) in 34 crossover trials (N = sample size minus number of withdrawals). Arrows mark two outlying trials. Reproduced with permission (8).*

Few trials with parallel groups used a similar utility measure (6), and it was also rarely reported in other surveys (42, 93).

QUALITY OF STATISTICAL ANALYSIS

In reviews of statistical analysis in medical research reports, frequent deficiencies and occasional errors (6, 41, 52, 94-102) and bias (6) have been reported. Some reviews are themselves questionable; in an editorial comment it was noted that the authors apparently had started out with the attitude that "each study was guilty of various errors and each had to prove beyond a shadow of doubt that in fact it was innocent" (103).

Bias in data analysis

Data analysis is to some extent subjective, as illustrated by the controversy over the University Group Diabetes Program trial (73). Choices must often be made, eg because of missing data and protocol violations, and one may use absolute values, change, percentage change, or geometric mean percentage change from baseline, or other transformations. Interestingly, discrepancies in analysis of potential importance for the interpretation of the report were found in 12 of the 31 NSAID trials published more than once (7).

In a letter from the Research Headquarters of a company to its local clinical trials monitor, it was argued at length that several side-effects might not be caused by the company's drug. The letter concluded: "Thus, even if one would admit that (the company drug) was not totally devoid of side effects, it would seem that only those in (two patients) might be attributable to (the company drug)". "On the other hand, there is little doubt that at least eight patients showed characteristic indomethacin side effects". The letter ended: "I am virtually certain that on close scrutiny of his data, Dr. (X) will arrive at the same conclusion – if he has not done it already".

The letter is the only example I have of internal company correspondence, received due to a request for information on withdrawals (8). The letter may be atypical, but, if not, it may explain why new drugs often appear to be better tolerated than control drugs. Further, such bias would tend to delay the time till atypical side-effects of new drugs become known. It gives food for thought that discrepancies between different reports in number of patients with side-effects or in number of side-effects were noted in five of the 31 NSAID trials that were subjected to multiple publication. They were inexplicable in four of them (7).

At a consensus conference in 1984, 31 of 36 academics agreed that in an industry sponsored trial, the manufacturer should not perform the final data analysis, whereas only 6 of 46 industrial representatives had the same opinion (104). However, since statisticians outside the industry may also be biased (73), double-blind trials should have blinded data analysis, which was used in only 1% of the NSAID trials (6).

A statistician who wishes to favour one drug over another has many options. Factors that may favour a new drug by increasing the number and the proportion of significant results in trial reports are listed in Table 3. Most of these were found and one or more of them were of importance in a quarter of the NSAID trials (6), but their relative weight could not be judged. In 12 of the trials (6%), significant results became nonsignificant on recalculation, and all errors favoured the new drug (6). In other surveys, claims based on erroneous results were found in 8% (96) and in 15% (97).

Confidence intervals (105) are a good alternative to the pollution of the literature caused by obsessive and biased significance testing, since estimation of the true difference is more relevant than null hypothesis testing.

Self-selected blinding

"There is none so blind as he who *will* not see" (106). Choice of a statistical test that assumes a less refined measurement scale than the one actually used may render significant results clearly nonsignificant and vice versa, depending on the cutpoint (107, 108).

Loss of information on effect variables was found in 26% of the NSAID trials, in 10% to such a degree that data on a continuous scale were reduced to a bi- or trinomial scale before presentation and analysis (6). For side-effects, the severity was registered in 26% of the trials, but usually only their frequency was analyzed. Finally, in 10%, no comparisons between, but only within the drug groups were made, ie only between baseline and subsequent values. In crossover trials, erroneous use of unpaired tests was noted in 10% for effect variables. For side-effects, unpaired tests were used equally often as McNemar's test or similar tests.

Such immense losses of information suggest bias, and I calculated significant differences favouring the control drug in two trials, and strongly suspected that correct analysis would have favoured this drug in a further 7 (6).

"Almost significant"

Confidence intervals or exact p values are rarely shown in negative trials (6, 19, 38, 43, 50-52), and the results are commonly interpreted as no important difference. In the NSAID trials, reported p values in the interval 5-10% for effect or side-effect measures mostly favoured the new drug ($p = 0.007$) (6). Thus, the decision to report a p value in this interval seems dependent on which drug it would favour.

Baseline comparability

Despite the randomization, important differences between the groups may exist, especially in small trials. Stratified analysis or adjustment were used in 9% of the NSAID trials (6) and in 20% in another survey (45), but whether the decision to use these procedures was biased could not be evaluated. It may be an advantage to adjust for baseline differences (45, 53, 73, 89, 105, 109-112), but results without adjustment should also be shown to facilitate meta-analyses.

Withdrawals

The handling of withdrawals can have a major impact on the results (113-118). However, withdrawals are often insufficiently described and even their number may be missing in trial reports (6, 19, 44, 50-52, 100). A distinction is often not made between patients withdrawn during the trial before the code is broken and patients excluded during data analysis, and the decision to exclude patients is almost never blinded (6, 118). Withdrawals were retained in the analysis of only 5% of the NSAID trials (6) according to the "intention to treat" principle, which is recommended to prevent bias (44, 47, 89, 113, 116-122). Correspondingly, withdrawals should be included in meta-analyses (8). Among NSAID trials published more than once, two were found in which patients who did not conform to the trial plan were not only omitted from the analysis, but from the entire publication (7). Thus, reports which do not mention withdrawals or exclusions, or that there were none, may be less reliable.

Table 3. *Factors that may increase the number and the proportion of significant results favouring a new drug. Reproduced with permission (6).*

1. Design bias
2. Selection of patients dissatisfied with control drug
3. Choice of dose
4. Selection of indices
5. Selective reporting among many variables
6. Ineffective blinding
7. Choice of statistical methods or no statistics at all
8. Handling of withdrawals
9. Handling of missing data and other uncertainties
10. Change in measurement scale before analysis
11. Choice of adjustment depending on result
12. Uneven distribution of prognostic factors
13. Wrong sampling unit for effect and side-effects
14. Wrong interpretation of within-groups analyses
15. Repeated testing on several groups or over time
16. Subgroup analyses
17. Selective reporting of $0.05<p<0.10$
18. Omission of significant results favouring the control
19. Wrong calculation
20. One-sided testing
21. Fraud
22. Publication bias

Wrong sampling unit and doubtful scale for effect measures

Use of a wrong sampling unit (123) violates the statistical assumption of independence between the observations. It may erroneously inflate the sample size, or may give undue weight to the results of one or a few atypical patients. Further, the assumptions for the measurement scale may be wrong, eg a patient with only one affected joint may have more pain than one with 20 affected joints.

In 63% of the NSAID trials (6), statistics on a wrong sampling unit for an effect variable were noted, most often as the average number of affected joints or the average of Ritchie's index (joint tenderness for 26 groups of joints is scored 0 to 3 and added) (25). Many joint indices were unexplained, which raises a suspicion of bias caused by construction of an index that "suits" the data.

The inflation record is held by a study, in which the sample size increased from 58 patients to 3,944 joints, and a p as low as 0.979126×10^{-8} was reported (6).

Wrong sampling unit for side-effects

In 39% of the NSAID trials, the number of patients with side-effects was not shown (6). A wrong sampling unit was used in 23%, mostly as incidence of separate side-effects, which in some trials resulted in 2-3 times as many side-effects as patients. This erroneous practice strongly disfavours the two commonest control drugs, indomethacin and aspirin, due to their conspicuous central nervous system side-effects, in addition to the gastrointestinal problems common for all NSAIDs (62). The occurrence of dyspepsia in 228 patients treated with indomethacin, usually 75 or 200 mg daily (68), was not dose-related (6% versus 10%), whereas headache clearly was (8% versus 36%) (69).

Wrong sampling units seem sometimes to be used on purpose to bias the results (6).

Regression towards the mean

Patients with a chronic fluctuating disease would be expected to enter a trial in bad periods, and thus improve independent of therapy (124). Similarly, when the worst affected joints are selected for examination (6), the trial will tend to show "improvement" in these. Due to this regression towards the mean effect (125, 126) and to patient and physician expectation, a claim of drug effectiveness in trials without a placebo control may be questionable. Such claims were made in 60% of the NSAID trials; however, the patients also "improved" on a randomized placebo ($p = 0.01$) (6).

Inflation of the type I error

The number of significant results in research reports will tend to become inflated when tests adequate only for two groups, such as the t-test, are used in trials with repeated looks over time or with several drugs or doses without adjustment of the significance level (127, 128). Inflation was noted in 23% of the NSAID trials (6), and in 17% (129) and 18% (45) of reports in other surveys.

Intake of test and rescue drugs

A varying intake of test drug might be acceptable in a pragmatic trial (89), whereas a differential intake of rescue drugs (eg analgesics in NSAID trials) may confound the comparison of the test drugs. Since drug intake is dependent on the trial's outcome for each patient, one cannot control for it by adjustment procedures, and misinterpretations may occur. If, for instance, the serum concentration of a drug is related to its effect, one may find that the lower the concentration, the larger the effect, but the explanation could be that patients in good periods take less of a toxic drug. Intake of drugs should be documented, which it rarely is (6, 50).

Bias in conclusion and abstract

Summarizing a report into an abstract involves judgement and exclusion of information. Data analysis is a similar process. Therefore, when an abstract is clearly biased, it is difficult to have much confidence in the rest of the report.

Doubtful conclusions were found in 72% of 149 analytical studies published in leading journals (95) and in 31 of 45 manuscripts on clinical trials submitted to the British Medical Journal (52). The abstract was misleading in 9 of 45 trials in surgery, and the conclusions were not supported by the data in 11 (100). In two-thirds of 171 reports on jejunoileal bypass for obesity, uncertainty was expressed about its advisability; however, among the remaining reports, more fatal cases occurred in the trials with a positive attitude ($p < 0.10$) (130). In a study of comparative trials, relatively fewer nonsignificant results were included in the summaries than significant ones (odds ratio 1:9.2) (45).

In the NSAID reports, significant results were omitted from the abstract more often when they favoured the control drug ($p = 0.08$) (6). Doubtful or invalid statements were noted in the conclusion or abstract of 76% of the reports. In 82 reports (42%), bias in the conclusion or abstract consistently favoured one of the drugs, which was the control drug in only one report and the new drug in the remaining 81 ($p = 3.4 \times 10^{-23}$) (6). Among multiple publications, the conclusion became more favourable for the new drug with time in three trials in which the conclusion varied (7).

It is difficult to view the average NSAID trial report as much else than an advertisement for the new drug.

READER BIAS

A review of others' work is to some extent subjective and I might have been preconceived towards finding bias in favour of the new drug. Therefore, I searched especially carefully for bias in favour of the control drug but the yield was poor, and the amount and varieties of bias in favour of the new drug was so enormous that reader bias could hardly be the only cause (6). Blinding of the drugs before reading the NSAID reports would probably have been futile, since it would anyway have been easy to guess which drug was new (5).

I hope I have presented my results in a way that enables the reader to draw his own conclusions.

GENERALIZATION OF THE FINDINGS

"Rheumatology offers medicine a unique opportunity for study of the methodology of clinical therapeutic trials" (131).

Do NSAID trials reflect the state of the art for randomized clinical trials, or is it a disaster area of clinical research? Unfortunately, there are several indications that the findings may be generalized, at least to some extent. Although surveys of trials in other areas have mainly been descriptive, similar results have been reported, eg a preponderance of positive outcomes and invalid conclusions. However, the mechanisms producing a biased literature need not be the same. In cancer trials, for instance, the lack of formal stopping rules with adjustment of the significance level may be important (46).

Extensive analytical reviews of other research areas should be performed to see whether the findings from the NSAID trials may be confirmed; to detect special biases, typical for the area in question; and to diminish the risk of biased meta-analyses.

REFERENCES

1. Andersen B, Holm P. Problems with p. Significance testing in medical research. Basle: Roche, 1984.
2. Fibiger J. Om serumbehandling af difteri. Hospitalstidende 1898; 6: 309-25.
3. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. Br Med J 1948; 2: 769-82.
4. Glass GV. Primary, secondary, and meta-analysis of research. Educ Res 1976; 5: 3-8.
5. Gøtzsche PC. Reference bias in reports of drug trials. Br Med J 1987; 295: 654-6.
6. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal, antiinflammatory drugs in rheumatoid arthritis. Controlled Clin Trials 1989; 10: 31-56.
7. Gøtzsche PC. Multiple publication of reports of drug trials. Eur J Clin Pharmacol 1989; 36: 429-32.
8. Gøtzsche PC. Patients' preference in indomethacin trials: an overview. Lancet 1989; i: 88-91.
9. Gøtzsche PC. Meta-analysis of grip strength: most common, but superfluous variable in comparative NSAID trials. Dan Med Bull 1989; 36: 493-5.
10. Gøtzsche PC. Review of dose-response studies of NSAIDs in rheumatoid arthritis. Dan Med Bull 1989; 36: 395-9.
11. Hampton JR, Julian DG. Role of the pharmaceutical industry in major clinical trials. Lancet 1987; ii: 1258-9.
12. Greenwald AG. Consequences of prejudice against the null hypothesis. Psychol Bull 1975; 82: 1-20.
13. Simes RJ. Publication bias. The case for an international registry of clinical trials. J Clin Oncol 1986; 4: 1529-41.
14. Glass GV, McGaw B, Smith ML. Meta-analysis in social research. London: Sage, 1981.
15. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H. Publication bias and clinical trials. Controlled Clin Trials 1987; 8: 343-53.
16. Anonymous. Manuscript guideline. Diabetologia 1984; 25: 6A.
17. Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. J Amer Stat Ass 1959; 42: 30-4.
18. Schoolman HM, Becktel JM, Best WR, Johnson AF. Statistics in medical research: Principles versus practices. J Lab Clin Med 1968; 71: 357-67.
19. Juhl E, Christensen E, Tygstrup N. The epidemiology of the gastrointestinal randomized clinical trial. N Engl J Med 1977; 296: 20-2.
20. Begg CB, Pocock SJ, Freedman L, Zelen M. State of art in comparative cancer clinical trials. Cancer 1987; 60: 2811-5.
21. Bland JM, Jones DR, Bennett S, Cook DG, Haines AP, Macfarlane AJ. Is the clinical trial evidence about new drugs statistically adequate? Br J Clin Pharmacol 1985; 19: 155-60.
22. Gilbert JP, McPeek B, Mosteller F. Progress in surgery and anesthesia: benefits and risks of innovative therapy. In: Bunker JP, ed. Costs, risks and benefits of surgery. New York: Oxford University Press, 1977: 124-69.
23. Cooperating Clinics Committee of the American Rheumatism Association. A seven-day variability study of 499 patients with peripheral rheumatoid arthritis. Arthritis Rheum 1965; 8: 302-34.
24. Fries JF, Britton MC. Some problems in the interpretation of clinical trials: longterm parallel study of fenoprofen in rheumatoid arthritis. J Rheumatol 1976; suppl 2: 61-6.
25. Ritchie DM, Boyle JA, McInnes JM et al. Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis. Q J Med 1968; 147: 393-406.
26. Mason RM, Barnardo DE, Fox WR, Weatherall M. Assessment of drugs in out-patients with rheumatoid arthritis. Evaluation of methods and a comparison of mefenamic and flufenamic acids with phenylbutazone and aspirin. Ann Rheum Dis 1967; 26: 373-88.
27. Boardmann PL, Hart FD. Clinical measurement of the anti-inflammatory effects of salicylates in rheumatoid arthritis. Br Med J 1967; 4: 264-8.
28. Lee P, Baxter A, Dick WC, Webb J. An assessment of grip strength measurement in rheumatoid arthritis. Scand J Rheumatol 1974; 3: 17-23.
29. Webb J, Downie W, Dick WC, Lee P. Evaluation of digital joint circumference measurements in rheumatoid arthritis. Scand J Rheumatol 1973; 2: 127-31.
30. Hansen TM, Keiding S, Lauritzen SL, Manthorpe R, Sørensen SF, Wiik A. Clinical assessment of disease activity in rheumatoid arthritis. Scand J Rheumatol 1979; 8: 101-5.
31. Ingpen ML. The quantitative measurement of joint changes in rheumatoid arthritis. Ann Phys Med 1968; 9: 322-7.
32. Wright V. Some observations on diurnal variation of grip. Cli Sci 1959; 18: 17-23.
33. Harkness JAL, Richter MB, Panayi GS et al. Circadian variation in disease activity in rheumatoid arthritis. Br Med J 1982; 284: 551-4.
34. Heyman ER. Variability of proximal interphalangeal joint size measurements in normal adults. Arthritis Rheum 1974; 17: 79-84.
35. Sasaki S. Clinical trials of ibuprofen in Japan. Report from the Drug Evaluation Committee, the official organ of the Japan Rheumatism Association. Rheumatol Phys Med 1970; suppl 10: 32-9.
36. Scott DL, Roden S, Marshall T, Kendall MJ. Variations in response to non-steroidal, anti-inflammatory drugs. Br J Clin Pharmacol 1982; 14: 691-4.

37. Light RJ, Pillemer DB. Summing up. The science of reviewing research. Cambridge: Harvard University Press, 1984: 63.
38. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. N Engl J Med 1978; 299: 690-4.
39. Reed JF, Slaichert W. Statistical proof in inconclusive negative trials. Arch Intern Med 1981; 141: 1307-10.
40. Juhl E, Christensen E, Tygstrup N. Hvordan tolkes negative kontrollerede kliniske undersøgelser? Ugeskr Læger 1982; 144: 2913-8.
41. Gram L, Bentsen KD, Parnas J, Flachs H. Controlled trials in epilepsy: a review. Epilepsia 1982; 23: 491-519.
42. Hemminki E. Quality of reports of clinical trials submitted by the drug industry to the Finnish and Swedish control authorities. Eur J Clin Pharmacol 1981; 19: 157-65.
43. Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials. Agenda for the Editor. Controlled Clin Trials 1980; 1: 37-58.
44. Meinert CL, Tonascia S, Higgins K. Content of reports on clinical trials: a critical review. Controlled Clin Trials 1984; 5: 328-47.
45. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 1987; 317: 426-32.
46. Pocock SJ. Size of cancer clinical trials and stopping rules. Br J Cancer 1978; 38: 757-66.
47. Peto R, Pike MC, Armitage P et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976; 34: 585-612.
48. Meinert CL. Toward more definitive clinical trials. Controlled Clin Trials 1980; 1: 249-61.
49. Altman DG. Statistics and ethics in medical research. Misuse of statistics is unethical. Br Med J 1980; 281: 1182-4.
50. Ambroz A, Chalmers TC, Smith H, Schroeder B, Freiman JA, Shareck EP. Deficiencies of randomized control trials. Clin Res 1978; 26: 280A.
51. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. N Engl J Med 1982; 306: 1332-7.
52. Emerson JD, McPeek B, Mosteller F. Reporting clinical trials in general surgical journals. Surgery 1984; 95: 572-9.
53. Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. Br Med J 1986; 292: 810-2.
54. International steering committee of medical journal editors. Multiple publication. Br Med J 1984; 288: 52.
55. Puder KS, Morgan JP. Persuading by citation: An analysis of the references of fifty-three published reports of phenylpropanolamine's clinical toxicity. Clin Pharmacol Ther 1987; 42: 1-9.
56. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983; 309: 1358-61.
57. Gøtzsche PC. Enalapril, atenolol, and hydrochlorothiazide in hypertension. Lancet 1986; ii: 38-9.
58. Karlowski TR, Chalmers TC, Frenkel LD, Kapikian AZ, Lewis TL, Lynch JM. Ascorbic acid for the common cold. A prophylactic and therapeutic trial. JAMA 1975; 231: 1038-42.
59. Huskisson EC, Scott J. How blind is double blind? And does it matter? Br J Clin Pharmacol 1976; 3: 331-2.
60. Byington RP, Curb JD, Mattson ME. Assessment of double-blindness at the conclusion of the beta-blocker heart attack trial. JAMA 1985; 253: 1733-6.
61. Gøtzsche PC. Ditiocarb in HIV infection. Lancet 1988; ii: 1024.
62. Gilman AG, Goodman LS, Gilman A. The pharmacological basis of therapeutics. New York: Macmillan, 1980.
63. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Br Med J 1988; 296: 401-5.
64. Hunninghake DB, Darby CA, Probstfield JL. Recruitment experience in clinical trials: literature summary and annotated bibliography. Controlled Clin Trials 1987; 8: 6S-30S.
65. Gøtzsche PC, Hvidberg EF, Juul P. Rational choice of dose: insufficient background knowledge? Ration Drug Ther 1986; 20: 1-7.
66. Gøtzsche PC, Andersen B, Karlsen FØ. Dosis-respons bias. Nord Med 1986; 101: 24-5.
67. Huskisson EC, Wojtulewski JA. Measurement of side effects of drugs. Br Med J 1974; 2: 698-9.
68. Hart FD, Boardman PL. Indomethacin and phenylbutazone: a comparison. Br Med J 1965; 2: 1281-4.
69. Boardman PL, Hart FD. Side-effects of indomethacin. Ann Rheum Dis 1967; 26: 127-32.
70. Multz CV, Bernhard GC, Blechman WC, Zane S, Restifo RA, Varady JC. A comparison of intermediate-dose aspirin and placebo in rheumatoid arthritis. Clin Pharmacol Ther 1974; 15: 310-5.
71. Lee P, Sturrock RD, Kennedy A, Dick WC. The evaluation of antirheumatic drugs. Curr Med Res Opin 1973; 1: 427-43.
72. Pocock SJ. Clinical trials. A practical approach. Chichester: Wiley, 1983.
73. Meinert CL. Clinical trials. Design, conduct, and analysis. New York: Oxford University Press, 1986.
74. Kirwan JR, Chaput de Saintonge DM, Joyce CRB, Currey LF. Clinical judgement analysis – practical application in rheumatoid arthritis. Br J Rheumatol 1983; 22 suppl: 18-23.
75. Hart FD, Huskisson EC. Measurement in rheumatoid arthritis. Lancet 1972; i: 28-30.
76. Deodhar SD, Dick WC, Hodgkinson R, Buchanan WW. Measurement of clinical response to antiinflammatory drug therapy in rheumatoid arthritis. Q J Med 1973; 42: 387-401.
77. Buchanan WW. Assessment of joint tenderness, grip strength, digital joint circumference and morning stiffness in rheumatoid arthritis. J Rheumatol 1982; 9: 763-6.
78. Bombardier C, Tugwell P, Sinclair A, Dok C, Anderson G, Buchanan WW. Preference for endpoint measures in clinical trials: results of structured workshops. J Rheumatol 1982; 9: 798-801.
79. Huskisson EC, Sturrock RD, Tugwell P. Measurement of patient outcome. Br J Rheumatol 1983; 22 suppl: 86-9.
80. McGuire RJ, Wright V. Statistical approach to indices of disease in rheumatoid arthritis. With reference to a trial of indomethacin. Ann Rheum Dis 1971; 30: 574-80.
81. Lee P, Dick WC. The assessment of disease activity and drug evaluation in rheumatoid arthritis. In: Buchanan WW, Dick WC, eds. Recent advances in rheumatology, part 2. London: Churchill Livingstone, 1976: 1-32.
82. Mainland D. The estimation of inflammatory activity in rheumatoid arthritis. Role of composite indices. Arthritis Rheum 1967; 10: 71-7.
83. Bellamy N. The clinical evaluation of osteoarthritis in the elderly. Clinics in Rheum Dis 1986; 12: 131-53.
84. Fjellström K-E, Goldberg L. Phenylbutazone in active periods of rheumatoid arthritis. Acta Med Scand 1957; suppl 320: 1-49.
85. Smyth CJ, Clark GM. Phenylbutazone in rheumatoid arthritis. J Chron Dis 1957; 5: 734-50.
86. Empire Rheumatism Council. A comparison of prednisolone with aspirin or other analgesics in the treatment of rheumatoid arthritis. Ann Rheum Dis 1959; 18: 173-88.
87. ARA Cooperating Clinics Committee. Aspirin in rheumatoid arthritis, a seven day, double-blind trial – preliminary report. Bull Rheum Dis 1965; 16: 388-91.
88. Donelly P, Lloyd K, Campbell H. Indomethacin in rheumatoid arthritis: an evaluation of its anti-inflammatory and side effects. Br Med J 1967; 1: 69-75.
89. Schwartz D, Flamant R, Lellouch J. Clinical trials. London: Academic, 1980.
90. Ridolfo AS, Mikulaschek WM, Gruber CM, Scholtz NE. Screening rapidly acting anti-inflammatory agents in patients with rheumatoid arthritis. Am J Med Sci 1973; 265: 375-9.
91. Tugwell P, Bombardier C. A methodologic framework for developing and selecting endpoints in clinical trials. J Rheumatol 1982; 9: 758-62.
92. Cooperating Clinics Committee of the American Rheumatism Association. A three-month trial of indomethacin in rheumatoid arthritis, with special reference to analysis and inference. Clin Pharmacol Ther 1967; 8: 11-37.
93. Hemminki E. Quality of clinical trials – a concern of three decades. Meth Inform Med 1982; 21: 81-5.
94. Badgley RF. An assessment of research methods in 103 scientific articles from two Canadian medical journals. Canad Med Assoc J 1961; 85: 246-50.
95. Schor S, Karten I. Statistical evaluation of medical journal manuscripts. JAMA 1966; 195: 145-150.
96. Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. Br Med J 1977; 1: 85-7.
97. Christensen E, Juhl E, Tygstrup N. Treatment of duodenal ulcer. Randomized clinical trials of a decade (1964 to 1974). Gastroenterology 1977; 73: 1170-8.
98. White SJ. Statistical errors in papers in the British Journal of Psychiatry. Br J Psychiatry 1979; 135: 336-42.
99. Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation 1980; 1: 1-7.
100. Evans M, Pollock AV. Trials on trial. A review of trials of antibiotic prophylaxis. Arch Surg 1984; 119: 109-13.
101. Avram MJ, Shanks CA, Dykes MHM, Ronai AK, Stiers WM. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg 1985; 64: 607-11.
102. MacKenzie CR, Charlson ME. Standards for the use of ordinal scales in clinical trials. Br Med J 1986; 292: 40-3.
103. Rimm AA, Mattingly RF. Editorial comment. Obstet Gynecol 1983; 62: 103-4.
104. Blum AL, Chalmers TC, Deutsch E et al. Differing attitudes of indus-

try and academia towards controlled clinical trials. Eur J Clin Invest 1986; 16: 455-60.
105. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. Br Med J 1983; 286: 1489-93.
106. Maxwell C. Some common errors in clinical trials. Clin Trials J 1969; May: 115-9.
107. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. Boston: John Wright, PSG Inc., 1982.
108. Moses LE, Emerson JD, Hosseini H. Analyzing data from ordered categories. N Engl J Med 1984; 311: 442-8.
109. Lavori PW, Louis TA, Bailar JC III, Polansky M. Designs for experiments – parallel comparisons of treatment. N Engl J Med 1983; 309: 1291-8.
110. Feinstein AR. An additional basic science for clinical medicine: III. The challenges of comparison and measurement. Ann Intern Med 1983; 99: 705-12.
111. Cupples LA, Heeren T, Schatzkin A, Colton T. Multiple testing of hypotheses in comparing two groups. Ann Intern Med 1984; 100: 122-9.
112. Egger MJ, Coleman ML, Ward JR, Reading JC, Williams HJ. Uses and abuses of analysis of covariance in clinical trials. Controlled Clin Trials 1985; 6: 12-24.
113. Chalmers TC, Smith H, Blackburn B et al. A method for assessing the quality of a randomized control trial. Controlled Clin Trials 1981; 2: 31-49.
114. Pocock SJ. Statistical aspects of clinical trial design. The Statistician 1982; 31: 1-18.
115. Jonge H de. Deficiencies in clinical reports for registration of drugs. Stat Med 1983; 2: 155-66.
116. May GS, DeMets DL, Friedman LM, Furberg C, Passamani E. The randomized clinical trial: bias in analysis. Circulation 1981; 4: 669-73.
117. Sackett DL, Gent M. Controversy in counting and contributing events in clinical trials. N Engl J Med 1979; 301: 1410-2.
118. Bhaskar R, Reitman D, Sacks H, Smith H, Chalmers TC. Loss of patients in clinical trials that measure long-term survival following myocardial infarction. Controlled Clin Trials 1986; 7: 134-48.
119. Feinstein AR. An additional basic science for clinical medicine: II. The limitations of randomized trials. Ann Intern Med 1983; 99: 544-50.
120. Dixon JS, Smith A, Evans SJW. Reporting clinical trials. Br J Rheumatol 1983; 22 suppl: 74-8.
121. Armitage P. Controversies and achievements in clinical trials. Controlled Clin Trials 1984; 5: 67-72.
122. Kirwan JR, Chaput de Saintonge DM, Joyce CRB, Currey HLF. Conclusions: a review of the symposium/workshop. Br J Rheumatol 1983; 22 suppl: 95-7.
123. Mainland D. Elementary medical statistics. Philadelphia: Saunders, 1963.
124. O'Brien WM. Indomethacin: a survey of clinical trials. Clin Pharmacol Ther 1968; 9: 94-107.
125. Galton F. Regression towards mediocrity in hereditary stature. J Anthropol Inst Great Britain Ireland 1885-6; 15: 246-63.
126. McDonald CJ, Mazzuca SA. How much of the placebo effect is really statistical regression? Stat Med 1983; 2: 417-27.
127. Sokal RR, Rohlf FJ. Biometry. 2nd ed. New York: Freeman, 1981.
128. Godfrey K. Comparing the means of several groups. N Engl J Med 1985; 313: 1450-6.
129. Gøtzsche PC, Forrest M, Hermann GG, Jørgensen PE, Andersen B. Videnskabelige artikler i Ugeskrift for Læger gennem 25 år. Analyse og tolkning af forsøgsresultater. Ugeskr Læger 1989; 151: 220-2.
130. Andersen T, Juhl E, Quaade F. Jejunoileal bypass for obesity – what can we learn from a literature study? Am J Clin Nutr 1980; 33: 440-5.
131. Buchanan WW, Smythe HA. Can clinicians and statisticians be friends? J Rheumatol 1982; 9: 653-4.

Gallstones

An epidemiological investigation

Torben Jørgensen

This review has been accepted as a thesis together with 8 previously published papers, by the University of Copenhagen, February 2, 1990, and defended on May 17, 1990.

Glostrup Population Studies.

Official opponents: Povl Riis, Professor, M.D., and Olaf Bonnevie, M.D.

Reprints: Torben Jørgensen, Carinaparken 29, DK-3460 Birkerød.

INTRODUCTION

Despite gallstones have been known from ancient time (*Beal* 1984), several uncertainties concerning frequency, clinical importance, and distribution of the condition remain.

When the present study started in 1982, the *prevalence of gallstones* in living populations had been assessed by screening procedures only in a Welsh community (*Bainton et al* 1976), in two American Indian tribes (*Sampliner et al* 1970, *Williams et al* 1977), and in a small rural Caucasian settlement in Canada (*Williams & Johnston* 1980). Mostly, gallstone prevalence was estimated in living populations simply by asking subjects whether they ever had suffered from gallstones (*Friedman et al* 1966), hereby identifying less than half of the total number of cases (*Bainton et al* 1976), or in autopsy studies (*Brett & Barker* 1976, *Lindström* 1977), where the prevalence could be overestimated as compared with a living population due to the association between gallstones and coronary heart disease (*Bergman et al* 1968) and cancer (*Lowenfels* 1980).

Symptoms and gallstones. The presence of upper abdominal symptoms, normally including pain, is the most common reason for treatment of gallstones; but which abdominal symptoms are specific to gallstones? Most of the literature describing abdominal symptoms in patients with gallstones does so without using a control group (*Lund* 1960, *Bouchier et al* 1968, *Gunn & Keddie* 1972). Screening studies have shown no substantial association between the presence of gallstones and upper abdominal pain or discomfort (*Price* 1963, *Bainton et al* 1976), but not much emphasis was given to a description of pain characteristics.

Factors associated with gallstone prevalence. Most studies on this subject are based on materials of clinical gallstones (mainly cholecystectomised subjects), which make up less than half of the total gallstone prevalence. It is questionable whether these patients with clinical gallstones are representative of the total gallstone population or whether a selection takes place. The iatrotropic stimulus may differ widely in subjects, and doctors may have different attitudes to examine subjects for gallstones, maybe depending on the presence of characteristics claimed to be associated with gallstones (*van der Linden* 1961). Already in 1946, *Berkson* warned against the use of hospital materials for epidemiological research, as the relative frequency of diseases in a group of patients who had entered the hospital differed from that of the whole population served by the hospital. This theoretical objection, called Berkson's fallacy, was later demonstrated empirically (*Roberts et al* 1978). It, therefore, seems reasonable to re-evaluate what factors are associated with gallstones, looking at gallstone prevalence in random populations.

In this thesis, results from the cross-sectional part of a planned cohort study of gallstones (I) are related to the problems outlined.

MATERIAL AND METHOD

2.1 SAMPLING AND INVITATION

The cohort of this study was an age- and sex-stratified random sample, comprising 4,807 men and women, born in 1922, 1932, 1942, and 1952. The cohort was drawn from the National Central