THE UNIVERSITY OF CALGARY

Publication Bias in Gastroenterological Research

by

Antje Timmer

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

JANUARY, 1999

Abstract

To assess whether publication bias is present in gastroenterological research, determinants of publication were studied in abstracts submitted to the American Gastroenterological Association. A random sample of 326 clinical trials (CCT), 455 abstracts on other clinical research (OCR) and 219 basic science reports (BSS) was evaluated. Publication rates were examined by a literature search and by a mail survey of abstract authors. Survey information was available on 499 abstracts. The overall estimated publication rate was 57%. A higher number of previous publications and multi-center status were associated with higher publication rates. Significant differences in publication of negative study results were identified only after correction for under-ascertainment and response bias. Statistically significant studies were published faster and in higher impact journals. The most frequent reason cited for non-publication was lack of time.

Acknowledgements

A number of people made important contributions to this project. First I would like to thank my supervisor, Dr. L. Sutherland, for making me aware of this interesting topic, and for his support in pursuing it.

Dr. R. Hilsden and Dr. C. Macarthur had substantial input into the development and validation of the quality scoring instrument which had grown into a little study by itself. The students in the Department of Community Health Sciences and in the GI-research group, and a number of staff members also contributed helpful advice in this matter.

Dr. P. Brasher gave advice on the statistical analysis. For help and advice in the data base searches I would like to thank Dr. John Cole of the Medical Library of the University of Calgary and Dr. Marlene Dorgan of the University of Alberta in Edmonton. Dr. D. Hailey was a further helpful member of my supervisory committee. Dr. K. Dickersin contributed substantially to the final version of this thesis by providing detailed and thoughtful comments on the manuscript.

I am particularly grateful to have had summer students helping me with the less gratifying aspects of this study: Rasheena Moorthy screened, cut and photocopied most of the abstracts and searched addresses for the mail survey. Diana Czechowski downloaded thousands of references from MEDLINE and Embase. Both students also helped with data entry.

For generous financial support I am obliged to the German Academic Exchange Service (DAAD), which had granted me a two year fellowship. In this context, I am grateful to KH Joeckel and H Goebell for supporting my application. The study was supported by grants from the Calgary Regional Health Authority and by Searle Inc. Canada.

This thesis is dedicated to Prof. Harald Goebell of Essen, Germany, who has had a major influence on my career by invoking my interest in epidemiology and gastroenterology.

Table of Contents

List of tables

List of figures

List of abbreviations

in alphabetical order:

AGA	American Gastroenterological Association
BSS	basic science study
CI	confidence interval
CCT	controlled clinical trial
DDW	Digestive Disease Week (annual meeting of the AGA)
epid.	epidemiology
GI	gastroenterology/gastrointestinal diseases
HP	Helicobacter pylori
HR	hazard ratio
ICCC	intra class correlation coefficient
IF	impact factor
IQR	interquartile range
N/W	North, West and Northwest
OCR	other clinical research
OR	odds ratio
phys	physiology
PI	principal investigator, senior investigator
S/E	South, East and Southeast
SE	standard error
T&D	therapy & diagnostics

I. Introduction

> Publication bias is the tendency on the parts of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings.
> Dickersin, 1990 (1)

Publication bias may arise from anywhere within the research and publication process - completion of a study, preparation and submission of a manuscript ("submission selection bias") and acceptance or rejection of the manuscript ("editorial selection bias") (2). A subjective component of publication bias is reflected in the tone of a presentation and in the interpretation of the evidence (presentation in "*advocacy style*" (3)). Citation bias, in the sense of selective citation of studies with positive results, may be viewed as another facet of the same problem (4,5).

Publication bias was first studied and acknowledged as a problem in the fields of education and behavioral sciences, as described in the historical review of the phenomenon by Kay Dickersin (1). A study by Sterling, published in 1959, found that 97% of all applicable articles in major psychology journals presented statistically positive results (6). This study was later updated and extended to three medical journals (NEJM, Am J Epidemiol, Am J Public Health) (7). It appeared that the problem had changed little over a thirty year period in the psychology journals. The medical journals had an only slightly lower percentage of positive outcome studies (85%). Another interesting early report from the behavioral sciences is an experimental study by Mahoney who presented 75 reviewers with manuscripts that were identical in introduction and methodology, but varied in the direction of the results and discussion sections. Inter-rater agreement was shown to be very low in this study. In addition, manuscripts with negative results were awarded significantly lower evaluations than those with positive results, and were

on average suggested for rejection, while studies with positive outcomes were usually accepted.

In the biomedical literature, publication bias has primarily been studied in the context of meta-analysis. Meta-analysis is the *process of using statistical methods to combine the results of different studies* (8), a technique used primarily in the context of randomized controlled clinical trials, but also in studies on causative effects (9,10). Commonly applied statistical methods for the synthesis of data are random or mixed effects models, such as the DerSimonian and Laird method (11). Meta-analyses based on these models aim to make inferences to a hypothetical population of studies, assuming that the studies at hand constitute a representative sample of this hypothetical underlying population. Publication bias thus forms a major threat to the validity of meta-analysis as it distorts the estimated effect when only published studies are included. Inclusion of unpublished material, on the other hand, may be problematic not only because of the considerable efforts required to identify these, but also, because lack of peer review may results in these studies being less rigorous in relation to scientific quality and validity (12).

In practical terms, publication bias leads to an inflation of the false positive rate of studies in the literature. This should be a concern in all fields of research. In meta-analysis of clinical trials, variation of effect sizes will be underestimated, and mean effects will be exaggerated (2). In epidemiological research, selective reporting may result in erroneous risk perceptions (13,14). An example of suspected publication bias with important public health implications concerns the relationship between smoking and lung cancer (15).

The roots of publication bias are complex. Dickersin cites the *vanity of thinking* (Hall, 1965), translating into *the notion of going to press only if one has something "big" to present*", and the 1621 statement of Francis Bacon that *the human*

intellect ...is more moved by affirmatives than by negatives (1). Mahoney equated publication bias with *confirmatory bias*, referring to *the tendency for humans to seek out, attend to, and sometimes embellish experiences which support or "confirm" their beliefs (16)*. Similarly, Wynder called publication bias a *most compelling* wish bias *adversely affecting good science,* defining *wish bias* as *the tendency on the part of the [...] investigator to reach a desired result (17)*. To this possibly *inherent part of human nature* (17), T.C. Chalmers less leniently adds a few more human traits that may be important in the creation of publication bias: ignorance, sloth and greed (18).

From the statistical view point, publication bias is rooted in the over-emphasis on p-values and significance testing as measures of evidence which have become traditional in most fields of research. A detailed review by Goodman discusses the historical debate surrounding the appropriateness of these measures for this purpose (19). The use of confidence intervals alleviates some of the problems of p-values by shifting the focus to effect estimation and by facilitating the interpretation of results in a wider context (20). Factually, however, although conveying more quantitative information, confidence intervals are usually used for significance testing in a similar way as p-values, thus do not prevent publication bias.

A third factor furthering the occurrence of publication bias is a high prevalence of inappropriately powered studies (2). Publication bias is directly related to sample size as the variability of study effects increases with decreasing power. Newcombe relates the problem of often insufficient sample sizes to the way research is organized and rewarded: Where individually published research output is the standard for promotion, wish bias will tend to be big and sample sizes will tend to be small (2,21).

Several statistical and graphical methods have been devised to retrospectively estimate the extent of publication bias in meta-analysis. These include sensitivity

analyses, by which effect sizes are compared between meta-analysis in- or excluding unpublished trials, comparable to the approach used by Simes in his review of ovarian cancer trials (22). Funnel plots take advantage of the association between sample size and publication bias: Effect estimates are plotted against a measure related to the variance of a study such as sample size or standard error (23). In the presence of publication bias a gap is expected in the area of non significant estimates in smaller studies within an otherwise funnel shaped scatter.

Publication bias has thus been demonstrated and appreciated as a problem in those areas of research where meta-analyses are commonly performed. I. Chalmers estimates that selective underreporting of research is probably more widespread and more likely to have adverse consequences for patients than the deliberate publication of falsified data, as for the latter there is the accepted protective mechanism of data replication (24). However, in other areas of research, the relevance of the problem is debated. While some consider the phenomenon inevitable in clinical trials (25), others suspect it to be primarily a perception of disappointed authors or *a problem made up by professional meta-analysts* (26,27). Occasionally, bias is even encouraged: *negative results have never made riveting reading* (BMJ editorial note, 1983 (28)), and *publication bias, like citation bias, is only part of subjective judgment that we should gladly enjoy to make meaningful contributions to medicine* (29).

One of the early descriptions of publication bias in the biomedical literature stems from the area of gastroenterology (30). Although publications of meta-analyses have been increasing exponentially over the last few years in this field (31,32), publication bias has not been systematically examined. It was our aim to study publication bias in clinical trials as well as in other types of research in gastroenterology. The following literature review will focus on possible research designs and on other possible predictors of publication bias in order to provide the rationale and background for the design of this study.

II. Literature review

> More has been written to complain about publication bias than to report results of studies undertaken to evaluate it.
> Dickersin, 1990 (1)

A. Assessment of publication bias - methodological approaches

1. *A posteriori* assessment of publication bias

Most reports on publication bias concern its impact on meta-analysis, usually in the context of a specific intervention or association. In addition to the previously mentioned strategies of funnel plotting or sensitivity analysis, several articles deal with methods to estimate or correct for the effect of publication bias in meta-analysis (33-35). Modeling may serve as a substitute for the inclusion of unpublished material. In this case, no information will be available on the characteristics and number of unpublished studies. Thus, while these methods may be helpful in the interpretation of a specific meta-analysis, no insight is gained into the etiology and nature of publication bias.

2. Cohort based studies - identification before publication

Cohort based studies aim at identifying studies at an early stage before publication and examine associations with publication by a comparison of relevant characteristics based on outcome. Examples include Chalmer's follow-up of summary reports to a register of clinical trials in perinatology (36), or the survey on trials registered with the International Cancer Research Data Bank by Simes (22). The latter study is a special case, as it was conceptualized as a strategy for complementing a specific meta-analysis. The cohort was restricted to trials of ovarian cancer, resulting in a very small sample size (n = 29). The impact of unpublished trials on the effect size was evaluated; assessment of predictors for publication was secondary, and hampered by the lack of power.

Beside clinical trial registries, ethics committee boards have been used to identify pre-publication cohorts of research projects, offering the advantage of a wider variety of research types. This source was used by Easterbrook et al. and by Dickersin et al. in their studies on determinants of publication (37,38). (These studies were later combined in a meta-analysis by Dickersin (39)). Similarly, an Australian study by Stern and Simes was based on projects submitted to a hospital review board (40). Taking advantage of these local review boards is likely to result in good follow up rates as compared to the more anonymous and geographically widespread national or international registries. On the other hand, the generalizability of the results may be compromised.

The most commonly used way to collect a cohort of pre-publication reports is based on abstracts submitted to scientific meetings. However, the majority of these studies following up meeting abstracts did not examine publication bias but restricted the analysis to the assessment of publication rates and time to publication (discussed below). A valuable exception is the study from Toronto/St. Jerome on the fate of abstracts submitted to a cancer meeting (41). A random sample of 197 abstracts of the 1984 ASCO meeting (American Society of Clinical Oncology) was studied for determinants of acceptance for presentation as well as for subsequent publication. In addition, content of articles was compared to the original abstracts, and citation practices were recorded (results will be discussed below).

Basing the cohort on meeting abstracts has some drawbacks, in particular the limited amount of information available from abstracts as compared to research proposals submitted to institutional review boards. In addition, it is expected that some submission bias may already be in play before abstracts are submitted to a meeting. This bias is aggravated, when only abstracts selected for presentation are available. On the other hand, medical meeting abstracts are usually easy to access. Major meetings are likely to attract reports fairly representative of the re-

search quality, types and topics in the respective field, increasing the generalizability within the area.

3. Cohort based studies - assessment of outcome

Techniques for assessing outcome in the cohort studies include information from clinical trial data bases (I. Chalmers et al. (36)), MEDLINE or other data base searches (41,42). Easterbrook et al., Simes, de Bellefeulle et al., Stern and Simes, and Dickersin et al. collected data from investigators on the publication status of research and reasons for non publication by mailed questionnaires (22,40,41) or personal interviews (38,43). A sufficiently complete investigator survey could be considered as a gold standard, as data base searches have been shown to be not sufficiently sensitive to the identification of published trials. A 1985 report found that MEDLINE searches identified only 30% to 60% of published trials as compared to the Oxford Database of Perinatal Trials (44). However, since then, indexing in MEDLINE and other data bases has undergone improvements. Also, the inclusion of complementary databases or of databases more specific to the focus of the respective study cohort may improve completeness of retrieval (45). In the 1992 follow up study by de Bellefeuille et al. on 197 cancer meeting abstracts, 12 additional trials were identified by contacting authors of apparently non-published abstracts. Added to 103 articles identified by Cancerline, the publication rate increased from 52% to 58%. The response rate in this survey was 55%. It was estimated that the Cancerline search identified 90% of all published studies. However, this estimation was based on the assumption that non responders did not have unidentified publications.

The survey of authors of abstracts both with and without data-base-identified subsequent publications is particularly important considering the notoriously poor response rates in mail surveys involving physicians as compared to general population surveys (46). The reference list of an article reporting a survey on chiropractors includes ten publications solely dealing with the problem of resistance

and annoyance in mail-surveyed physicians (47). However, most of the cohort studies following up abstract authors or investigators submitting to review boards had surprisingly high response rates, possibly related to the relatively exclusive selection of the respective survey population (Scherer, ophthalmology abstracts: 89%; Simes, ovarian cancer review board submissions, 63%; Stern and Simes, hospital review board submissions, 70%) (22,40,42).

4. Cross-sectional studies

In addition to these cohort studies, two studies were identified that used mail surveys in a cross-sectional approach (48,49). Particularly interesting is the study by Dickersin, who identified unpublished studies by approaching authors of published clinical trials. With respect to the reasons of non-publication on the parts of the investigators, this study represents the main source of information available for the development of this thesis project.

B. Outcome measures

1. Publication rates

> Almost every scientist working today gets his work published, somewhere, once he decides "to write it up"; maybe it will be in the bulletin of the Podunk Country Medical Society rather than in the journal with international prestige or readership.
> Comroe, 1976 (21).

Rates of publication obviously depend on what is chosen as the denominator, e.g. at which stage of research a cohort was assembled, on duration of follow-up or on how the outcome was assessed. I. Chalmers et al. reported publication rates of 58% to 78% (depending on subspecialty) in his perinatal data base study (36). Easterbrook et al. found that 73% of studies submitted to a review board were published as abstracts. Of these, 61% were later published in full (37).

Publication rates following abstract submissions have been assessed in a large number of studies from various specialties, all of which used data base searches for outcome assessment (41,42,50-55) (list not exhaustive). Publication rates ranged from 32% to 77%; the average was around 50%. A meta-analysis by Scherer, including 2391 abstracts from ten follow-up studies, found a weighted mean of 51% (95% CI 45%-57%) (42). Mean follow-up rates for abstracts submitted to the meeting of the American Gastroenterological Association (AGA) were calculated as 49% (56).

2. Time to publication

Time to publication is often cited as the median time to publication. There are some problems with this approach which will be discussed in more detail later. For one, median time obviously depends on the time point chosen as the reference point, on the eventual publication rate, and on the duration of the follow-up. Most studies following up abstracts used a minimum follow up time of two years. In the meta-analysis by Scherer, all studies showed a similar time pattern of publication, reaching a plateau at about 24 to 30 months (42). Thereafter, only a few abstracts were published. Juzych described a somewhat shorter time profile in more detail, observing a lag phase of four months following a meeting, and a plateau from month 17. Median time to publication was 13 months. The annual probabilities for publication, if not published before, were 13% for year three, 5% for years four and five (52). For AGA-abstracts, the median time from presentation to publication has been calculated to be 1.7 years $\pm$ 1 year (follow up period 1991-1996) (56).

Recently, there has been additional discussion over whether delay in publication may be an indicator of publication bias. Stern and Simes found a delay for quantitative studies with negative results (40). Median time from approval by the ethics committee to publication was 4.8 years for studies with significant results, 8 years for studies with null results and "not reached" for studies with indefinite re-

sults. The numbers were similar in the subgroup of clinical trials. The authors conclude that time to publication is another important factor in the examination of publication bias. The finding was recently replicated by Ioannidis who compared time to publication in efficacy trials in AIDS, and found median times from patient enrollment to publication to be 4.3 years if results were positive and 6.4 years, if results were negative (57). Apparently, this difference was largely due to a difference in the time from completion of patient follow-up to publication (1.7 vs. 3 years). Based on these results, Ioannidis coined the term *publication lag* or *time lag bias* (58) which he considered a distinct phenomenon with *complementary implications.*

3. Impact of journal

There is some indication that publication bias may be more prevalent in higher impact journals: Easterbrook et al. found a mean IF of 1.6 for published studies with positive findings, as compared to 0.9 for studies with negative results (43). Simes also noticed that studies with positive findings were more likely to be published in high impact journals (22). On the other hand, in a study comparing articles from emergency medicine journals (two journals, 177 articles) to articles from the NEJM and JAMA (211 articles), no significant difference in the proportion of studies with statistically negative results was found (15% to 16%) (59).

C. Determinants of publication

> The main determinant of what is or is not published therefore seems to be the scientist, for it is he who decides to become or not become an author.
> Comroe, 1976 (21).

1. Editorial vs. submission bias

In studies from the areas of psychology and education, significant editorial bias has been shown, as summarized in Dickersin's review on the history of publica-

tion bias (1,38). In medicine, however, little evidence for editor based bias has been found, even though whole meetings have been devoted to the subject of peer review in biomedical publication (Chicago, 1989 and 1993; proceedings published in JAMA 263 and 272). Studies that evaluated reasons for not publishing in the context of publication bias relied on investigators' judgment regarding reasons for rejection. Generally, rejection of a manuscript seems to be an infrequent reason for failure to publish, accounting for only up to 11% of unpublished projects (43,49).

More important are the factors influencing investigators' motivations to complete and submit research projects. In the Toronto study as well as in Dickersin's meta-analysis, lack of time or insufficient priority were the main reasons quoted by investigators (39,41). The Australian study by Stern and Simes found lack of funding, lack of support by colleagues and problems with patient accrual to be prominent factors, depending on the stage of research at the time of abandonment (100 respondents) (40). Similarly, Dickersin et al. in their cross-sectional study on clinical researchers stratified reasons for non publication by stage of abandonment (49). Of 271 unpublished studies, 36% had been stopped before completion (most frequent reason sample size problems, 27%), and 50% were completed but never submitted (most frequent reason negative results! 34%). Other factors were lack of interest (12%) and poor methodology (4%). Of note - this survey identified a record holder with 34 unpublished projects for a single investigator.

2. Direction and strength of effect studied

This variable is the primary exposure in the study of publication bias and has been discussed in greater detail in the introduction section. Associations of strength of effect as well as of factors requiring more subjective judgment (such as biased interpretation of study results) with publication rates might all be considered effects of publication bias. Factually, however, studies on the phenome-

non were usually restricted to the examination of the direction of the results as a predictor of publication. There is some inconsistency about how to define significance or direction of results. De Bellefeuille et al. divided results into "positive", "negative" and "neutral" based on a statistical significance level of 95% (41). Stern and Simes differentiated "significant ($p < 0.05$)", "showing a non significant trend ($0.05 \leq p < 0.1$)", and "non - significant ($p \geq 0.1$)". They were, however, not entirely consistent, as they also reported results based on a not previously defined category of "indefinite results" (40). Dickersin et al. used a classification including "trends" or tendencies separately for either treatment group in their survey of investigators (49). In the cohort studies following up institutional review board submission, as well as in Scherer's study on ophthalmology abstracts, results were simply divided into "significant" and "not significant" (39,42). Easterbrook et al. combined the p-value oriented classification into "positive", "negative" and "trend" with a rating of the clinical or scientific relevance of the study (37). In contrast, in the study by I. Chalmers et al., the differentiation was based on the authors' conclusions regarding the superiority ("positive study"), or harmfulness ("negative study") of the test treatment, with neutral meaning that there was no real difference (36). These differences have to be kept in mind, when results from different studies are compared. Also, in particular when non RCT's are included, statements on or evidence of statistical significance are not always apparent, and misclassification may arise.

The OR for publication of statistically significant results as compared to negative results was 6.2 (95% CI 2.2 to 16.9) in Dickersin's meta-analysis (39). Also, preference for positive results was found by Stern and Simes (HR 2.3, 95% CI 1.5-3.7) (40), Easterbrook et al. (OR 2.3, 95% CI 1.3-4.3) (37) and De Bellefeuille et al. (74% vs. 32% publication rate, $p < 0.001$) (41). In addition, De Bellefeuille et al. found that positive results were also associated with higher rates of abstract acceptance at a cancer meeting. Scherer, on the other hand, found only a weak association with publication (OR 1.2, 95%CI 0.9-1.6). I.

Chalmers et al., using their different classification (see above), did not discover a significant association (36).

3. Research type and design

The majority of studies on publication bias (and other determinants of publication) are restricted to controlled clinical trials. Data are rare on uncontrolled clinical research, and none are available for basic science. However, there is some evidence that design rigor may be inversely related to the occurrence of publication bias. Easterbrook et al. found publication bias significantly smaller in clinical trials vs. observational studies, in studies with concurrent vs. non concurrent controls, and in randomized vs. non randomized trials (37).

Randomization has been demonstrated to be associated with weaker treatment effects (60). Similarly, Berlin found that a negative association between sample size and strength of effect (indicative of publication bias) was more obvious in non randomized trials as compared to randomized studies (61).

A possible explanation for differences in susceptibility to publication bias depending on research type and design features is that trials involving more commitment of time and resources are less easily discarded, while trials in less controlled settings, such as phase II or other more exploratory trials have more potential for selective reporting. Dickersin, Juzych and Scherer with their respective co-workers, on the other hand, did not find associations between design and publication rates (38,42,52).

4. Methodological quality

> If the characteristics that determine publication are related to study quality then the selection bias incurred by only studying published literature is acceptable, even desirable.
> Dickersin, 1990 (1)

The quality of the underlying research, together with scientific or clinical relevance will ideally be the decisive determinant of research publication. However, assessment of methodological quality is associated with a number of problems, rendering it the variable that is most difficult to account for.

First, methodological quality of research is an ill-defined term and describes a concept rather than a single variable. Like publication bias, it has so far primarily been a concern of meta-analysts who consider it a function of how well bias is controlled by the study design and analysis (62). Factors to be considered include subject selection, methods of treatment allocation, blinding, appropriate sample size estimations, minimization of measurement bias by reproducible definitions of outcome and exposure variables, control of confounding etc. Separate examination of these single components leads to several problems involving multiple testing and multivariate modeling. A number of summary scores have therefore been developed to assess the quality of controlled clinical trials (63), as well as modifications appropriate to epidemiological research, i.e. controlled observational studies of causation (64). The use of summary scores on the other hand is by itself associated with several problems, including insufficient validation, lack of a gold standard, or the pretense of objectivity (quality scores are *perhaps the most insidious form of subjectivity masquerading as objectivity* (65,66).

Second, the methodological quality of research can only be assessed to the extent that information on this is available from its report or via investigator survey. This is particularly problematic when assessment is based on summary reports, as it is controversial as to how well the quality of research is reflected by its short report (67,68). Comparing abstracts to investigator information or full articles has revealed considerable inconsistencies. Ignorance of basic study features is not uncommon when authors of abstracts are surveyed (41,42). Quality of reporting

and quality of research are thus not separable, and have been combined in this research into a variable of "formal report quality".

Third, there is no literature on the assessment of formal quality in basic science research or less controlled clinical studies. Fourth, most quality scores require adaptation to a specific intervention, and cannot be applied simultaneously to diverse study topics (e.g. (62)). Finally, even in controlled clinical trials, no instrument is available to assess formal quality of abstracts, as previous guidelines and checklists for the composition of abstracts were restricted to the assessment of completeness of reporting, neglecting features of the underlying research (69,70).

Consequently, few studies on publication bias included quality as a covariate. An exception is the cohort study by I. Chalmers et al. on perinatal data base summary reports. He used a very simplified version of T.C. Chalmers' well known instrument for full reports of clinical trials, incorporating method of treatment assignment, intent to treat analysis and concealment of allocation (resulting in a measure obviously very specific to RCT's and oblivious to completeness of reporting) (36,62). No association with publication rates was found. Reasons discussed by the authors included insufficient power for modest effects, or insufficient information available from the short reports precluding meaningful assessment of quality.

5. Sample size

> Publication bias is endemic and will remain so as long as the sample sizes commonly used in research are too small and the methods used to assess adequacy of sample size are deficient.
> Newcombe, 1987 (2)

Sample size is directly related to susceptibility to publication bias, on the grounds of statistical reasoning and a variety of associated risk factors, e.g. the easier discarding of smaller trials that did not yield the desired results. Berlin found that there was a strong inverse relation between sample size and treatment effect (61). Most methods for the *posteriori* assessment of publication bias in meta-analysis are based on this association (see above). It has been argued that small studies are *per se* of inferior quality. There is, however, controversy on the justification of rejecting studies solely on the basis of power and sample size (61,71).

In the cohort studies on publication bias, the effect of sample size on publication was not consistent: Dickersin, Easterbrook and Stern and their respective coworkers did not find an association of publication rates with sample size (37-40). Scherer, on the other hand, found an OR of 1.5 (95% CI 1.1-1.9) for studies with above the median sample size. Obviously, there are problems with the comparability of sample size across different interventions and study designs, and other measures such as power or standard errors may be more valuable parameters. Unfortunately, use of these parameters is usually not feasible as they are rarely reported (72).

6. Presentation of abstracts at scientific meetings

The association of abstract presentation with subsequent publication is a consistent finding in all studies that have included this parameter (41,54,55). It is not clear whether this reflects the appropriateness of abstract selection procedures, an effect of motivation through appreciation, or the necessity of at least a preliminary report which is required for presentation. Abstract acceptance is probably not an extraneous risk factor for publication but could be considered an intermediate factor (*a step in the causal chain* between risk factor and outcome (73)). It must therefore not be treated as an independent predictor. Specifically, it should not be included as a possible confounder in a multivariate model.

7. Country of origin, language of publication

The country of origin as a predictor of publication has not been examined, as all follow-up studies on publication bias recruited their cohorts from abstracts to national meetings or from submissions to local or national review boards and registries. Language and origin, however, have been a focus in debates pertaining to study identification for meta-analysis. Language restrictions in systematic reviews are often applied, possibly based on the idea that important and well done studies will be published in English - while the true underlying reason is more likely one of convenience. Two studies were able to demonstrate that inclusion of foreign language publications may have an impact on effect sizes, and that exclusion could not be justified on the grounds of qualitative inferiority (74,75). But even within English language publications national biases may be apparent, as demonstrated in citation analyses (76). A recent study justified national bias by the observation that "certain countries" produced only statistically significant results, suggesting that publication bias could be diminished by treating studies from these countries "with particular caution" (77). This study, however, was poorly designed and neglected findings from a more thoughtful report by Egger, who found that non-English publications reported higher rates of trials with nil results than English language publications of the same authors (78). It is likely that authors tend to submit to national rather than to international journals, for example when statistically negative results lead to the anticipation of editorial rejection. The association of negative study results with language of publication or, inversely, of positive results with origin of authors, may therefore be considered as another indicator of publication bias.

8. Seniority of author, institutional prestige

Prior publication by the investigator was found to be the best predictor of publication in a small study on menstrual cycle research (48). Similarly, a controlled study on the effects of blinding in the peer review process found a significant correlation of the number of previous publications by the senior author with the

score awarded by reviewers, but only in blinded peer review (79). It was concluded that manuscripts by more senior authors were in fact better than those of less experienced authors.

Institutional prestige was graded on the received monetary value of grants funded by the NIH in a study by Garfunkel, and was found to be without influence on publication rates (80). On the other hand, in a follow up of cardiology abstracts, Goldman found publication rates to be associated with the ranking of the submitting institution, ranging from 69% for "Top Five US Medical Schools" to 35% for "foreign or not affiliated with US Medical School" (55).

9. Source of funding

Dickersin et al. found in both their cohort study and meta-analysis, that external funding was a predictor of publication, the adjusted OR being 3.0 (95% CI 1.8 - 4.8) as compared to either no or to internal funding (38,39). This finding was confirmed in the study by Stern and Simes (40). A study by Davidson supports suspicions that industry may withhold unfavorable results by detecting a significantly lower percentage of trials without statistically significant results in publications of research sponsored by industry as compared to those with other sources of funding (81).

10. Number of centers involved

Multi-center status was found to be a significant contributor to prediction of publication in the cohort study by Dickersin et al. (adjusted OR 2.1, 95% CI 1.2-3.6) as compared to single center studies (38). This was independent of sample size.

11. Gender bias

While in editorial processes, gender differences such as faster review times and more rejections by female reviewers and editors as compared to men have been detected (82), no evidence was found for bias based on the sex of investigators.

However, with respect to studies on publication bias, only in the study by Dickersin et al. had this variable been included in the initial model (38).

D. Research rationale

Based on the review of the literature, publication bias can be considered a significant problem in the communication on research. In the field of gastroenterology, the occurrence of publication bias and determinants of publication have so far not been systematically studied. Most previous studies on publication bias have focused primarily on the implications for meta-analysis of controlled clinical trials. However, publication bias is likely to affect any type of research. Because of the study designs used, geographical restrictions and/or focus on certain subspecialties of medicine it is doubtful whether these results may be generalized to other fields of medicine.

The annual meeting of the American Gastroenterological Association is the most important venue for the presentation of research in gastroenterology, attracting researchers and clinicians from around the world. For the 1998 Digestive Diseases Week in New Orleans, more than 6,000 abstracts were submitted - all of which are published in a supplement to *Gastroenterology*. Four thousand presentations included work from 41 countries, covering all aspects of gastroenterological research (83). Approximately half of these abstracts will eventually be published as full articles (56).

Abstracts submitted to this meeting thus form a suitable basis for the collection of a pre-publication cohort of GI-research projects, representative of international research interests. A follow-up study of these abstracts was expected to yield important information on the determinants of publication in gastroenterology, surpassing previous research by adding an international context, and by the parallel study of different types of research. The study was therefore based on a

random sample of abstracts, including all research types represented in the overall population of submitted abstracts. However, as no information on determinants of publication in basic science was available, there were doubts about the general applicability of some of the concepts which had previously been primarily studied in clinical trials. The design took this into account by stratifying for research type, with each stratum having sufficient power by itself for the detection of the association of interest (publication bias). The thesis may therefore be conceptualized as a combination of studies: The section on controlled clinical trials studies represents an analysis of cause and effect relationships which were in hypothesis and design supported by previous publications, but now extended to a new specialty. The section on basic science at the other end of the spectrum may be considered a pilot study, serving to collect information on determinants of publication in this understudied area, possibly sparking further research, while expectations as to the answer to the study question were modest.

III. Study objectives

This thesis was based on the hypothesis that studies showing statistically positive results are more likely to result in full publications than those with negative or indifferent results (publication bias).

The primary objective was therefore to examine a cohort of abstracts submitted to gastroenterological meetings for evidence of publication bias. In addition to crude publication rates, time to publication and prestige of target journals were studied.

To take into account possible confounding or interacting variables, other potential determinants of publication, as described in the literature review were also evaluated. Descriptive analyses examined reasons for non publication, where applicable.

IV. Methods

The study was submitted to the Conjoint Medical Research Ethics Board of the University of Calgary Medical Faculty and the Foothill's Hospital Research and Development committee and received scientific and ethical approval.

A. Study design

The study was designed as a retrospective cohort study based on abstracts submitted to AGA-meetings.

B. Definitions

For the purpose of this study, the following definitions were used:

Publication bias: the selective submission and acceptance of studies based on the direction of the study results, i.e. a bias in favor of studies with statistically significant results.

Direction of the study results: The terms "positive" and "negative" for study results are used in analogy to their use in medicine, where "positive" means affirming, and "negative" means not affirming, the presence of the organism or condition in question (84), or, in this case, of the association or effect under study. It is not of interest whether the diagnosis, organism, effect or association is harmful or beneficial. Statistical significance testing is used to decide whether a study is considered to affirm or to not affirm the effect in question. In other words, affirmation is assumed if the null hypothesis (hypothesis of no effect) is rejected.

Statistical significance: *(85)*

- Positive results, positive outcome: statistical significance was achieved for the main outcome according to the study question, or, in the presence of multiple outcomes, for the majority of outcomes. Statistical significance is assumed when $p < 0.05$, or when the 95% confidence interval excludes the reference

value, or when the authors state that statistical significance was achieved. It is not taken into account whether a statistical difference is in favor or against the test treatment, or whether a statistically significant association is positive or negative.

- Negative results, nil results, negative outcome: statistical significance was not achieved for the main outcome or, if multiple outcomes were used, for the majority of outcomes. A study outcome is assumed negative when $p \geq 0.05$, or when confidence intervals include the reference value, or when the authors state that statistical significance was not achieved. This includes results considered by the authors to show a "trend" or tendency towards the effect or association.
- Equivocal studies, mixed results, unclear results: includes all studies that used analytical set-ups (controlled studies or before-after-comparisons), where results could not be classified as positive or negative according to the definitions used. For example, multiple results are reported, but are of mixed significance and it is not clear what the primary question was. Also, results were classified as equivocal, if test results were not reported and/or if a statement on statistical significance was missing.

Descriptive (non analytical) studies included case reports, case series, incidence/prevalence studies, qualitative research, as well as descriptions of procedures and instrument validations.

Research type:

- Controlled clinical trial (CCT): any prospective study examining the effects of a therapeutic or diagnostic intervention in humans in a parallel controlled or cross-over design
- Other clinical research (OCR): any other study using the intact human (or groups of intact humans) as the unit of analysis. This group is very heterogeneous and was therefore sub-classified as follows:

- *Therapy & Diagnostics*: studies other than CCT that examine effects of treatment or diagnostic procedures in humans. This includes, for example, uncontrolled clinical trials (phase I, II or IV) (86), meta-analyses and cost-effectiveness or cost-benefit analyses.
- *Epidemiology:* any observational study analyzing or describing distributions, health attributes or determinants of health related states or events
- *Human physiology:* studies that examine patho-physiological or biochemical mechanisms in the intact diseased or healthy human. This includes observational studies where description of disease characteristics required invasive or costly lab-based diagnostic procedures such as 24-hour pH evaluation, breath tests or motility recordings.

- Basic Science Studies (BSS): any study, performed in a laboratory setting, where the unit of analysis is not the intact human

Country groups: countries were grouped based on geopolitical regions after considering similarities found when exploring the sampling frame (acceptance rates, distribution of research types). France was considered a part of N/W-Europe. Australia contributed too few abstracts to form a separate group (n = 20), and was therefore combined with N/W Europe.

Response rates in the abstract author survey (46,87):
Two proportions were of interest: response rates of authors, and proportion of abstracts for which completed questionnaires were available. To avoid any confusion, detailed definitions are provided as follows:

- *Eligible authors:*. All authors who were contacted as abstract authors, or who were recommended for information on a particular abstract by a previously contacted author. Authors who considered themselves not to be the appropriate contact but provided an alternative, were erased from the eligible list. Similarly, contacted investigators who had erroneously been assigned an ab-

stract they had not authored were dropped from the list if they notified us about this.

- *Responders*: Authors who completed and returned the questionnaire(s). In addition, authors were included who refused to complete the questionnaire, but provided information on the outcome, e.g. by sending a re-print.
- *Non responders*: All authors to whom questionnaires were mailed and who did not return them. Undeliverable mailings are excluded. It is not taken into account (because it is not known) whether non responding authors actually received the questionnaires, and whether authors were actually eligible. Also included as non responders are authors who refused to give any information on the outcome, unless they recommended contacting a specific alternative author.
- *Undeliverable* mail: mail that was returned undeliverable, and for which no alternative or updated addresses could be identified subsequently. Authors for whom mailings remained undeliverable were not considered in the calculation of response rates.
- *Response rates* refer to the number of authors in the mail survey and are calculated using the following formula:

  $$n_{responders} / [n_{responders} + n_{non\ responders}]$$

  (Undeliverable mailings were not included in the denominator.)
- *Abstracts covered* by the survey/survey abstracts: all abstracts for which information on the outcome (publication) was provided by at least one of the authors.
- The *proportion of abstracts covered* was calculated as follows:

  $$n_{abstracts\ covered} / [n_{abstracts\ covered} + n_{abstracts\ not\ covered}]$$

Response bias: "misrepresentation of the target population by the respondent population in terms of the prevalence of risk factors and/or the prevalence of subsequent incidence of [the outcome]" (88). The target population was in this

case defined as all abstracts in the sample, and the respondent population as the abstracts covered by the survey.

Publication: full publication or acceptance for publication by a peer reviewed journal. "Full" implies the existence of sections on methodology used, results, and discussion. "Peer review" is defined for this purpose as a formal process of judging submitted articles for scientific and technical merit by other scientists in the same field (8) and includes the option for rejection of a manuscript

Formal abstract quality: comprises the quality of reporting (completeness) and the quality of the underlying research (appropriateness of measures undertaken to improve the internal validity of the study) as evident from the report

Journal descriptors

- Impact factor: *a measure of the frequency with which the "average" article in a journal has been cited in a particular year or period.* The impact factor of a journal is calculated by dividing the number of current year citations to the source items published in that journal during the previous two years. (Garfield, 1994 (89)).
- The impact factor is in this study used as an approximation for prestige of a journal. Other qualities of journals were category (general medical, GI, other clinical specialties, basic science) and journal origin. The residence of the editor in chief and the address for manuscript submission were taken as indicators for journal origin if origin was not evident from the journal title.

Principal or senior investigator: whoever feels that this term applies to him or her. A definition was not given as concepts may vary geographically. Any other abstract author was classified as co-investigator.

C. *Subjects: Inclusion and exclusion criteria*

All abstracts submitted to the 1992 to 1995 annual meetings of the American Gastroenterological Association were eligible. These abstracts are published in a supplement to *Gastroenterology*, including those rejected for presentation at the meeting. Abstracts that did not report empirical data or that could not be classified as clinical research or basic science studies were excluded.

D. *Data collection*

1. Overview

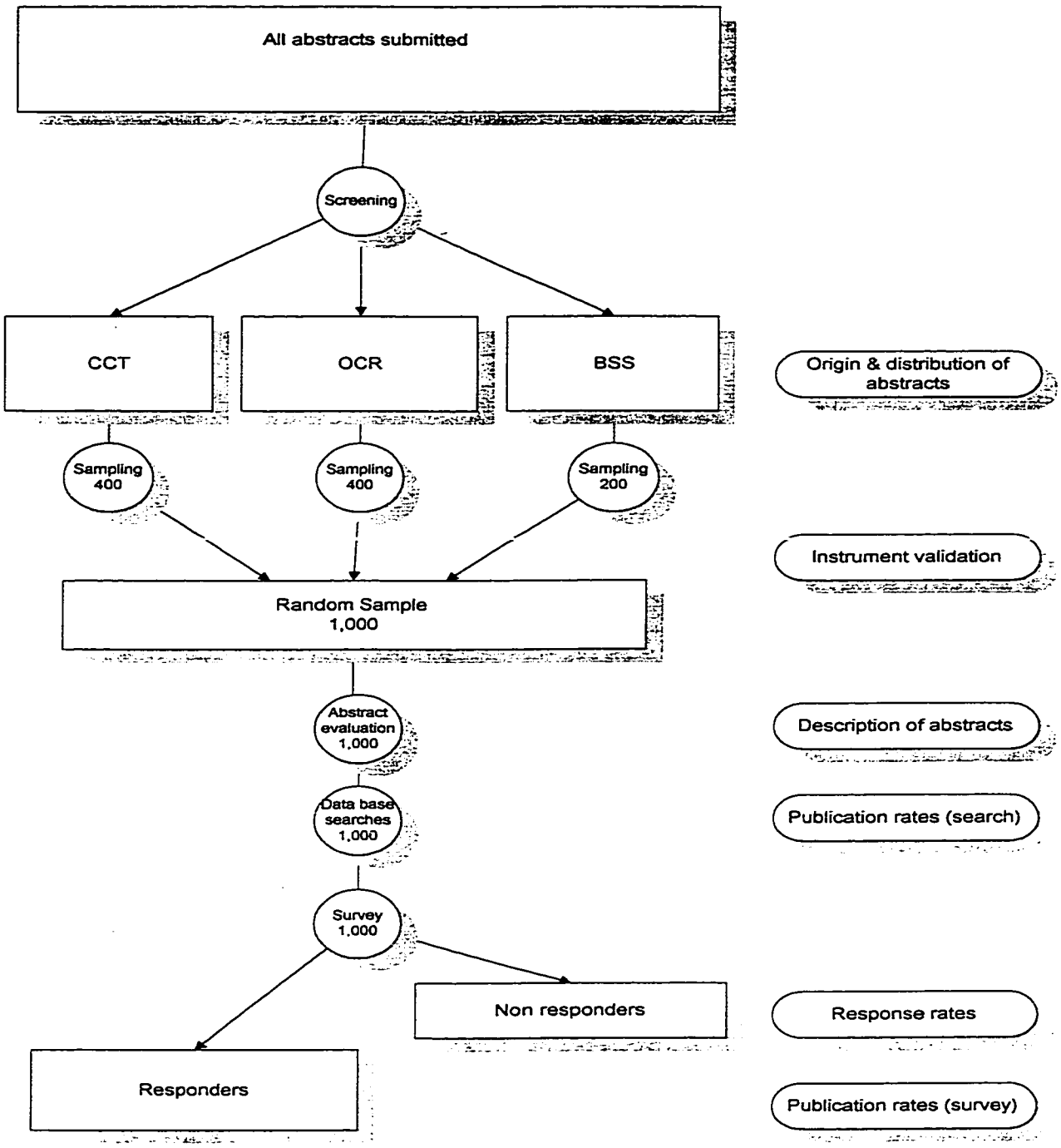
Figure 1 demonstrates the sampling procedures, and the different stages of data collection and analysis (for numbers see figure 2 and 3 - results section). Groups of abstracts are characterized by square boxes, while circles refer to procedures. The corresponding phases of the analysis are shown on the right side of the figure.

2. Abstract Screening and Sampling

All abstracts were screened by summer students for research type, country of origin and whether the abstract was accepted for presentation. The data were entered into a data base (SPSS) and computer-generated random numbers were assigned to all abstracts. A random sample was drawn stratified by year of submission and study type (CCT: n = 400, OCR: n = 400, BSS: n =200). The disproportionate sampling served to increase the number of CCT in the sample, while decreasing the number of BSS abstracts. Stratification by submission year was proportional and intended to find a balanced compromise between problems with recall by abstract authors with increasing time since submission vs. incomplete publication rates for more recent abstracts.

Sample abstracts were copied and title sections cut off in order to blind evaluation for acceptance of abstract, authors and affiliations.

Figure 1: Overview - data collection



3. Abstract evaluation

Abstracts were evaluated by a single observer in random order, using a validated evaluation form (appendix 1). To ensure consistency of evaluation, 10% of every 100 consecutively evaluated abstracts (overall 80) were re-evaluated by the same evaluator after 4 to 6 weeks (intra-observer reliability). The study type was reviewed and corrected where necessary. Other items recorded were study design, topic, sample size, statistical significance of study results, perceived direction of results, intervention tested where applicable and whether funding was disclosed. In addition, a previously validated 19 item quality scale was applied, examining completeness of reporting and features like randomization, blinding, use of control groups, control of confounding and appropriateness of statistical methods (appendix C).

4. Data base search

For all abstracts, MEDLINE was searched to identify full publications. Searches were based on first and last authors, and covered citations from the year of the meeting up to and including the most recent data base update at the time of the search. Publications were matched to abstracts based on titles, and, where necessary, by comparison of MEDLINE abstracts to meeting abstracts.

To identify and quantify potential sources of under-ascertainment, several additional search strategies were used for subsets of abstracts:

1) *Extension of search period* - search period beginning one year before presentation: all 1992-1994 abstracts
2) *Extension to other authors* - MEDLINE search based on all abstracts authors: 5% of 1992 - 1994 abstracts, or until lack of additional retrieval was felt to warrant abandonment of this strategy
3) Simultaneous search of *additional databases,* using identical search strategy and cross checks:

- Embase, Cancerlit were searched for all 1995 abstracts, plus a random sample (10%) of 1992-1994 abstracts
- HealthStar, Cinahl were searched for all 1995 abstracts, plus a random sample (10%) of 1992-1994 abstracts, or until lack of additional retrieval was felt to warrant abandonment of this strategy
- Biosis was searched for all 1995 basic science abstracts, plus an equally sized sample of 1995 clinical research (CCT and OCR)
- The Cochrane library was searched for all 1995 abstracts on CCT

As these subsets were considered random samples of the full sample with the exception of year of submission, all identified publications were included in the analysis, irrespective of source of retrieval. MEDLINE, Embase, Cancerlit, HealthStar, Biosis and Cinahl were accessed on line via the Health Knowledge Network through the University of Calgary and the University of Alberta (Edmonton). For the Cochrane search, the CD-ROM published by the Cochrane Collaboration was used (edition 1998 II) (90). For all publications, date of publication, first author, name of journal and source (data base) were recorded. Impact factors and journal type were assigned based on the science citation index (91). Prestige of journal was based on the distribution of impact factor values for all publications identified (high impact = impact factor within the highest tertile).

Time to publication was calculated as the time in months between the AGA meeting and the date of the journal issue. Negative values were substituted with the value 1 (i.e. full publications were available within the first month after the meeting). Time of follow up was defined as the time between meeting and data base search.

5. Author Survey

A four page questionnaire was sent to authors of abstracts (appendix 3). This questionnaire covered the following:

- study characteristics, such as design, randomization and sample size
- statistical significance of study results, perceived relevance
- source of funding
- characteristics of both the person surveyed and of the principal investigator: age, sex, number of previous publications
- current status of research

and, where applicable:

- publication - journal and date
- number of submissions
- reasons for rejection
- reasons for non - publication

The time required to complete the questionnaire was estimated as five minutes. An information sheet was included to clarify the terminology used (appendix 4). In addition, authors received a copy of the abstract the questionnaire was referring to, and an addressed return envelope. For the US-American part of the survey, return envelopes were randomized to postage or no postage. Otherwise, envelopes were stamped wherever stamps were available (Germany, France, UK, Ireland, Canada, Belgium, The Netherlands, Australia and Japan).

Addresses were identified from the abstracts, the AGA-address book, and various other sources such as the Internet, MEDLINE, and membership and medical association catalogues. Primarily, only the last author listed was addressed. Two written reminders were sent out to non-responders. The respective intervals between mailings were six to eight weeks for overseas mailing and four to six weeks for North-American mailings. In addition, alternative abstract authors, i.e. those listed first, were contacted at the time of the second reminder.

E. Instrument validation

As no universally applicable instrument was available to allow the calculation of a summary score for formal quality of abstracts, a score was developed and validated to estimate the methodological quality of submissions to scientific meetings.

Item selection and instrument structure were based on previously published instruments for the assessment of full papers in drug trials and epidemiological research (10,62,63), in particular the one by Cho and Bero due to its applicability to various forms of clinical research (92). In this instrument, items are assessed on a three point scale (fully met, partially met, not met). A summary score is calculated by dividing the overall number of points achieved by the number of possible points after subtraction of non applicable items. Adaptation of this scale for the use in abstracts took into account the published guidelines for the composition of structured abstracts of original research (69,93). For content validity, the items were discussed with investigators in basic science, clinical research and epidemiology. A detailed manual was developed to clarify standards within the large variety of research types (appendix 6).

Inter-rater reliability of the final instrument was assessed by independent evaluation of 100 randomly sampled AGA abstracts by two observers with training in both gastroenterology and epidemiology. For test- retest reliability (intra-rater agreement), 85 additional abstracts were evaluated twice by the same observer, after an interval of four to six weeks. This part was identical with the procedure used to monitor consistency of ratings in the main study (see above).

To test whether the instrument is suitable for describing abstract quality (construct validity) (94), two hypotheses were formulated:

1. Abstract quality scores are predictive of abstract acceptance by the AGA

2. Abstracts globally considered to be of exemplary formal quality by methodologically trained reviewers score better than abstracts that are not considered exemplary.

Comparisons for 1) were based on the sample used for reliability testing after calculation of average scores for both observers. High quality was defined as a score within the highest tertile (average of both observers). Exemplary abstracts were chosen out of a sample of 30 abstracts by post graduate students in Community Health Sciences/Epidemiology.

Finally, aspects of instrument sensibility as defined by Feinstein (95) were evaluated by the Canadian responders of the main survey, and by staff members of the Faculty of Medicine, using a mailed questionnaire. This questionnaire was adapted from Oxman (96) and included items such as the suitability of the instrument for the purpose of measuring formal quality as perceived by the respondents (appendix 7).

For the reliability testing, the full evaluation form for abstracts was used (appendix 1). Thus, items of the abstract evaluation which were not part of the quality scoring instrument (study type, sample size, acknowledgment of funding, statistical significance, perceived direction of results) were included in the validation process.

The questionnaire for abstract authors was pre-tested by ten staff members, and modified accordingly. A pre-run in Canadian authors of abstracts did not necessitate further modifications of the mailing procedure.

F. Statistical analysis

1. Instrument validation

Mean quality scores were calculated and reported with standard errors (SE). Agreement between summary scores was assessed by the calculation of intra class coefficients (97). For the assessment of single items, Cohen's kappa was used, with kappa > 0.4 indicating sufficient (fair to good) agreement (98,99). Sub group analysis was based on research type.

Differences in scores between accepted and rejected, and between exemplary and other abstracts were compared by Student t-tests for independent samples. Results from the sensibility test are quoted as mean scores out of ratings on a scale from 1 to 7. Mean scores ≥ 5 were considered to indicate satisfaction with the respective item.

2. Sample size and power

A priori sample size calculations were based on the comparison of studies with negative vs. studies with positive outcome with respect to publication rate. Sufficient power was required to examine the association at question separately by research type, as there were doubts whether these strata would be combinable. The year of abstract was not expected to influence publication rates when a time dependent analysis was used. Feasibility considerations assumed a sample size of 1,000 abstracts as the maximum possible within the one year planned for completion.

Fifty percent of the submitted abstracts were expected to be published in full (56). A 20% difference in publication rate was considered a relevant difference, translating into an OR of 2.3 for a 40% vs. 60% ratio (0.43 for a negative association). A pre-test was done on 50 abstracts to estimate the percentage of studies with nil results by research type. Given a 1:2 ratio of negative vs. positive

outcomes, the required sample size, assuming $1-\beta = 0.8$, $\alpha = 0.05$, two -sided, is about 85 in the smaller group, overall 252 (based on Altman's nomogram, standardized difference 0.4 (100)). In the search based section, the study on 400 abstracts had a power of 0.95. For the survey, a response rate of 67% was assumed, resulting in a power of > 0.80 (n = 267 per stratum).

For basic science abstracts, a 1:4 proportion of negative to positive outcomes was assumed. Due to the lack of previous data on the publication patterns in this area, and uncertainty about the applicability of the descriptors used, a higher difference was considered relevant for practical reasons (30%). Based on this, a total of 150 basic science studies was required to achieve a power of 80%. Thus, a sample of 400 CCT, 400 OCR and 200 BSS was drawn. This corresponded to proportions of 43% of all CCT submitted, 5% of OCR and 2% of BSS.

3. Data analysis

All tests of significance were two-sided with $\alpha = 0.05$. Analysis was performed using the software SPSS (101).

Abstract characteristics and information from the author survey were described by frequency tables, using chi-squared tests for the examination of possibly interrelated variables. Explorative analysis included the examination of predictors of abstract quality and abstract acceptance by logistic regression. Continuous variables (score, time to publication, impact factor) were, after graphical exploration, categorized based on tertiles to enable examinations of trends. Where adjoining categories showed similar behavior, binary variables were created: high impact = IF $\geq$ 3; high quality = score $\geq$ 0.63.

Agreement between information from different data sources was compared and agreement reported as the percentage of cases with perfect agreement, sup-

plemented by Cohen's kappa (98,99). Retrieval rates for the survey abstracts were calculated using the following formula:

$$n_{\text{publications identified by search (survey abstracts only)}} / n_{\text{publications identified by survey}}$$

Crude publication rates were calculated separately based on the data base search and the author survey, and are presented as percentages with 95% confidence intervals (binomial exact). Crude rates corrected for responder bias and under-ascertainment were estimated for the full sample by multiplying the data base search derived publication rate by the reciprocal of the retrieval rate:

$$\text{publication rate}_{\text{search (full sample)}} * [1/\text{retrieval rate}]$$

To calculate crude OR, and the effect of responder bias on OR, two by two tables were constructed, excluding all descriptive and equivocal studies. Based on the search results, response bias was estimated using the method described by Austin and Criqui (88,102): response rates are compared as a cross tabulation of outcome by exposure. The resulting OR represents an error term, which links the biased and the corrected OR as follows:

$$OR_{\text{responders}} = \text{error term} * OR_{\text{full sample}}$$

For multivariate analysis, logistic regression with stepwise backwards elimination was used to examine predictors of publication. Two approaches were used:

- a model based on the abstract evaluations and the data base searches
- a model based on the survey information.

Descriptive studies were excluded in all models. Model fit was examined by the Hosmer-Lemeshow-test (94,103).

For separate analyses by study type, Cox proportional hazards models were applied (due to interaction of time to publication with study type, this method was not used for the combined group). Goodness of fit was examined using likelihood

ratios. In addition, another Cox regression model examined the predictors of early publication by using all data censored at two years.

Annual publication rates were calculated as cumulative and conditional rates based on actuarial life tables, based on the search results (100).

Predictors of time to publication were explored using chi squared tests. Kaplan Meier curves were used to demonstrate effects of different variables on time to publication graphically. To examine the association between statistical significance and prestige of a journal, multiple logistic regression was applied to the subgroup of published studies, calculating adjusted OR for publication in high impact journals.

4. Missing variables

Information from the abstracts was expected to be often incomplete, especially with respect to the reporting of absent features. Therefore, for most variables, no distinction was made between "attribute not reported" and "attribute not present". The null value was assumed in both cases, and missing values were thus not expected. With respect to the main exposure under study (statistical significance of study results), a missing value was coded as "equivocal/not clear" (see definitions).

Corresponding questions in the survey were formulated in analogy, e.g. answer options for randomization or blinding included only "yes" [1] and "no/not applicable" [0], and missing values would be coded as [0] (see appendix). It is appreciated that this procedure constitutes a compromise, as the resulting null or equivocal category will contain a mix of actual values (104).

Exceptions from the procedure included sample size, funding and demographics of investigators. For these variables, cases with missing values were excluded from the respective analysis ("complete subject analysis") (104).

V. Results

A. Overview

First, the results of the quality score development and validation will be presented (section B). This section is separate from the main study and served to provide the instrument for the abstract evaluations.

The description of the study cohort (sample abstracts) is primarily based on the abstract evaluations (section C). This section will include an analysis of determinants of abstract quality and abstract acceptance. The section on survey information will focus on differences between non responders and responders, and will also provide information on the comparability of survey vs. abstract information (section D).

The sample sizes for the analyses resulting from the sampling and exclusion procedures will be presented in flow charts, separately for the abstract evaluation/ data base search approach and the survey based analysis (figure 2 and 3).

Section E contains the main analysis which consists of publication rates and determinants of publication. Results are presented in the form of synopses, contrasting the results from the two approaches. The determination of crude publication rates are complemented by analyses on the effects of bias. At the end of this section the focus is on secondary outcomes such as time to publication and impact of target journals.

Lastly, reasons for non - publication will be described (section F)

B. Instrument Validation

The average time needed to evaluate an abstract was 4:09 minutes (range 2:33 to 6:09). Scores ranged between 0.28 and 0.95, and were approximately normally distributed for both observers. Mean scores and agreement between observers were best for controlled clinical trials, poorest for basic science abstracts (table 1). Individual item analysis showed fair to good agreement for the majority of items including statistical significance and direction of results (kappa 0.41 to 0.74). Problematic were the questions on the quality of the research question and the outcome measure, and support of the conclusions (kappa 0.08 to 0.19). Test-retest reliability was excellent, with identical mean scores for first and second evaluations (0.57, SE 0.01) and an intra-class coefficient of 0.85. All single items had good intra-rater agreement (kappa 0.54 to 0.85).

Accepted abstracts had significantly better scores than rejected submissions (0.61 vs. 0.54, $p = 0.001$). This translated to an OR of acceptance of 2.7 (95% CI 1.3 to 5.6) for high quality abstracts (score $\geq$ 0.65). When grouped by research type, a tendency to better scores in accepted abstracts was evident for all groups. However, this only reached significance for clinical trials (0.67 vs. 0.52, $p < 0.001$). Scores were significantly higher for abstracts considered exemplary as compared to those not considered exemplary (mean scores 0.67 vs. 0.51, $p = 0.01$).

Sensibility questionnaires were completed by 13 observers. Overall, the instrument was considered useful and acceptable by methodologists and clinical researchers (table 3). In particular, it was felt that the instrument was suitable to measure both components of formal abstract quality (quality of underlying research methodology and quality of reporting). Considered problematic were the amount of time and effort needed, and the applicability to basic science. Also, it

was felt that the often limited information available from abstracts might be insufficient in some cases. Basic scientists found the instrument generally less useful, mainly due to the time and effort needed.

Table 1: Inter-rater reliability with intra class correlation coefficients

	n	mean 1 (SE)	mean 2 (SE)	ICC
Overall score	100	.58 (.01)	.61 (.01)	.69
CCT	42	.63 (.02)	.63 (.02)	.81
OCR	39	.57 (.02)	.59 (.02)	.67
BSS	19	.50 (.02)	.61 (.03)	.60

Table 2: Test-retest reliability with intra class correlation coefficients

	n	mean 1 (SE)	mean 2 (SE)	ICC
Overall score	85	.57 (.01)	.57 (.01)	.85
CCT	30	.66 (.02)	.67 (.02)	.84
OCR	34	.51 (.02)	.53 (.02)	.61
BSS	21	.54 (.02)	.51 (.03)	.54

Table 3: Sensibility ratings

	Basic Scientists (range)	Clinical Researchers/ Methodologists
Number of reviewers	4	9
1. Wide applicability	3.5 (1-5)	5.7 (3-7)
2. Use in basic science	2.5 (1-5)	5.4 (4-7)
3. Use in clinical research	4.5 (1-6)	5.8 (4-7)
4. Clarity and simplicity	3.3 (1-5)	5.7 (5-7)
5. Time and effort needed	2.3 (1-5)	5.6 (5-7)
6. Adequate instructions	3.8 (1-5)	5.8 (5-7)
7. Information available from abstract	4.0 (3-5)	4.6 (4-5)
8. Need for subjective decisions	4.8 (3-6)	5.4 (4-7)
9. No redundancy	4.3 (3-5)	6.3 (5-7)
10. Comprehensiveness	5.5 (2-7)	5.9 (3-7)
11. Response options	6.0 (4-7)	5.9 (3-7)
12. Discriminative power	5.8 (5-7)	5.7 (5-7)
13. Measures methodology	5.5 (4-7)	5.7 (4-7)
14. Measures reporting	5.3 (3-7)	5.9 (5-7)
Mean rating	4.3 (3.3-5.7)	5.7 (4.9-6.5)

Ratings on a scale from 1 (unacceptable) to 7 (fully met).

Mean > 5: item acceptable

C. Abstract evaluation: description of cohort

1. Overview

The absolute numbers of abstracts available in each subgroup is presented in Figure 2. Possibly confusing is the fact that the evaluation of the abstracts resulted in occasional revisions of the study type and subsequently in changes in the proportions of the three categories, as reflected in the chart. Further changes in sample size were due to the exclusion of descriptive studies from any analysis including the main exposure (statistical significance of the study results).

2. Origin and distribution of abstracts

To create the sampling frame, all abstracts submitted for the 1992-1995 AGA meetings were screened by a summer student for research type. During this four year period, 17,205 abstracts were submitted. Contributions were divided almost evenly between clinical and basic science research. 5.4% were CCT. The percentage of unclassifiable studies was below 1%. Abstracts were submitted from 65 different countries. However, the top 10 contributors together accounted for 87% of all abstracts submitted. Forty three percent of all abstracts originated in the US, followed by Japanese (10%) and German (8%) contributions. Country distribution varied by research type, with the US and Japan particularly strong in basic science, while 67% of all CCT originated in Europe.

3. Description of abstracts in cohort

Research type was revised during the abstract evaluation in the study sample (n = 1,000). Hereby, screening was found to have been particularly sensitive to CCT on the expense of specificity (table 4). This resulted in deviations from the original stratum sizes, in particular in a lower number of CCT: The sample included 326 CCT, 455 OCR and 219 BSS abstracts.

Figure 2: Abstracts available for analysis

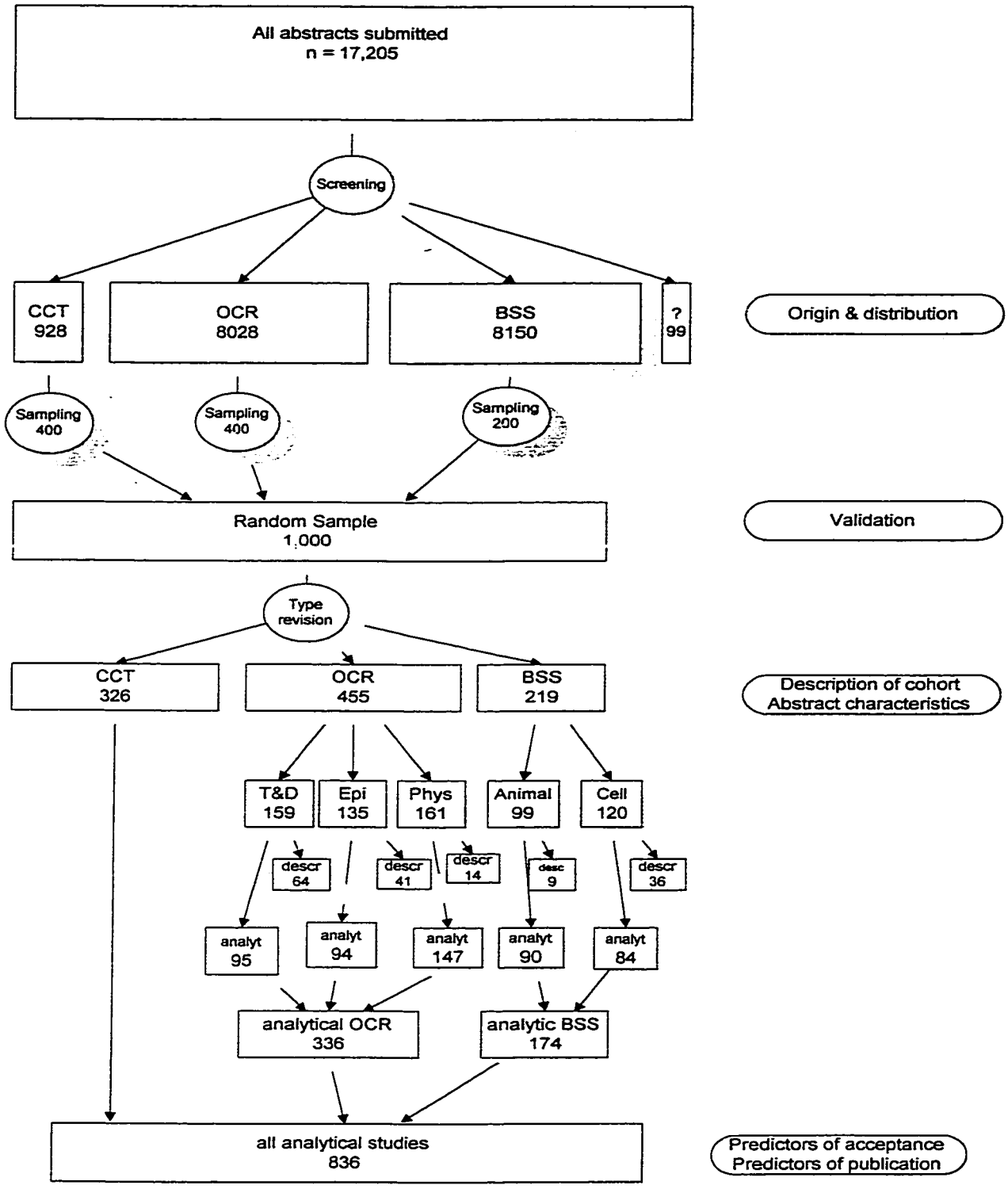


Table 4: Accuracy of screening

	Sensitivity	Specificity
CCT	94%	82%
OCR	79%	93%
BSS	84%	98%

Research topics varied. CCT most often studied proton pump inhibitors (22%), followed by H2-blockers and antibiotics. 15% of all CCT examined HP-eradication schemes, overall 46% examined acid- and/or reflux related diseases. Among the T&D reports, studies of endoscopic procedures were relatively prominent (18%). In epidemiology, diseases of the liver, in particular viral hepatitis, were the most frequent topic (21%), while GI-motility was the number one interest in studies of human physiology (33%). The most common topics in BSS were, similar to CCT, patho-physiological processes of the upper GI mucosa (21%).

Classification of study design and statistical significance was particularly difficult in BSS abstracts, as comparisons were often qualitative only, or a number of experiments were reported without obvious intention to compare the effects between interventions. In this group, 39% of the trials were parallel controlled, 11% were cross-overs, 30% before-after comparisons, and the rest observational (controlled or descriptive). A similar mixture was found in the designs used in human physiology studies and T&D. Table 5 gives an overview for study sample sizes (median number of subjects per group) and distribution of statistical significance by the different research types.

Table 5: Description of study characteristics

	CCT	T&D	Epid.	Phys	BSS
Total number	326	159	135	161	219
Analytical studies	326	95	94	147	174
Med. sample size (IQR)	23 (12-54)	22 (15-33)	38 (15-96)	10 (8-15)	6[1] (4-10)
Positive results	43%	38%	46%	48%	37%
Negative results	31%	12%	14%	12%	4%
Equivocal	26%	51%	40%	41%	59%
Mean quality score (95% CI)	0.65 (0.63-0.66)	[2]	0.55 (0.52-0.57)	0.54 (0.53-0.56)	0.51 (0.50-0.53)

Completeness of reporting was generally not very satisfactory (table 6). With the exception of CCT, a well defined research question or study objective was only given in about half of the abstracts. This occasionally impeded the decision on how to classify the statistical significance of a study. A reason for concern is the frequent failure to report attrition (and reasons for attrition) in CCT> Striking in BSS was the frequent failure to report the sample size (omitted in 46% of studies in intact animals).

[1] studies in intact animals only (n=99)

[2] n/a (part of criterion for differentiation between T&D and CCT)

Table 6: Completeness of reporting[3]

	CCT	T&D	Epid.	Phys	BSS
Description of study question / objective	68%	43%	48%	59%	53%
Design evident and appropriate[4]	-	-	49%	51%	54%
Description of subject characteristics	70%	32%	24%	23%	45%
Method of subject selection	14%	18%	26%	2%	-
Method of random allocation	0	-	-	0	0
Definition of outcome measure	50%	36%	34%	53%	48%
Sample size reported	100%	97%	100%	97%	54%[5]
Power calculations or CI for neg. results	8%	4%	4%	6%	3%
Reporting of statistical tests	28%	23%	36%	29%	9%
Exact p-values or confidence intervals	25%	17%	30%	16%	7%
Attrition and reasons for attrition	15%	22%	16%	2%	1%
Results reported in sufficient detail	68%	64%	57%	69%	48%
Conclusions supported by results	56%	37%	33%	43%	47%
Funding acknowledged	13%	8%	3%	10%	16%
Number of abstracts	326	159	135	161	219

Striking was the high percentage of trials with statistically negative results in the group of CCT (31%), which is statistically different from the proportion of negative results in T&D. Within the CCT, a tendency to higher rates of positive results with higher sample size was not statistically significant.

In a multiple logistic regression model, it was found that reports with equivocal and reports with negative results were inversely associated with good abstract

[3] relative frequencies relate to the number of "where applicable"

[4] n/a (part of criterion for differentiation between T&D and CCT)

[5] relates to studies in complete animals only

quality (table 7). T&D reports were expected to be of artificially low formal quality due to the study type categorization used. However, OCR reports were inversely associated with good abstract quality even after exclusion of T&D studies.

Table 7: Factors associated with abstract quality

	n	OR	95% CI
Statistical significance			
positive	318	1	
negative	136	0.5	0.3 - 0.8
equivocal	287	0.3	0.2 - 0.4
Research type			
CCT	326	1	
OCR	241[6]	0.3	0.2 - 0.4
BSS	174	0.2	0.1 - 0.3

n = 741; excluded: descriptive studies, T&D studies

Model c2 137.8 df 4 $p < .0001$

Hosmer and Lemeshow Goodness-of-Fit Test: c2 6.1; 5df; $p = .3$

The following variables were not associated with abstract quality: study size, topic, source of funding, year of submission, country group of origin

4. Predictors of abstract acceptance

The overall acceptance rate was 62%. Logistic regression analysis on all 836 abstracts (excluding descriptive studies) showed that formal quality, statistical significance of results, research type and country of origin were associated with abstract acceptance (table 8).

[6] excluded: T&D studies (n=95)

Table 8: Predictors of abstract acceptance

	n	OR	95% CI
Abstract quality score			
< 0.63	539	1	
≥ 0.63 (highest tertile)	297	1.5	1.1 - 2.1
Statistical significance			
positive	354	1	
negative	147	0.5	0.3 - 0.8
equivocal	335	0.7	0.5 - 1.0
Research type			
CCT	326	1	
OCR	336	0.6	0.5 - 0.9
BSS	174	0.4	0.3 - 0.7
Country group			
USA/Canada	311	1	
N/W Europe, Australia	279	0.4	0.3 - 0.6
S/E Europe	105	0.4	0.2 - 0.7
other countries	141	0.3	0.2 - 0.5

n = 336; excluded: descriptive studies

Model $\chi 2$ 60.4 df 8 $p < .0001$

Hosmer and Lemeshow Goodness-of-Fit Test: c2 7.2; 8df; p = .5

The model retained the same variables when reports with equivocal results were excluded. All OR and the model fit were almost identical. In particular, the negative association between negative study results and acceptance for presentation was sustained (OR 0.5, 95% CI 0.3-0.8. n = 501). A separate model for CCT also found similar OR for statistical significance and formal abstract quality. In addition, in this group, cross-over studies had a lower chance for acceptance as compared to trials with parallel design (OR 0.4, 95% CI 0.3 - 0.8).

D. Author survey

1. Overview

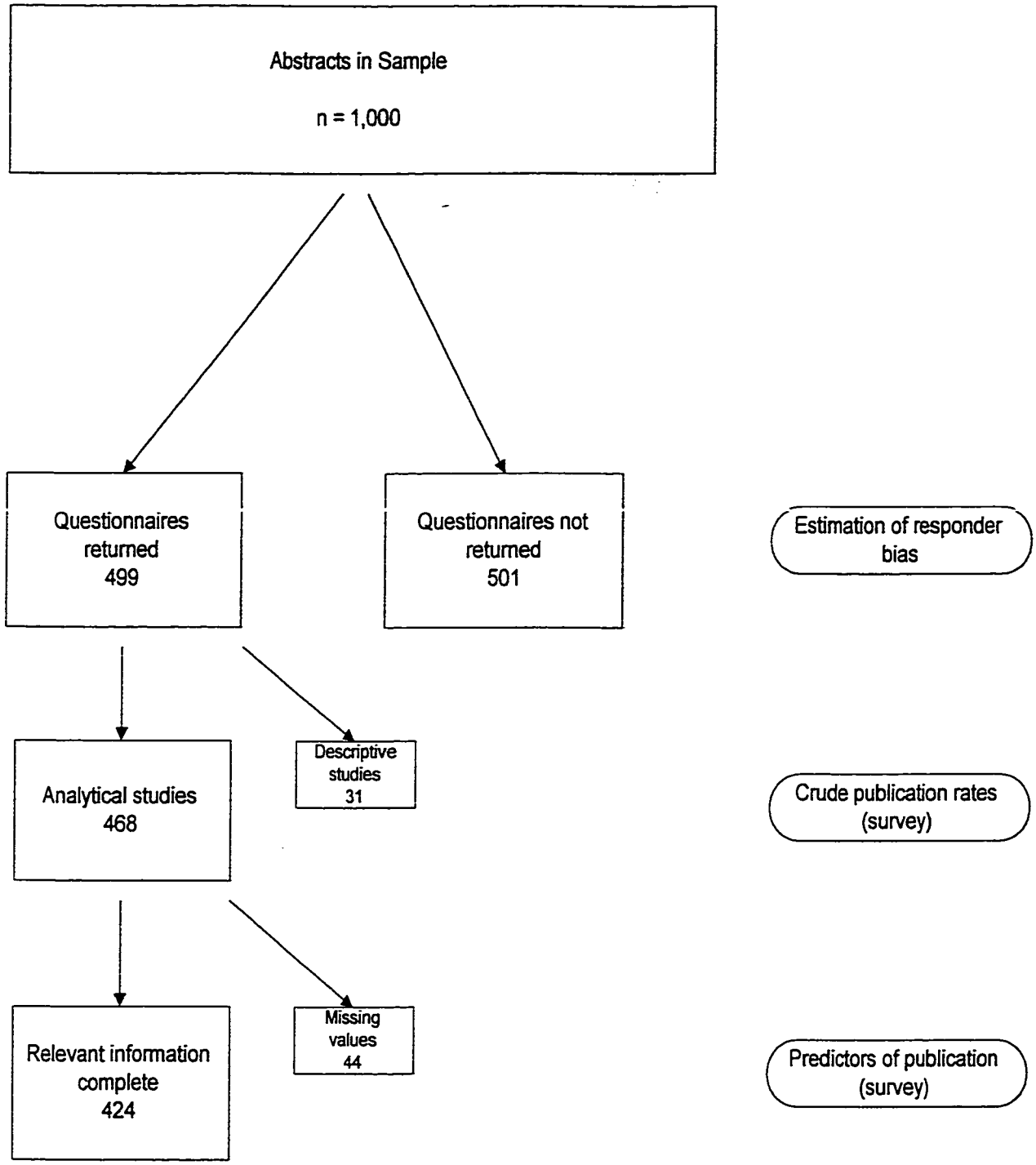
A schematic overview of the number of abstracts available for inclusion into the analysis is presented in Figure 3. Please note that the number of abstracts equals the number of questionnaires (after exclusion of doubles), but is not directly comparable to the number of authors surveyed, as presented in tables 9 and 10. In the case of non-response, alternative authors were addressed, increasing the number of authors. On the other hand, an author may have been listed on several abstracts. Reductions in the number of available abstracts were due to non response (n = 501), exclusion of descriptive studies for some analyses (n = 31), and to insufficient information on relevant items (n = 44, please refer to table 12).

2. Response rates and responder characteristics (Tables 9 and 10)

Mailings took place between October 1997 and July 1998 and included overall 1182 authors. The majority of authors (81%) received one abstract only; however, some investigators had up to six abstracts. The mailing was complicated by a two-week postal strike in December 1997. In addition to 58 first mailings which were returned undeliverable, 106 reminders were returned undeliverable while first mailings obviously remained missing in the mail. In 36 cases, authors, when reminded, notified us that they had not received previously received a questionnaire. Eventually, 102 of 1182 addresses remained unidentified (8.6%).

The overall response rate at the time of this thesis analysis (July 1998) was 40.8% (35.3% for first mailings, 8.4% following first reminders, 1.8% following second reminders). Response rates varied depending on time of mailing ($p = 0.005$), on author rank ($p < 0.001$) and on country group ($p = 0.002$) (table 9).

Figure 3: Number abstracts available for analysis in survey approach



Authors with more than one abstract also had slightly higher response rates (p = 0.09). Proportions of abstracts received by country are displayed in table 10, along with the corresponding response rates. Inclusion of return postage (randomized for US, October to March mailings only, n = 197) was not associated with better response rates. There were three refusals.

Table 9: Response rates

	authors in survey	responses (%)
Country group		
US/Canada	426	171 (40.1)
NW-Europe, Australia	379	175 (46.2)
SE-Europe	140	57 (41.4)
Other countries	135	37 (27.4)
Time of first mailing		
Oct 97 - Dec 97	493	246 (49.9)
Jan 98 - Mar 98	272	98 (36.0)
Apr 98 - Jul 98	315	97 (30.8)
Rank of author		
first on abstract	518	170 (32.8)
last on abstract	557	268 (48.1)
other or missing	5	3 (60.0)
Number of abstracts per author		
one	884	884 (39.6)
two or more	196	91 (46.4)
Postage (randomized)[7]		
yes	110	50 (45.5)
no	87	39 (44.8)
Total	1080	441 (40.8)

[7] US, October to March mailings only, n = 197

Table 10: Distribution of abstracts and questionnaires by country

Country	abstracts in sample	Response %	quest. received number (%)
US	349	38.0	161 (46.1)
Germany	106	53.8	64 (60.4)
France	86	37.7	38 (44.2)
Italy	73	36.9	40 (54.8)
UK	67	44.6	37 (55.2)
Japan	63	32.1	27 (42.9)
Canada	41	62.2	25 (61.0)
The Netherlands	26	53.3	14 (53.8)
Switzerland	26	35.0	10 (38.5)
Spain	22	34.8	11 (50.0)
Australia	20	68.4	15 (75.0)
Denmark	12	67.7	7 (58.3)
Belgium	9	50.0	5 (55.6)
Finland	9	33.3	5 (55.6)
Ireland	9	25.0	3 (33.3)
Poland	9	50.0	5 (55.6)
South Africa	9	33.3	5 (55.6)
Sweden	8	25.0	3 (37.5)
Israel	7	50.0	5 (71.4)
Greece	5	100.0	3 (60.0)
India	5	0	0
Austria	4	50.0	3 (75.0)
Mexico	4	33.3	3 (75.0)
Turkey	4	60.0	2 (50.0)
other	27	24.2	8 (29.6)
Overall	1000	40.8	499 (49.9)

Table 11: Comparisons of abstracts of responders vs. of non responders

	responders	non responders
Type of research (p = 0.8)		
CCT	158 (31.7)	168 (33.5)
OCR	229 (45.9)	226 (45.1)
BSS	112 (22.4)	107 (21.4)
Year of submission (p = 0.3)		
1992	120 (24.0)	130 (25.9)
1993	118 (23.6)	131 (26.1)
1994	124 (24.8)	127 (25.3)
1995	137 (27.5)	113 (22.6)
Acceptance for DDW (p = 0.001)		
accepted	341 (68.3)	290 (57.9)
rejected	158 (31.7)	211 (42.1)
Origin (p = 0.05)		
US/Canada	186 (37.3)	204 (40.7)
NW-Europe, Australia	169 (33.8)	148 (29.5)
SE-Europe	68 (13.6)	55 (11.0)
other countries	76 (15.2)	92 (18.2)
Stat. significance (p = 0.1)		
positive	189 (45.0)	165 (39.7)
negative	77 (18.3)	70 (16.8)
equivocal	154 (36.7)	181 (43.5)
Quality of abstract (p = 0.8)		
high quality	166 (33.3)	161 (32.1)
other	333 (67.7)	340 (67.9)
Total number	499	501

Table 11 shows the distribution of study and abstract characteristics for the projects covered in the survey, as compared to those not covered due to non re-

sponse or unidentifiable addresses. Based on abstract evaluation data, differences were found for abstract acceptance, which was higher in responders, and for country of origin. Response rates also varied by research topic (lowest: HP, 38%; highest: IBD, 67%, $p = 0.05$ over eight strata).

3. Completeness and validity of information in the survey

For most items in the questionnaire, information was complete in more than 90% of questionnaires (table 12). However, the proportion of missing values was high for any item that required recall of dates. Sample size was dropped as a possible determinant, as completeness was relatively low (83%), and the usefulness of this variable in the context doubtful. Due to the high proportion of missing values for date of publication, a time dependent analysis could not be performed on this data set. After exclusion of all cases with insufficient information in other relevant items (see table 12), 424 of 468 analytical studies (90.6%) were available for inclusion in the survey based multivariate analysis.

Comparing the information given by survey respondents to that from the abstract evaluation, there were significant inconsistencies (table 13). Some of these are obviously related to insufficient information from the abstracts. In particular, the source of funding had been so rarely mentioned that this factor could only be assessed for the survey projects.

Most surprising was the inconsistency with respect to abstract acceptance. Eighty eight (88) of 158 abstracts listed in the abstract volume as rejected for presentation, were reported by the authors as posters (81 abstracts) or talks (7 abstracts). Three authors reported that their abstracts had been rejected, while based on the abstract evaluation, they were accepted. Assuming the information from the abstract evaluation is correct, this would translate into a false positive rate of 51%, and a false negative rate of 1%.

Table 12: Completeness of information

	questionnaires (%)
Total	499
rank of abstract author	498 (98.0)
abstract acceptance	477 (95.6)
sample size	415 (83.2)
perceived relevance	468 (93.8)
number of centers*	485 (97.2)
source of funding*	483 (96.8)
number of previous publications*	458 (91.8)
current state of research	481 (96.4)
outcome (publication)*	499 (100.0)
variables with restricted applicability:	
• statistical significance (analytical studies only, n = 468)*	468 (100.0)
• date of publication (published studies only, n = 329)	137 (41.6)
• time to completion (completed studies only, n = 428)	173 (40.4)
• all variables included in survey model (n = 468)	424 (90.6)

* variables considered relevant

Similarly, there was some inconsistency with respect to statistical significance (table 14). Agreement was only 57% (kappa 0.3). Inconsistencies concerned primarily the equivocal abstracts, of which almost two thirds (64%) were eventually considered as having resulted in positive outcomes by authors. Exclusion of equivocal studies (by either classification) would result in a total agreement of 88%, with kappa = 0.7. Only 17 projects (8%) with equivocal abstracts were classified as statistically negative by the abstract authors. Of the 76 projects initially considered statistically negative, almost a third ended up as positive outcome studies.

Table 13: Comparison of abstract evaluation and survey

	by survey	by abstract evaluation
Total number	499	499
Acceptance for DDW		
talk	100 (20.0)	
poster	310 (62.1)	341 (68.3)[8]
rejection	67 (13.4)	158 (31.7)
not reported/ not known	22	-
Statistical significance		
positive	322 (69.5)	187 (40.4)
negative	64 (13.8)	75 (16.2)
mixed	77 (16.6)	201 (43.4)
descriptive	31	31
Source of funding		
government	122 (24.6)	16 (15.5)
industry	130 (26.2)	62 (60.2)
both	11 (2.2)	4 (3.9)
neither/other	233 (47.0)	21 (20.4)
not reported	3	396

Within the clinical trials, the situation was better. Agreement was 67%, kappa 0.6. However, as many as 27% of the studies that had statistically negative results at the time of presentation, were now considered statistically significant by the authors. On the other hand, almost a third of the equivocal abstracts were eventually labeled as having "negative" results. Therefore, again, inconsistencies were mainly due to changes in the "equivocal group". Exclusion of the group would result in 85% agreement, with kappa = 0.7. In spite of the considerable

[8] abstract evaluation could not differentiate between talk and poster presentation

changes in direction of outcome, the percentage of statistically negative clinical trials remained relatively high with 29% (OCR 6% negative, BSS 5% negative).

Table 14: Statistical significance: survey vs. abstract evaluation

	abstract			
survey	positive	negative	equivocal	Total
positive	168	23	131	322
negative	5	42	17	64
equivocal	15	11	56	82
Total	188	76	204	468

excluded: descriptive studies (n = 31);

kappa = 0.3.

Table 15: Statistical significance: survey vs. abstract evaluation, CCT only

	abstract			
survey	positive	negative	equivocal	total
positive	67	14	17	98
negative	3	33	10	46
mixed	3	5	6	13
total	73	52	33	158

kappa = 0.5

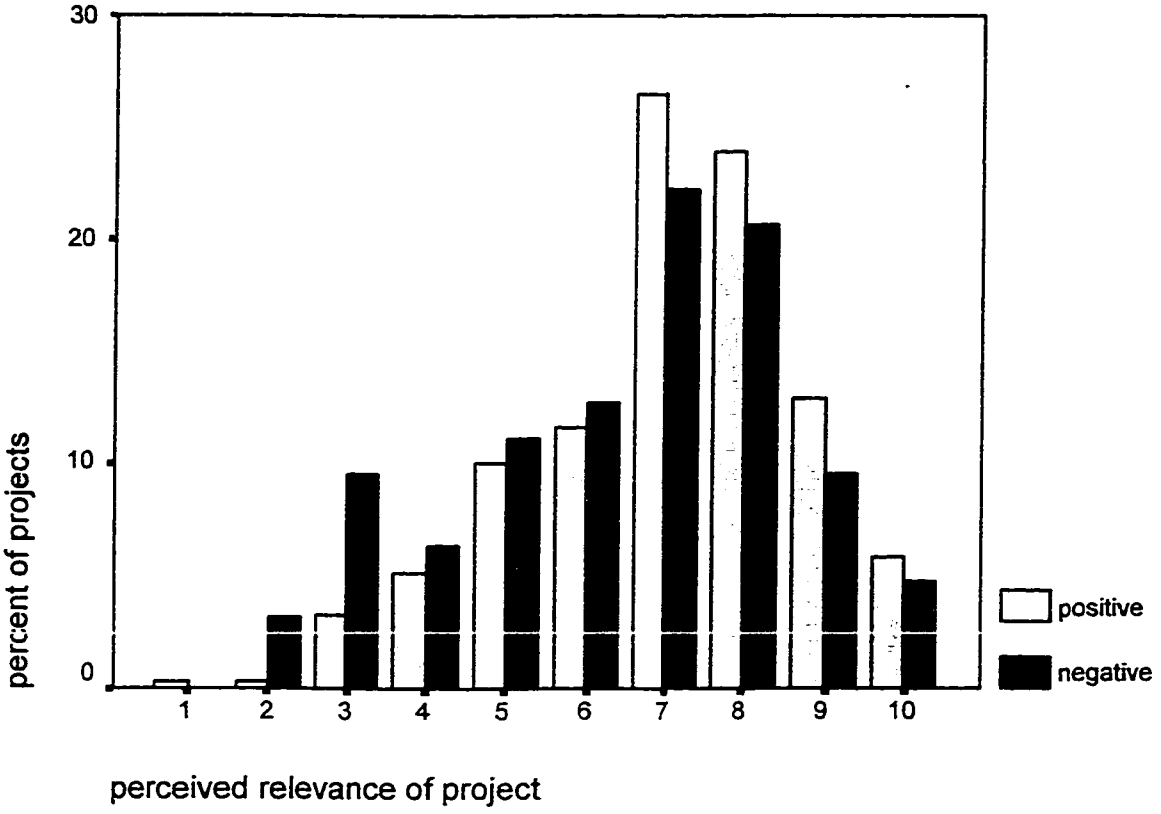
Agreement for design features like randomization, double blinding and use of placebo was 78% to 80% (kappa 0.2 for randomization, 0.6 for use of placebo, and 0.6 for double blinding). Disagreement was mainly due to failure to report an item in the abstract: of 24 clinical trials for which randomization was not reported in the abstract, 16 had actually been randomized.

A few variables could only be assessed by the author survey. These included the variables on PI demographics and on the perceived relevance of the project. Overall, 378 of 498 (75%) of the respondents were PI. There were no significant differences between PI and responding co-investigators with respect to age or sex, 93% being male and 60% between the age of 35 and 50. However, PI had significantly more publications (77% vs. 50% with more than 10 publications). Female PI were younger and had fewer publications than male PI.

The majority of authors rated their research projects as 7 or 8 out of 10 for relevance (median and mode 7, interquartile range 6 to 8). Generally, the older a respondent was, and the more publications he had, the more important he considered his research ($p = 0.001$). Projects were considered more important when abstracts had been accepted for presentation. Similarly, there were generally higher relevance ratings in positive outcome studies vs. statistically negative studies and equivocal studies (figure 4); however, this was not statistically significant.

Despite trying to record funding source during the abstract evaluation, information on this was absent in 89%. In the questionnaire, this item was 97% complete. The source of funding varies by research type, with industry funding more prominent in CCT (63% of CCT), and funding by a governmental agency more common in BSS (55% of BSS). Generally, only one study center was involved in a study project (76%; 2-5:15%, > 5: 9%).

Figure 4: Perceived relevance by statistical significance of results[9]



E. Publication of abstracts

1. Crude publication rates

Data base searches and survey were performed between October 1997 and July 1998. Follow up times were similar for the two modes of assessment. The minimum time was 34 months (median 50 months, maximum 75 months).

In the data base search (n = 1,000), full publications were identified for 458 projects (publication rate 45.8%). Based on the survey information, 329 of 499 abstracts were followed by full publication (65.9%). The calculation of data base

---

[9] excluding descriptive studies and, for clarity of display, studies with equivocal results

derived publication rates for the subgroup of surveyed abstracts identified two reasons for this difference:

For the surveyed abstracts 266 publications were identified, 63 were missed. This translates into a retrieval rate of 80.9%. In addition, there was evidence for responder bias in favor of published studies: Investigators were more likely to respond if they had their projects published. The data base derived publication rate for surveyed abstracts was 53.3% (95% CI 48.8% to 57.8%). This compared to a publication rate of only 38.3% (95% CI 34.0% to 42.7%) for abstracts of non-responders. Based on this information, the overall publication rate corrected for under-ascertainment and bias was estimated to be 56.6%.

There were no significant differences in the retrieval rates in different subgroups, although there was a tendency to lower retrieval in BSS (73%) as compared to CCT (87%). Interestingly, the country of origin was not of importance (range 79% to 84%). Missed publications comprised the whole spectrum of journals with respect to impact and origin. Only seven of the missed publications were published in a language other than English.

Table 16 demonstrates differences in publication rates by different abstract characteristics. The corrected estimates take into account the stratum specific response bias and under-ascertainment. Generally, trends seemed independent of the mode of outcome assessment. Higher publication rates were found for multi-center projects and for abstracts that had been accepted for presentation at the meeting. OCR abstracts did significantly worse than CCT and BSS. In addition, there was a positive correlation between time since abstract submission and publication rate.

Publication rates based on the statistical significance of the study results are shown for both abstract evaluation and author survey because of the previously

demonstrated inconsistency (table 17). Again, estimates were corrected by taking into account the under-ascertainment of publications by the data base searches. Several aspects are interesting in this table.

Table 16: Crude publication rates by abstract characteristics

	search		survey		corrected estimate
	n	pub. (%)	n[10]	pub. (%)	% (95% CI)
Total	1,000	458 (45.8)	499	329 (65.9)	56.6 (53.5 - 59.7)
Research type					
CCT	326	170 (52.1)	158	110 (69.6)	59.8 (54.3 - 65.2)
OCR	455	189 (41.5)	229	140 (61.1)	51.9 (47.2 - 56.4)
BSS	219	99 (45.2)	112	79 (70.5)	61.6 (54.9 - 68.1)
Abstract Year					
1992	250	118 (47.2)	120	84 (70.0)	62.0 (55.7 - 68.0)
1993	249	127 (51.0)	118	79 (66.9)	60.1 (53.9 - 66.4)
1994	251	111 (44.2)	124	85 (68.5)	56.1 (50.0 - 62.4)
1995	250	102 (40.8)	137	81 (59.1)	48.6 (42.5 - 55.2)
Country group					
USA/Canada	390	179 (45.9)	186	126 (67.7)	58.4 (53.4 - 63.4)
N/W Europe.	317	152 (47.9)	169	117 (69.2)	57.8 (52.1 - 63.2)
S/E Europe	123	53 (43.1)	68	42 (61.8)	54.8 (45.2 - 63.5)
other	141	74 (43.5)	76	44 (57.9)	62.4 (53.9 - 70.4)
Acceptance					
no	369	112 (33.3)	158	81 (51.3)	37.8 (32.7 - 42.8)
yes	631	335 (53.1)	341	248 (72.7)	65.5 (61.6 - 69.2)
Multicenter					
no	934	413 (44.3)	439	281 (64.0)	55.0 (51.8 - 58.3)
yes	66	45 (68.2)	45	37 (82.2)	79.1 (67.0 - 87.9)

[10] numbers may not add up to 499 due to missing values (see table 12)

First, inconsistencies in the corrected rates were highest with respect to the equivocal studies. For studies with positive and negative results, on the other hand, survey derived publication rates as well as corrected estimates hardly differed based on the mode of exposure assessment (positive 67.5% vs. 67.6%; negative 50.3% vs. 51.9%). There is a relatively high publication rate for negative outcome studies based on the data base searches (59.4%).

Table 17: Crude publication rates by statistical significance

	search		survey		corrected rate
	n	published	n	published	% (95% CI)
Total	836	392 (46.9)	468	310 (66.2)	57.7 (54.2 - 61.0)
Abstract evaluation					
positive	354	180 (50.8)	188	134 (71.3)	67.5 (62.3 - 72.4)
negative	147	69 (46.9)	76	47 (61.8)	50.3 (42.0 - 58.7)
equivocal	335	143 (42.7)	204	129 (63.2)	51.5 (45.8 - 56.8)
Author survey:					
positive	322	177 (55.0)	322	230 (71.4)	67.6 (63.8 - 71.4)
negative	64	38 (59.4)	64	41 (64.1)	51.6 (38.7 - 64.2)
equivocal	82	37 (45.1)	82	39 (47.6)	39.0 (28.4 - 50.4)

In the following two tables, studies with equivocal results are excluded to enable the calculation of risk measures from two by two tables. (Also, equivocal results are not considered relevant to the examination of RR or OR of negative vs. positive results).

In table 18 the effects of misclassification of exposure and outcome on relative risks are examined by contrasting evaluation/data base search and survey based information cross wise. Generally, RR are closest to the null value for the search based outcome assessment. Classification based on the survey information yields slightly lower RR. Correction for under-ascertainment results in another

very small correction of the RR away from the null, while correction for responder bias decreases the relative risk further. These trends seem independent of the classification of exposure. However, all changes are minimal and well within the possibility of chance.

Table 18: RR for publication of studies with negative results (vs. positive)

Exposure assessment	Outcome Assessment		
	search	survey	corrected rates
abstracts	1.08 (0.86-1.35)	0.90 (0.74-1.10)	0.75 (0.63-0.89)
survey	0.92 (0.76-1.13)	0.87 (0.71-1.06)	0.76 (0.59-0.98)

Another approach specifically examines the effect of responder bias on OR[11]: table 19 demonstrates the response rates by outcome and exposure (abstract and data base search derived information). In accordance with the results presented in table 11, the marginal response rates indicate no effect of the exposure. There is, on the other hand, a significantly higher response rate for authors with publications as compared to those who did not have their abstracts followed by publication (outcome response bias). However, looking more closely, there were inconsistencies within the inner cells, suggesting that the response bias in favor of published abstracts was more pronounced in authors of negative studies.

[11] The calculations by Austin & Criqui are based on OR. Our study is a cohort study. RR, as reported in table 17 are therefore the appropriate measure of size of effect. However, as we will be using logistic regression, we will eventually report results as OR, so that the use of OR as well as RR, depending on the occasion, seems justified.

Table 19: Response rates by stat. significance and publication status (%)

	published	not published	all
negative	65.2 (52.8 - 76.3)	41.0 (30.0 - 52.7)	53.4 (44.0 - 60.7)
positive	56.1 (48.5 - 63.4)	50.6 (42.9 - 58.2)	52.5 (48.0 - 58.7)
neg & pos	58.6 (52.2 - 64.8)	47.6 (41.3 - 54.0)	53.1 (48.6 - 57.5)

Based on the equation by Austin and Criqui, the resulting error term (OR of response rates) is 1.4. This means that the OR based on the responder abstracts overestimated the OR by the factor 1.4. This is confirmed by the calculation of OR for publication of negative results vs. positive results by responder status:

OR (responders) = 1.2 (0.7 - 2.1) OR (non responders) = 0.5 (0.3 - 1.0)
OR (all) = 0.9 (0.6 - 1.3)

In words, for responders, the OR for publication of negative vs. positive results was > 1 (i.e. in this group, studies with negative results were more likely to be published than studies with positive results). For non responders, the association was reverse. The combined group shows an OR slightly below 1, i.e. studies with negative results were in fact less likely to be published than studies with positive results when the responder bias was corrected. The confidence intervals in all groups, however, included 1; all effects may just be due to chance.

Publication rates could not be corrected for those variables where information was only available from the survey (table 20). The correlation between the mode of abstract presentation as well as the perceived relevance with the rate of publication was striking. The age of the investigator was also significantly associated with publication, but this could be shown by stratified analysis to be an effect of confounding by the number of previous publications.

Table 20: Crude publication rates (subgroups based on survey information)

	n[12]	pub	% (95% CI)
Total	499		
Source of funding[13]			
government	133	93	69.9 (61.4 - 77.6)
industry	141	95	67.4 (59.0 - 75.0)
neither/other/not reported	236	152	64.4 (57.9 - 70.5)
Abstract presentation			
no presentation	67	29	43.3 (31.2 - 56.0)
poster	310	208	67.1 (61.6 - 72.3)
talk	100	75	75.0 (65.3 - 83.1)
Sex of investigator			
male	425	290	68.2 (63.6 - 72.6)
female	34	22	64.7 (46.5 - 80.3)
Age of investigator			
< 35	66	40	60.6 (47.8 - 72.4)
35 - 50	285	188	66.0 (60.1 - 71.5)
> 35	126	90	71.4 (62.7 - 79.1)
Number of previous publications			
0-10	115	64	55.7 (46.1 - 64.9)
>10	340	240	70.6 (65.4 - 75.4)
Perceived relevance			
not so important (1-5)	111	48	43.2 (33.9 - 53.0)
important (6-7)	173	121	69.9 (62.5 - 76.7)
very important (8-9)	184	142	77.2 (70.4 - 83.0)

[12] numbers may not add up to 499 due to missing values (see table 12)

[13] categories not mutually exclusive

2. Multivariate analysis: predictors of publication

Cox regression models could not be calculated for the full sample because of evidence for interaction of study type with time to publication and violation of the proportional hazards assumption. Therefore, logistic regression was used. Results are reported in table 21. Variables which were only significant predictors in one of the two models are included in the complementary display in italics for comparison.

Based on the data base search results and the information from the abstracts, only multi-center status was found to be significantly associated with full publication. In the survey model, statistical significance and number of previous publications were retained as significant predictors. However, only studies with mixed results did significantly worse than statistically positive studies. The OR for nil results were very similar in the two models, being insignificantly lower than the null value (0.8, 0.7). Research type, funding and origin were not significant in either model.

Because of the smaller sample size and the high proportion of missing data on the time to event in the survey, separate analyses by study type using Cox regression analysis were performed only on the search data. For the subgroup of clinical trials, multi-center status and quality of original abstracts were found to be predictive of publication: studies cooperatively conducted by more than three centers had an HR of 2.0 (95% CI 1.4 to 3.0), and high quality abstracts had an HR of 1.4 (95% CI 1.0 to 1.9) (table 22).

For basic science studies, both abstract acceptance ($p < 0.001$) and multi-center status ($p = 0.04$) were predictive of eventual publication on univariate analysis (log ranks). A Cox regression analysis, which did not include abstract acceptance, failed to identify any significant predictors of publication. Similarly, Cox regression analysis was unsuccessful in OCR.

Table 21: Predictors of publication

	Search OR (95% CI)	Survey OR (95% CI)
Number of abstracts	836	423
Statistical significance		
positive	*1*	1
negative	*0.8 (0.6 - 1.2)*	0.7 (0.4 - 1.2)
equivocal	*0.7 (0.6 - 1.0)*	0.4 (0.2 - 0.7)
Previous publications of PI		
less than 10	not available	1
more than 10		1.7 (1.1 - 2.8)
Multicenter status		
1-3 centers	1	*1*
> 3 centers	2.8 (1.6 - 4.9)	*2.0 (0.9 - 4.5)*

Data base search based: Model $\chi 2$ 17.1, df 3, $p < .001$. Hosmer and Lemeshow Goodness-of-Fit Test: $\chi 2$ .24; 3df; p = .97

Survey based: n = 423; Model $\chi 2$ 13.2;df 1; $p < .001$; Hosmer and Lemeshow: skipped (d.f. <1)

Table 22: Predictors of publication: CCT only

	n	HR (95% CI)
CCT (n = 326)		
Multi-center status		
1-3 centers	280	1
> 3 centers	46	2.0 (1.4 - 3.0)
Formal abstract quality		
Average or below (< 0.63)	149	1
High (≥ .63)	177	1.4 (1.0 - 1.9)

Model $\chi 2$ 22.5, df 2, $p = < 0.001$. Likelihood ratio (Goodness-of-Fit): $\chi 2$ 37.6; 16 df; $p < 0.001$

3. Time to publication

Time to event analyses could only be performed on the data base searches. Cumulative and conditional publication probabilities (the probabilities of publication if not yet published) are shown in table 23. Publication rates were highest in year two. If a project was not published within four years, the chances of subsequent publication were only about 3% per year.

Table 23: Annual publication probabilities

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Cumulative	15.4%	30.7%	39.9%	44.5%	46.5%	47.2%
Conditional	15.4%	18.1%	14.1%	8.9%	3.4%	2.9%

Median time from meeting to publication could not be calculated due to the low overall publication rate. For clinical trials only, (publication rate 52.1%), the median time was 44 months (95% CI 27-60). It was longer in negative outcome trials (54 months) as compared to trials with positive results (33 months). Studies with equivocal results did not reach 50% publication (figure 5).

Within the published studies, time to publication was found to be associated with year of submission, indicating ongoing publication activity after year three, the minimum time of follow up. Also, there was an association with study type. Within the first year, relative publication rates are highest for BSS, lowest for CCT (45% vs. 25% of eventually published studies), while in the long run, CCT ended up with a higher publication rate. This phenomenon is graphically demonstrated by crossing lines (Kaplan Meier method, figure 6).

Figure 5: Time to publication by statistical significance

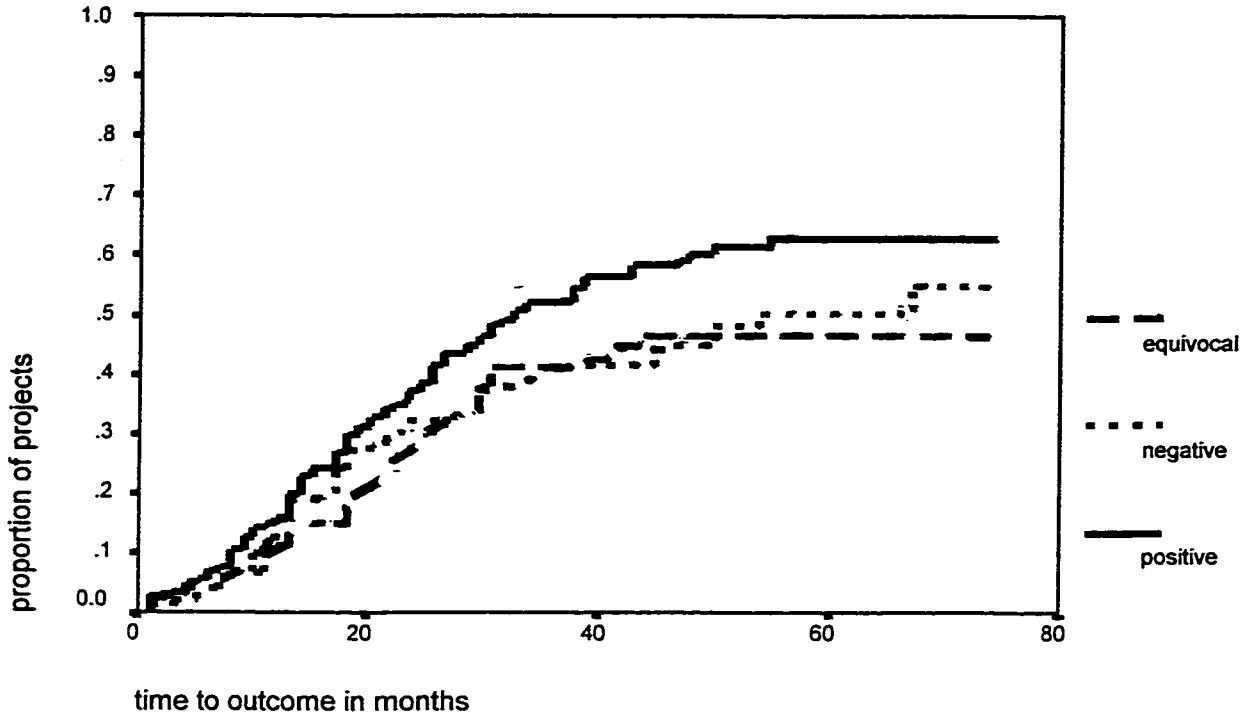
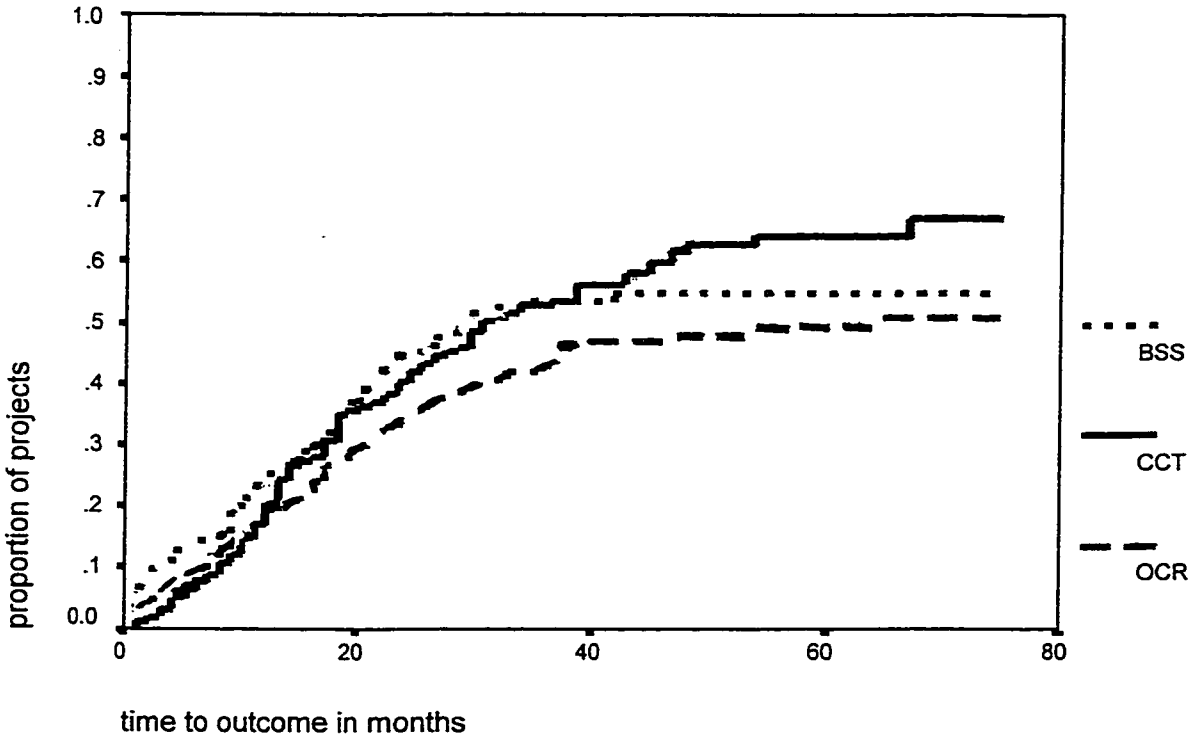


Figure 6: Time to publication by research type



In a logistic regression analysis comparing projects published within two years to projects published thereafter, there were significant associations with country group (p = 0.03). Specifically, SE-European projects were published faster than North-American manuscripts (OR 6.6, CI 1.7 - 25.7). However, the confidence interval was huge, and overlapped widely with OR for all other country groups except North-America. Other factors including significance of study results could not be demonstrated to be associated with differences in time to publication within the subgroup of published studies. Also, journal characteristics did not seem to be important: Impact factor, origin and type of a journal were without significant influence on the time to publication.

4. Journals

The majority of the identified publications appeared in gastroenterological journals, most often *Gastroenterology* (n = 39), *Gut* (n = 34) and *Digestive Diseases and Sciences* (n = 32). The leading basic science journal was the *American Journal of Physiology* (n = 23). Four percent (n = 41) of all research projects submitted as abstracts were published in general medical journals.

US based journals were prominent when the origin of journals was examined (60%). A further 18% appeared in journals originating in the UK. Only fifteen publications (3.3%) in languages other than English were identified. Table 24 gives an overview on the profile of journals in which projects were published, broken down by study type. The median impact factor was 2.1 (IQR 1.2-3.3). A third (33.3%) of all publications were in journals with an impact factor of 3.0 or more; these were defined as high impact publications.

On chi-squared tests of published projects, publications following abstracts accepted at the DDW were more likely to be in high impact journals than those following rejected abstracts (39% vs. 19% of publications, p < 0.001). Also, formal quality of the initial abstract was important: the proportion of high impact publica-

tions was 23% for low, 32% for average and 43% for above average abstracts ($p = 0.005$). Publications of studies with positive results had higher impact factors than those with negative results (38% vs. 26%; mixed results: 33%). However, statistical significance was not achieved for this variable, nor for study type, sample size, source of funding, multi-center status, and country of origin.

Table 24: Journal profiles of publications by research type (% of projects)[14]

	CCT	OCR	BSS	Total
published	170 (52.1)	189 (42.1)	99 (45.2)	458 (45.8)
non-English	8 (2.4)	5 (1.1)	2 (0.9)	15 (1.5)
Journal category				
general medical	23 (7.1)	17 (3.7)	1 (0.5)	41 (4.1)
GI/ hepatology	121 (37.1)	124 (27.3)	27 (12.3)	272 (27.2)
other clinical	24 (7.4)	43 (9.5)	26 (11.9)	93 (9.3)
basic science	1 (0.3)	5 (1.1)	45 (20.5)	51 (5.1)
Journal prestige				
impact factor > 2	104 (31.9)	110 (24.2)	46 (21.0)	260 (26.0)
impact factor > 3	62 (19.0)	62 (13.6)	31 (14.2)	155 (15.5)
Journal origin				
US - based	88 (27.0)	115 (25.3)	74 (33.8)	277 (27.7)
UK - based	46 (14.1)	31 (6.8)	7 (3.2)	84 (8.4)
European /intern.	14 (4.3)	28 (6.2)	5 (2.3)	34 (3.4)
national (non US/UK)	22 (6.7)	15 (3.3)	12 (5.5)	62 (6.2)
total	326	455	219	1,000

On multiple regression, only high abstract quality at the DDW was found to predict publication in a high impact journal (OR 2.0, 95% CI 1.3 to 3.1). The association was even stronger, when the analysis was restricted to clinical trials (OR

3.9, 95% CI 2.0-7.8). In addition, in this group, industrial funding was associated with a lower proportion of high impact publications (OR 0.6, 95% CI 0.2-0.8). Statistical significance of study results, research type, country group of origin, funding by a governmental agency and sample size were not predictive.

Table 25: Top three journals by type (% of all abstracts)

Rank	CCT	OCR	BSS
1.	Gut: 18 (5.5)	Dig Dis Sci: 18 (4.0)	AJP: 21 (9.6)
2.	APT: 16 (4.9)	Gastro: 17 (3.7)	Gastro: 9 (4.1)
3.	AJG: 14 (4.3)	Gut: 16 (3.5)	EJP: 4 (1.8)

Gastro: Gastroenterology; APT: Am J Physiol; APT: Al Pharm Ther; EPT: Eur J Pharmacol

Of all projects initially submitted to the DDW, 34 (3.4%) made it into the leading gastroenterological journal (*Gastroenterology*, IF 8.2). Exploratory analysis did not find any factors significantly associated with publication in *Gastroenterology*, in particular no evidence for publication bias, bias based on origin of authors or a preference for one of the research types.

F. Reasons for non-publication

For 143 projects which were not published, information about the current stage of the research was available (table 26) . Eighty five (59%) had been abandoned while 58 were still in progress. It is of note that 17 projects were still in the experimental (data collection) phase three to six years after abstract submission. In general, projects were not published because they were not written up and submitted (41% of non published projects). Only 22 projects (15%) were not published due to editorial rejection, a further seven were still in the submission-publication process.

[14] referring to all (published + unpublished) abstracts in subgroup

Table 26: Current state of research

Phase last completed	abandoned	in progress	total (%)
(data collection not completed)	16 (18.8)	17 (29.3)	33 (23.6)
data collection	8 (9.4)	3 (5.2)	11 (7.7)
data analysis	39 (45.9)	31 (53.4)	70 (40.6)
manuscript submission	22 (25.9)	7 (12.1)	29 (19.9)
Total	85	58	143

Table 27: Reasons for non publication

	a reason[15]	main reason[16]
number of respondents	109	109
lack of time	73 (67.0)	45 (41.3)
(co-)investigator left	32 (29.4)	12 (11.9)
lack of interest	25 (22.9)	4 (3.7)
sample size/recruitment problems	23 (21.1)	5 (4.6)
limitations in methodology	20 (18.3)	5 (4.6)
unimportant results	13 (11.9)	3 (2.8)
external problems	12 (11.0)	6 (5.5)
rejection anticipated	10 (9.2)	2 (1.8)
publication no aim	10 (9.2)	3 (2.8)
negative results	9 (8.3)	3 (2.8)
side effects/ethical problems	5 (4.6)	1 (0.9)
equipment/software problems	2 (1.8)	1 (0.9)
no main reason given	-	19

[15] more than one answer possible

[16] one answer only

Reasons for non publication were given by 109 authors (table 27). By far the most frequently mentioned was lack of time (67%), which was the main reason for non publication in 41%. "Co-investigator left" ranked second (29%, main reason in 12%), followed by lack of interest (23%), which was, however, rarely considered the main reason (4%). Also rarely decisive were technical or methodological problems pertinent to the project. Three projects were not published due to negative results, in a further six percent the lack of statistical significance was mentioned as one of several reasons.

Of the projects not published, none had so far been submitted to more than one journal. Published projects had been accepted by the first journals approached in 85%. Only three percent had been submitted to three journals or more (up to five). Reasons for editorial rejection were given by 110 respondents, including those of papers published after resubmission (table 28). Negative results were given as a reason for rejection in nine cases (not identical to those mentioned above).

Table 28: Reasons for rejection

	n (%)
methodology problems	37 (33.6)
no interest in topic	33 (30.0)
sample size problems	23 (20.9)
redundancy, lack of originality	21 (19.1)
negative results	9 (8.2)
unimportant results	7 (6.4)
formal requirements not met	5 (4.5)
total	110[17]

[17] Sum ~= total, as multiple responses possible

G. Posteriori *power calculations*

The following power calculations are restricted to the analysis of OR for publication of negative outcome studies as compared to those with positive outcomes. Equivocal and descriptive studies were excluded. Beforehand, a 20% difference in risk (OR 0.43) had been considered the relevant smallest detectable OR for clinical studies.

In the search based approach, the observed numbers of available abstracts were reduced due to the unexpectedly high proportion of equivocal studies, in particular in OCR and BSS. In these subgroups, the analyses had insufficient power. In the survey, the low response rate further decreased the numbers. However, for the combined groups, the power was sufficient in both approaches.

Table 29: Probability to detect relevant OR (% power)

assessment	pub:not pub	% neg (not pub)[18]	OR = 0.4	OR = 0.5
search	180:174	29.3	98.3	89.4
CCT	108:131	41.4	90.9	72.5
OCR	109: 81	21.6	47.7	32.1
BSS	35:37	9.7	15.2	11.2
survey	230:92	16.6	95.8	84.5

alpha = 0.05, two sided; excluded: equivocal studies, descriptive studies

[18] percentage of negative results among those unpublished studies with either negative or positive outcome

VI. Discussion

A. Review of the findings

This study examined determinants of publication based on a cohort of 1,000 abstracts submitted to the annual meetings of the American Gastroenterological Association. In particular, the occurrence of publication bias was of primary concern. A novel aspect of the study was its comprehensiveness: the follow up included all types of research submitted in abstract form to a large international gastroenterological meeting. The study was therefore expected to provide representative information on submission and publication processes and characteristics in the field of gastroenterology in an international context. A further unique feature of the study design was the combination of a follow up by data base search, complemented by a mail survey, with analyses based on a detailed evaluation of all abstracts. This design was chosen in the view of the different advantages and disadvantages of mail surveys and data base searches, and enabled us to examine and quantify sources of bias introduced by either of these methods.

In all, 326 clinical trials, 455 abstracts from other clinical research and 219 basic science abstracts submitted between 1992 and 1995 were followed up for a minimum of three years. Data base searches were complete for all 1,000 abstracts. Through the survey, further information was received on the fate of 499 abstracts.

The publication rates were 45.8 % (95% CI 42.7 to 48.9) for all abstracts based on the data base search and 65.9% (95% CI 61.6% to 70.0%) for the 499 abstracts represented in the mail survey. Crude publication rates were highest for clinical trials (52%). Taking into account a retrieval rate of 81% in the data base search and a bias in the survey response which favored published abstracts, the estimated overall publication rate was 57%. These numbers are well in accor-

dance with figures reported from other medical societies, where ranges between 32% and 77%, with a mean of 51% were reported (42). Our data base search results, though somewhat lower, were also compatible with the publication rate of 48.9 (95% CI 46.9 to 50.9%, p = 0.1) reported by Duchini, who had followed up 2512 DDW abstracts submitted in 1991 for a period of five years (56). In contrast to us, he found higher publication rates for basic science as compared to clinical research (54% vs. 47%). However, this study was only published in abstract form, and information on sampling and search procedures are not available. Also, no attempt was made in this study to examine and control for predictors of publication other than country and research type.

Predictors of publication were assessed separately by the two methods of data collection. The analysis based on data base searches and abstract evaluation included all analytical studies in the sample. In a second approach, the estimation of odds ratios for publication was based on the information collected in the author survey. (The methodological issues relating to these different approaches and outcomes will be discussed below).

Based on the data base search results, only multi-center status was found to predict subsequent publication. The importance of this variable confirms previous findings, in particular by Dickersin et al. (38). This association did not reach statistical significance in the survey based model; however, the number of multi-center studies was low, and the confidence intervals were very large.

The OR for publication of negative outcome studies in the search based analysis was 0.8, which did not significantly differ from 1.0 (for positive results). In the survey based analysis, studies with equivocal results were found to be significantly less likely to be published compared to positive results, but the OR of 0.7 for negative results did not reach statistical significance. There was thus no firm evidence for publication bias in either model, which is in contrast to the evidence

prevailing in the literature (1,3). Possible explanations for this inconsistency will be discussed below.

On the other hand, the finding of lower publication rates in equivocal studies as compared to positive outcome studies is consistent with the reports in the few other studies where a similar distinction was used (40,41).

In the survey based analysis, the number of previous publications of the principal investigator was also found to predict subsequent publication. Seniority of author as measured by the number of previous publications has not often been included in studies evaluating publication bias, as this variable requires information from authors. However, the association had been previously described by Sommer in a study on menstrual cycle research (48). The implications of this are not quite clear. It has been suggested that manuscript quality correlates with seniority of author in blinded abstract review (79). Thus the association between seniority and chance for publication may simply reflect that experienced researchers have better chances for publication and continue to do so due to the quality of their work. However, our evaluation of meeting abstracts did not support this assumption, as there was instead a tendency for better formal quality in less experienced researchers. More realistic journal selection is an alternative explanation. A survey comparing criteria for selecting journals between first and subsequent submissions found that for first submissions, authors tend to aim for higher prestige of journals, while with subsequent submissions, odds for acceptance become more important (105).

Separate analyses by study type were only performed based on the search data. Here, abstract quality was found to predict publication in clinical trials. This is a novel finding. Abstract quality had so far rarely been examined, and assessment had been restricted to only a few key features (36). While several problems are associated with the use of this variable, the finding may indicate that a good ab-

stract reflects the quality of the underlying research (67), at least in clinical studies.

No predictors of publication were found for clinical research other than controlled trials or for basic science, but the power in these subgroups was low.

Examining the effect of statistical significance on secondary outcomes, some indirect indication for publication bias could be identified. Median time to publication was shown to be longer in clinical trials with negative results as compared to positive outcome trials. However, we would hesitate to interpret these findings as an independent feature of publication bias, as suggested by Ioannidis (57), as some methodological problems are involved in this approach (see below). In addition, impact factors were higher in published studies with positive results, as compared to studies with statistically negative or equivocal outcomes, confirming previous reports suggesting that susceptibility to publication bias may be increased in higher impact journals (22,37).

A very consistent finding in all studies surveying investigators of unpublished studies is the importance of the investigator in the decision of whether a project gets published. We found more data to support what can by now be considered an established fact: failure to publish in the vast majority of cases rests with the investigator. The most frequent reason given is lack of interest or insufficient priority (1,40,41,49). We tried to evaluate to which degree this reflects anticipation of rejection by editors, as an indirect effect of editorial bias; however, only two investigators quoted this as the primary reason for non submission. Overall 10% admitted that anticipated rejection had played a role in the failure to publish.

While many studies used abstracts to scientific meetings to calculate rates of subsequent publication, few data exist on the determinants of abstract acceptance. This is unfortunate as abstract acceptance has been shown to be an im-

portant predictor of subsequent publication (41). We confirmed the relevance of abstract acceptance for publication, noting even higher rates for oral presentations as compared to posters. The association was consistent in all types of research, even though obvious misclassification problems occurring in the survey may have weakened the effect. We have, however, not treated this factor as an independent variable in the prediction of publication. Rather, the significance of abstract quality, and the evidence for publication bias in abstract selection procedures, which have been described before (14,41) indicate that abstract acceptance is an outcome related to publication.

The country of origin was a factor which had not been examined before as previous studies were restricted to US or other national meetings or registries. While with respect to abstract acceptance rates, a bias in favor of North-American contributions was identified, this variable was not important in the prediction of publication. This is somewhat surprising as a bias in retrieval rates in favor of English language publications and in particular of US-American journals (resulting in an underestimation of non-US publication rates), was expected.

B. Threats to the validity and limitations of the study

In the conceptualization of this study it was acknowledged that there is probably no way to assess the variables of interest without introducing bias. It is a unique feature of the design that the key variables were simultaneously assessed by complementary approaches: abstract evaluation and data base searches on the one hand, a survey of authors on the other hand. While this special design is to some extent a compromise owing to the particular susceptibility of the study subject to various forms of bias, it offers an excellent opportunity for a detailed evaluation of the strengths and weaknesses of the methods used, and how they may affect the validity of the study results.

In view of the study results as summarized in the previous section, two main questions have to be addressed in this context: How are discrepancies between the two models explained, and how do the study findings fit into the context of the existing literature on publication bias?

1. Threats to the internal validity: data base approach

In the model based on abstract evaluation and data base searches, assessment was complete for all abstracts in the random sample. Selection bias was therefore of no concern in this part of the study. However, there is evidence for misclassification of both exposure and outcome. Specifically, the comparison between survey data and abstract evaluation results revealed substantial inconsistencies with respect to the direction of the study results. There are several possible causes for this.

First, abstracts may report preliminary findings that are not necessarily representative of the final study results. In addition, information from the abstracts was often insufficient to reliably assess the outcome, and misinterpretation is possible. The high proportion of abstracts for which the direction of results could not be unequivocally determined ("equivocal" results, overall 45%), demonstrates these difficulties.

Studies comparing the information from meeting abstracts to those in the corresponding subsequent full paper shed additional doubts on the reliability of abstracts: Weintraub et al. examined 33 papers in pediatric surgery together with their corresponding abstracts, and found changes in the title (27%), shifts in emphasis (27%) and increases in sample size (21%) (106). They also reported that many papers had "mathematically incompatible" numbers as compared to the abstracts, suggesting lack of critical review by authors for accuracy, duplications or frank mistakes at the stage of abstract submission. The problems were more pronounced in basic science where control animal groups were often smaller in

the paper as compared to the abstract. Weintraub et al. felt that abstracts had occasionally only a "faint resemblance" to the subsequent paper. On the other hand, Bellefeuille et al. found "good to excellent correlation" between the conclusions of an article and of the respective preceding meeting abstract in 15 out of 18 phase III trial reports (41). Generally, problems in the reliability of abstracts have to be appreciated, and meeting abstracts should possibly be considered as "work in progress" rather than as a correct summary of the underlying research as it would be presented in a manuscript (107).

With respect to our study, the direction of the study results as classified based on the abstract evaluation can only serve as an approximation to the eventual study outcome. Unfortunately, the additional information available from the survey can not serve as a gold standard to estimate how well the information from the abstracts represents the eventual direction of the study results, due to several reasons which will be discussed below. There is, however, no indication or reason to assume that misclassification depended on the outcome of interest (publication), especially, as this was not known at the stage of the abstract evaluation. It can, therefore, be considered non differential misclassification, which would have biased the association under study towards the null.

Unfortunately, the misclassification of the exposure in the abstract evaluation was accompanied by a misclassification of the outcome based on the data base searches. The comparison of data base search results and survey information in the sub-sample for which survey information was available, revealed errors especially in the form of under-ascertainment: 26% of the abstracts which were followed by publication had been classified as "not published". The retrieval rate[19] of 81% is very similar to that calculated for the study by De Bellefeuille et

---

[19] survey based publication rate divided by data base based publication rate - ignores false positives

al. when corrected for non-response (see above, 82%). (41). The absolute numbers in the De Bellefeuille study were too small to analyze differences between missed and identified publications (103 identified; 12 missed). It was an advantage of the large sample size in our study that even when restricted to the surveyed abstracts these analyses could be performed (266 identified; 85 missed). Fortunately, however, no factors associated with under-ascertainment were identified. This is thus another source of non differential misclassification bias, leading to a further weakening of a possible association between statistical significance of study results and publication.

In conclusion, in view of the assumed substantial non differential misclassification, the failure to detect publication bias in spite of the high power of the study does not exclude that an association is actually present. Several findings actually suggest an effect of statistical significance on publication, even though statistical significance was not achieved: In figure 5, time to publication appeared to be shorter for studies with statistically positive results, as demonstrated by Kaplan Meier curves. In addition, studies with statistically significant results were more often published in higher impact journals as compared to negative outcome studies (38% vs. 26%).

2. Threats to the internal validity: survey approach

The problems arising from the limited information from the abstracts and the under-ascertainment of publications by data base searches had been anticipated in the planning of this study. Therefore, a survey of authors was conducted to verify the information collected and to estimate the extent of misclassification.

Contrary to previous workers, we surveyed all authors, irrespective of whether publications were identified or not. We were thus able to examine not only the extent to which publications were missed (26%), but also the proportion of publications that were identified by the data base search but were not confirmed by

the author survey (8%). This latter proportion is very low considering the fact that in the data base search, some error in the linking of publications to abstracts is expected. The survey information can be considered to be reasonably reliable with respect to the outcome, although some caution may be advised.

It was hoped that authors would also give more reliable information on the exposure (direction of the study results), in particular in those cases where the assessment of statistical significance had not been possible e.g. due to limited information ("equivocal results"). Also, where preliminary results were presented in the abstract, the author information would be expected to be more representative of the study results at the stage of manuscript preparation and submission. It is indeed striking that the proportion of "equivocal results" in the abstract evaluations was much lower in the author survey (19% vs. 45%). Of the study results initially rated "positive", 90% were confirmed as such by the survey, compared to a consistent classification of negative results in 56%. These results are plausible when a high proportion of preliminary reports or poorly reported abstracts is assumed. On the other hand, it would probably be overly optimistic to consider the author information as a gold standard with respect to the exposure classification. In particular, the frequent unwillingness of authors to provide exact data on publication dates indicate a certain lack of diligence in the completion of the questionnaires. Also, the striking inconsistencies between author information and abstract evaluations with respect to abstract acceptance at the meeting are reason for concern - for this variable, a false positive rate of 57% could be calculated for the author information. Unfortunately, we had no means to objectively determine the true distribution of the exposure "statistical significance", as obviously, both sources of information were subject to misclassification. However, misclassification bias seems to have been a greater problem in the abstract and data base based analysis as compared to the survey.

Rather, the predominant problem in the survey was the low response rate. We had predicted to receive information on about 67% of abstracts, based on the idea that several authors were available per abstract. This seemed realistic based on the reports on physician cooperation with surveys where an average publication rate of 54% had been reported (46). However, unfortunately, our response rate was considerably lower (41%; 50% of abstracts covered). One reason for the low response was the unreliability of available addresses. Abstracts dated back as early as 1992 . The information given on author affiliations was often sparse. The postal strike seems to have contributed to further delays, misdirection and loss of mail. A large number of reminders was returned undeliverable while first mailings with identical addresses were not. This indicates that undeliverable mail was often not successfully redirected. Also, there were frequent requests for questionnaires when prompted by reminders. Obviously, a substantial proportion of addressees had not received the questionnaires.

The inability to determine the cooperation rate (the proportion of participants out of all people who were contacted and found to be eligible (87)) rather than a more conservative response rate (proportion of participants out of all people who were selected), is a problem of mail surveys in particular when the proportion of non-contacts is suspected to be high. In an interesting study by Sandler and Holland, up to 24% of letters to fictitious occupants of correct addresses (university departments in several centers in North Carolina) were never returned as undeliverable (13% if "reminders" were used). Incorrect addresses, on the other hand, led to 100% return as "undeliverable". The problem is even larger in an international survey, where differences in the procedures and the reliability of postal services may be very variable. The non-response proportion can therefore be suspected to be substantially inflated in our survey.

It is of note that the response was much better within the country (Canada, 62%), and for authors within the same area of interest as the senior surveyor (Dr. Suth-

erland; IBD, 67%). This suggests that national surveys in a restricted field of interest, or in a small geographically defined area, are easier to perform, and are more likely to achieve higher response rates than an international project like this. Examples, as discussed in the introduction section, include the Australian study on hospital board submissions (70%), or the follow up study on American ophthalmology abstracts (89%) (40,42).

It is not the response rate, but the extent and direction of selection bias which is the main problem when considering the effects of non response. Due to the parallel use of an alternative data collection method with 100% completeness, the magnitude and effect of a possibly existing response bias could be estimated. Bivariate analysis revealed that publication rates were higher among the responders. In contrast, response rates did not seem to be associated with statistical significance of results. However, even if study losses due to non response are random under individual hypotheses, the joint interaction between outcome, exposure, and study participation may lead to interaction, in particular, if losses are large (108). Therefore, pure outcome response bias, which would leave OR's unaffected (88), can only be assumed after examination of individual response rates (interior cell values), as illustrated in table 19. In fact, in our study, in spite of identical marginal response rates based on statistical significance, OR for publication of negative outcome studies differed significantly between responders and non responders (1.2 vs. 0.6). Based on the formula presented by Austin and Criqui (see methods section), the responder OR was found to overestimate the magnitude of the association by 43%. In the presence of a (truly) negative association, this corresponds to a distortion of the association towards the null.

This error term can only be calculated if information on outcome and exposure is complete for both responders and non responders. Therefore, to estimate the effect of response bias in the survey based analysis, the search and abstract based classifications were applied. This represents, of course, a simplification,

using the abstract based exposure classification and the search based publication rates as approximations to the survey based classifications.

Similarly, in the calculation of corrected measures of association, an identical magnitude of under-ascertainment had been assumed for publications of responders and non-responders. This assumption of absence of bias seem reasonable. However, random variation was neglected, and confidence intervals for the corrected rates must therefore be considered optimistically/ artificially narrow. Generally, all corrections resulted in only slight changes in the estimates, the important fact to note is the consistency in the direction of change (every attempt to correct for any bias resulted in lower, i.e. stronger OR's) .

3. Further limitations - both approaches

An additional problem in both approaches may have been insufficient control for confounding due to limited information on possibly important variables:

The quality of the underlying research was the variable most difficult to assess in the prediction of publication, especially as the information available from abstracts and survey was limited. The follow-up study on summary reports by I. Chalmers et al. is one of the rare examples where a summary variable for quality was included (36). However, the study failed to detect a significant association. The measure used had been restricted to very few features, and substantial problems due to the insufficient information were reported. Fortunately, we succeeded in developing a measure which could be shown to be reliable and valid for the assessment of formal abstract quality. It is, however, questionable how well the abstract represents the underlying research, or even, how comparable formal abstract quality will be to formal quality of a subsequent full paper. Previous research indicates that often considerable inconsistencies exist (41,68). This was confirmed in our study. For example, agreement for randomization, double blinding and use of placebo was poor, when comparing author survey informa-

tion to the information from abstracts. The score has therefore to be considered as an approximation for research quality. This may explain why the association between quality and publication was weak for clinical trials (OR 1.4), and was not detected in the full sample analyses based on data base searches.

Neither model found evidence for an effect of sample size. This factor had been examined by several investigators, most of which also failed to establish this variable as a predictor of publication (38-40,57). This may seem surprising, as the interdependence of sample size and publication bias would have led to the expectation of an impact of study size on chances for publication. However, the interdependence only holds within cohorts of studies with comparable interventions or associations under study. The more varied the research topics and designs, the less meaningful are differences in study size. Power, or the standard error relative to the treatment effect, would have been more meaningful. Unfortunately though, this information is rarely available, even in full reports (72). In our sample abstracts, power calculations for negative results were given in only 3% to 8% of the cases. Of interest is the follow-up study based on submissions to review boards by Ioannidis, where information about target sample sizes and problems in accrual was available (57). However, the Ioannidis study was restricted to phase II and III trials in AIDS, facilitating the comparability of sample size requirements. In this report, larger studies (n > 1,000) took longer to complete, but, once completed, were published faster (HR for publication 2.8, 95% CI 1.2 to 6.6).

Funding is another example of a variable which may be more difficult to assess in a more comprehensive study. Funding may have different implications in different research environments, and in different research areas. In particular, the associations between funding and a pressure to publish may not be comparable between different countries. Thus, a lack of association found in this international study does not contradict the importance of external funding for publication rates

of US or Australian research, as demonstrated by Dickersin et al. and by Stern and Simes (38-40).

A positive factor in the discussion of the results of both analyses is the large sample size. Initial sample size considerations took into account some losses due to non - response, as well as the option for separate analyses based on study types. Therefore, although the response rate in the survey was even lower than expected, the power for the detection of relevant publication bias was still around 80%.

Finally, one important determinant of publication was not studied in this thesis, nor by any other study on publication bias: scientific quality beyond formal methodology and completeness of reporting. The originality of the research question, the relevance of the project for science and/or clinical medicine and the ethical dimensions involved can only be assessed by content experts. These characteristics are hopefully relevant in the chances for publication of a project. Whether aspects of this may be associated with features like sample size and, consequently, statistical significance of study findings and susceptibility to publication bias, is not clear. Confounding can thus not be excluded completely.

4. Problems affecting the generalizability of studies on publication bias

Some additional aspects have to be considered when assessing the generalizability of the results, and their interpretation in the light of the existing literature. An important problem is an inconsistency in the definitions used, in particular with respect to the primary exposure of interest (direction of study results), as discussed above. An overemphasis on p-values has been suggested to be at the root of the problem "publication bias" (2). Most definitions of publication bias are therefore based on the presence or absence of statistical significance (1,8). They are, in addition, less likely to be subject to subjective error as compared to classifications based on perceived difference, as used by I. Chalmers et al. (36).

Some investigators of publication bias have used "trend" or tendency as additional categories, e.g. defined by a p value between 0.05 and 0.1 (40). This approach was not feasible in our study, as information in the abstracts is limited and p-values exceeding 0.05 are usually not reported (and remembered) quantitatively. In other previously reported studies, a dichotomous classification has been used, with negative outcome trials comprising any study without statistically significant results (38,42). "Negative studies" in this case would thus include all studies where the direction of the study results could not be determined due to insufficient reporting or use of multiple outcomes with mixed results (termed "equivocal" in our study). "Negativity" in these cases, therefore, seems to reflect formal quality rather than statistical significance. Assuming that quality is a predictor of publication, the resulting misclassification would inflate the association between direction of outcome and publication, in particular when the proportion of studies with equivocal results is as large as in our cohort. We were able to demonstrate that studies with equivocal results are less likely to be of good formal quality as compared to studies with either negative or positive results. In addition, based on the survey the majority of equivocal studies was later reclassified as positive outcome studies. The distinction between negative and equivocal results, was therefore maintained. This has possibly resulted in a more conservative estimation of the effect as compared to other studies.

In addition, the use of logistic regression rather than survival analysis may have led to a loss in sensitivity, as time to publication has been suggested to be associated with a delay in publication. In fact, the discussion around the so called "time lag bias" seems to represent a shift in focus which has occurred during the last year. Both the Australian study by Stern and Simes in the BMJ in September 97 and the study by Ioannidis in JAMA (January 98) (40,57) based their conclusions on differences in the median time to publication. In our study, the low overall publication rate prevented us from calculating meaningful estimates for median time to remission. A standard reference text for the use of survival analysis

clearly discourages the use of median time to event comparisons in these circumstances (109)[20], and it is therefore surprising how much emphasis was put on the finding of different survival times in other studies (40,57,58,110,111). In deed, while actual publication rates are not reported in the studies by Ioannidis and Stern and Simes, medians were reported as "not reached", or with confidence intervals including infinity, shedding doubts about the appropriateness of using this measure. In our analysis, we found the median time to publication for clinical trials with negative results to be longer than the upper limit of the confidence interval for positive outcome trials. In analogy to the reports discussed above, this could be interpreted as evidence for a time lag bias. We do not, however, consider this a distinct phenomenon, but rather, a *built-in feature for the time dimension of publication bias* (111).

5. Limitations of the study: summary and conclusions

In summary, both approaches to the study of publication bias (survey and data base based analysis) were subject to methodological problems that may have introduced bias. Our design, with its double approach, offered the rare opportunity to study and to a certain degree quantify the susceptibility of the different methods to different sources of bias. However, it is difficult to assess the degree to which the methodological problems may have influenced the study results. In the case of the data base based analysis, the main concern was misclassification, which was expected to have weakened the effect, perhaps explaining the lack of statistical evidence for the association under study. In the survey, while

---

[20] *"Median survival times are very unreliable unless the death rate around the time of the median survival is still high. Even in quite extensive data, median survival times time can be very inaccurate. Although median survival times are widely cited, they should therefore be treated with great caution, except for diseases in which nearly everyone dies, the data are extensive and the life table falls rapidly through the whole region between 70% and 30% alive [...]. Average survival times can be far worse and should almost never be cited"* (in this case "survival" times would refer to time to publication as opposed to time to death).

some misclassification of exposure can not be excluded, the main problem was the low response rate. There is some indication that selection bias, if present, was directed towards the null.

In conclusion, even though the different models based on the data base search and on the survey yielded somewhat different results, the two approaches should be seem as complementary rather than contradictory. All problems discussed most likely resulted in a reduced likelihood to detect the effects under study. There is, therefore, reason to consider the associations found to be valid.

C. Implications of the study findings and conclusions

In summary, we have found only indirect evidence for publication bias in gastroenterology, as implicated by lower acceptance rates at the DDW, slower publication rates and lower impact of journals for negative outcome studies. The phenomenon is present in all types of research, although in basic science, selection based on statistical significance seems to take place primarily at a stage prior to abstract submission as suggested by the very low percentage of statistically negative studies in the sample.

Several limitations to the methodology used in the examination of publication bias were identified and to some extent quantified. In particular, incomplete retrieval by data base based literature searches and limited information available from summary reports have to be considered in secondary data collection, while biased response rates and incomplete and sometimes unreliable information can be a problem in mail surveys. This confirms the appropriateness and necessity of the use of complementary strategies.

It was shown that caution is needed with respect to meeting abstracts. Information can not be considered very reliable, as inconsistencies with information from the survey were frequent, even and in particular with respect to the direction of the main outcome. Substantial improvements are possible with respect to the formal quality of abstracts. Addresses provided with the abstracts were found to be often insufficient to enable contact with the authors to clarify methods and findings.

The findings imply that publication bias is not likely to be of high magnitude, but has to be considered when evaluating the literature in gastroenterology. The problems concerns clinicians and clinical researchers as well as basic scientists.

VII. Bibliography

1. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990;263:1385-9.
2. Newcombe RG. Towards a reduction in publication bias. Br Med J 1987;285:656-9.
3. Begg CB, Berlin JA. Review: publication bias and dissemination of clinical research. J Natl Cancer Inst 1989;81(2):107-15.
4. Katerndahl DA. Citation bias: supporting your case in the extreme [editorial]. Fam Pract Res J 1994;14(2):107-8.
5. Gotzsche PC. Reference bias in reports of drug trials. Br Med J 1987;295:654-6.
6. Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. J Am Stat Assoc 1959;54:30-4.
7. Sterling TD, Rosenbaum WL, Weinkam JJ. Publication decisions revisited: the effect of outcome of statistical tests on the decision to publish and vice versa. Am Stat 1995;49:108-12.
8. Last JM. A dictionary of epidemiology. 3rd ed. New York, Oxford, Toronto: Oxford University Press; 1995;
9. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science. Epidemiological reviews 1992;14:154-76.
10. Friedenreich CM. Methods for pooled analyses of epidemiologic studies. Epidemiology 1993;4:295-302.
11. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. Stat Med 1989;8(2):141-51.
12. Cook DJ, Guyatt G, Ryan G, Clifton J, Buckingham L, Willan A, McIlroy W, Oxman AD. Should unpublished data be included in meta-analyses? Current convictions and controversies. JAMA 1993;269(21):2749-53.

13. Trichopoulos D, Zavitsanos X, Koutis C, Drogari P, Proukakis C, Petridov E. The victims of Chernobyl in Greece: induced abortions after the accident. Br Med J 1987;295:1100
14. Koren G. Bias against the null hypothesis in maternal-fetal pharmacology and toxicology. Clin Pharmacol Ther 1997;62(1):1-5.
15. Bero LA, Glantz SA, Rennie D. Publication bias and public health policy on environmental tobacco smoke. JAMA 1994;272(2):133-6.
16. Mahoney MJ. Publication prejudices: an experimental study of confirmatory bias in the peer review system. Cog Ther Res 1977;1:161-75.
17. Wynder EL, Higgins IT, Harris RE. The wish bias. J Clin Epidemiol 1990;43:619-21.
18. Chalmers TC, Frank CS, Reitman D. Minimizing the three stages of publication bias. JAMA 1990;263(10):1392-5.
19. Goodman SN. p-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993;137(5):485-96.
20. Walter SD. Perman JA, Rey J (eds.). Statistical Analysis: significance tests versus confidence intervals. In: Clinical trials in infant nutrition. Philadelphia: Lippincott-Raven Publishers; 1998; p. 47-65.
21. Comroe JHJ. Publish and/or perish. Am Rev Resp Dis 1976;113:561-5.
22. Simes RJ. Confronting publication bias: a cohort design for meta-analysis. Stat Med 1987;6(1):11-29.
23. Light RS; Pillemer DB. Summing up: the science of reviewing research. Cambridge: Harvard University Press; 1984. 65p.
24. Chalmers I. Underreporting research is scientific misconduct. JAMA 1990;263:1405-8.
25. Maxwell C. Clinical trials, reviews and the journal of negative results. Br J Clin Pharmacol 1981;1:15-8.
26. Sharp DW. What can and should be done to reduce publication bias. The perspective of an editor. JAMA 1990;263:1390-1.

27. de Melker HE, Rosendaal FR, Vandenbroucke JP. Is publication bias a medical problem? Lancet 1993;342:621
28. Minerva. News and notes. Views. Br Med J 1983;287:1886
29. Vandenbroucke JP, Rosendaal FR. Publication bias [letter; comment]. Lancet 1994;343(8889):119
30. Chalmers TC, Koff RS, Grady GF. A note on fatality in serum hepatitis. Gastroenterology 1965;49:22-6.
31. Auperin A, Pignon J-P, Poynard T. Review article: critical review of meta-analyses of randomized clinical trials in hepatogastroenterology. Al Pharmacol Ther 1997;11:215-25.
32. Chalmers TC, Lau J. Randomized control trials and meta-analyses in gastroenterology: major achievements and future potential. [Review]. Ann NY Acad Sci 1993;703:96-105;:discussion 105-.
33. Gieser LJ, Olkin I. Models for estimating the number of unpublished studies. Stat Med 1996;15:2493-507.
34. Begg CB. A measure to aid in the interpretation of published clinical trials. Stat Med 1985;4:1-9.
35. Begg CB, Mazumbar M, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088-101.
36. Chalmers I, Adams M, Dickersin K, Hetherington J, Tarnow-Mordi W, Meinert C, Tonascia s, Chalmers TC. A cohort study of summary reports of controlled trials. JAMA 1990;263:1401-5.
37. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991;337:867-72.
38. Dickersin K, Min YI, Meinert C, Meinert CL. Factors influencing publication of research results: follow up of applications submitted to two institutional review boards. JAMA 1992;267:374-8.
39. Dickersin K, Min YI. Publication bias: the problem that won't go away. Ann NY Acad Sci 1993;703:135-46.

40. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. Br Med J 1997;315:640-5.
41. De Bellefeuille C, Morrison CA, Tannock IF. The fate of abstracts submitted to a cancer meeting: factors which influence presentation and subsequent publication. Ann Oncol 1992;3:187-91.
42. Scherer RW, Dickersin K, Langenberg P. Full publication or results initially presented in abstracts. A meta-analysis. JAMA 1994;272(2):158-62.
43. Easterbrook P, Berlin J. Meta-analysis [letter]. Lancet 1993;341(8850):965
44. Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Perusing the literature: comparison of MEDLINE searching with a perinatal trials database. Control Clin Trials 1985;6:306-17.
45. Woods D, Trewheellar K. MEDLINE and Embase complement each other in literature searches [letter]. Br Med J 1998;316:1166
46. Asch DA, Jedrziewski MK, Christakis NA. Response rates to mail surveys published in medical journals. J Clin Epidemiol 1997;50(10):1129-36.
47. Hay DA. A mail survey of health care professionals: an analysis of the response. J Can Chiropr Assoc 1996;40(3):162-8.
48. Sommer B. The file drawer effect and publication rates in menstrual cycle research. Psychol Women Q 1987;11:233-42.
49. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith HJ. Publication bias and clinical trials. Control Clin Trials 1987;8:343-53.
50. Meranze J, Ellison N, Greenshow DE. Publications resulting form anesthesia meeting abstracts. Anesth Analgesia 1982;61:445-8.
51. Yentis SM, Campbell FA, Lerman J. Publication of abstracts presented at anaesthesia meetings. Can J Anaesth 1993;40(7):632-4.
52. Juzych MS, Shin DH, Coffey JB, Parrow KA, Tsai CS, Briggs KS. Pattern of publication of ophthalmologic abstracts in peer-reviewed journals. Ophthalmology 1991;98:553-6.

53. Juzych MS, Shin DH, Coffey J, Juzych L, Shin D. Whatever happened to abstracts from different sections of the association for research in vision and ophthalmology? Inv Ophthalmol Vis Sci 1993;34(5):1879-82.
54. McCormick MC, Holmes JH. Publication of research presented at the pediatric meetings. Change in selection. Am J Dis Child 1985;139(2):122-6.
55. Goldman L, Loscalzo A. Fate of cardiology research originally published in abstract form. N Engl J Med 1980;303:255-9.
56. Duchini A, Genta RM. From abstract to peer-reviewed article: the fate of abstracts submitted to the DDW. [Abstract] Gastroenterology 1997;112:(4)A12
57. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 1998;279(4):281-6.
58. Jadad AR, Rennie D. The randomized controlled trial gets a middle-aged checkup [editorial; comment]. JAMA 1998;279(4):319-20.
59. Moscati R, Jehle D, Ellis D, Fiorello A, Landi M. Positive-outcome bias: comparison of emergency medicine and general medicine literatures. Acad Emerg Med 1994;1(3):267-71.
60. Sacks HS, Chalmers TC, Smith H. Sensitivity and specificity of clinical trials: randomized vs. historical cohorts. Arch Intern Med 1983;143:753-5.
61. Berlin JA, Begg CB, Louis TA. An assessment of publication bias using a sample of published clinical trials. J Am Stat Assoc 1989;84:381-92.
62. Chalmers TC, Smith HJ, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. Control Clin Trials 1981;2:31-49.
63. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16(1):62-73.

64. Morris RD, Audet A-M, Angelillo IF, Chalmers TC, Mosteller F. Chlorination, chlorination by-products, and cancer: a meta-analysis. Am J Public Health 1992;82(7):955-63.

65. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 1994;140(3):290-6.

66. Greenland S. Quality scores are useless and potentially misleading. Reply to "Re: A critical look at some popular analytic methods". Am J Epidemiol 1994;140(3):300-1.

67. Panush RS, Delafuentte JC, Connelly CS, Edwards NL, Greer JM, Longley S, Bennett F. Profile of a meeting: how abstracts are written and reviewed. J Rheumatol 1989;16:145-7.

68. Relman AS. News reports of medical meetings: how reliable are abstracts? N Engl J Med 1980;303:277-8.

69. Haynes RB. More informative abstracts: current status and evaluation. J Clin Epidemiol 1993;46(7):595-7.

70. Taddio A, Pain T, Fassos FF, Boon H, Ilersich AL, Einarson TR. Quality of non structured and structured abstracts of original research abstracts in the *British Medical Journal*, the *Canadian Medical Association Journal* and the *Journal of the American Medical Association*. Can Med Ass J 1994;150:1611-5.

71. Powell-Tuck J, MacRae KD, Lennard-Jones JE, Parkins RA. A defence of the small clinical trial: evaluation of three gastroenterological studies. Br Med J 1986;292:599-602.

72. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272(2):122-4.

73. Rothman KJ; Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott-Raven; 1998.

74. Gregoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? J Clin Epidemiol 1995;48(1):159-63.

75. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, Le Lorier J, Liberati A, Linde K, Penna A. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet 1996;347(8998):363-6.
76. Campbell FM. National bias: a comparison of citation practices by health professionals. Bull Med Libr Assoc 1990;78(4):376-82.
77. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. Control Clin Trials 1998;19:159-66.
78. Egger M, Zellweger-Zaehner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997;350:326-9.
79. Fisher M, Friedman SB, Strauss B. The effects of blinding on acceptance of research papers by peer review. JAMA 1994;272(2):143-6.
80. Garfunkel JM, Ulshen MH, Hamrick HJ, Lawson EE. Effect of institutional prestige on reviewers' recommendations and editorial decisions. JAMA 1994;272(2):137-8.
81. Davidson RA. Source of funding and outcome of clinical trials. J Gen Intern Med 1986;155-8.
82. Gilbert JR, Williams ES, Lundberg GD. Is there gender bias in JAMA's peer review process? JAMA 1994;272(2):139-42.
83. AASLD, AGA, ASGE, SSAT. Program notices. DDW: Program, May 17-20, 1998, New Orleans, LA 1998;12-4.
84. Yahoo. Online Medical Dictionary. 1998; (Anonymous)
85. Dickersin K. Confusion about "negative" studies. N Engl J Med 1990;322(15):1084
86. Friedman LM; Furberg CD; DeMets DL. Fundamentals of clinical trials. 3rd ed. St. Louis: Mosby; 1996.

87. Slattery ML, Edwards SL, Caan BJ, Kerber RA, Potter JD. Response rates among control subjects in case-control studies. Ann Epidemiol 1995;5:245-9.

88. Criqui MH. Response bias and risk ratios in epidemiologic studies. Am J Epidemiol 1979;109(4):394-9.

89. Garfield E. The impact factor. ISI-essays 1994;http://(www.isinet.com/): essay7.html

90. The Cochrane Collaboration. The Cochrane Library. The Cochrane Collaboration & Update Software Ltd. 1998/II

91. Garfield E. SCI journal citation reports; a bibliometric analysis of science journals in the ISI database. Philadelphia: Institute for Scientific Information, Inc. 1995.

92. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. JAMA 1994;272(2):101-4.

93. Froom P, Froom J. Deficiencies in structured medical abstracts. J Clin Epidemiol 1993;46(7):591-4.

94. Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariate methods. 2nd ed. Belmont, California: Duxbary Press; 1988;

95. Feinstein AR. The theory and evaluation of sensibility. In: Feinstein AR. Clinimetrics. New Haven, London: Yale University Press; 1987; p. 141-66.

96. Oxman AD, Guyatt G. Validation of an index of the quality of review articles. J Clin Epidemiol 1991;11:1271-8.

97. Streiner DL, Norman GR. Health Measurement Scales. A practical guide to their development and use. Oxford New York Tokyo: Oxford University Press; 1989; 8, Reliability. p. 79-96.

98. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33:613-9.

99. Seigel DG, Podgor MJ, Remaley NA. Acceptable values of kappa for comparison of two groups. Am J Epidemiol 1992;135(5):571-8.
100. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991.
101. SPSS for Windows [computer program]. SPSS Inc. 7.0. SPSS Inc; 1995;
102. Austin MA, Criqui MH, Barrett-Connor E, Holdbrook MJ. The effect of response bias on the odds ratio. Am J Epidemiol 1981;114:137-43.
103. Christensen E. Multivariate survival analysis using Cox's regression model. Hepatology 1987;7:1346-58.
104. Greenland S, Rothman KJ. Fundamentals of epidemiologic data analysis. Handling of missing data. In: Rothman KJ, Greenland S (eds.) Modern Epidemiology. 2nd ed. Philadelphia, Lippincott-Raven; 1998p. 207-8.
105. Frank E. Authors' criteria for selecting journals. JAMA 1994;272(2):163-4.
106. Weintraub WH. Are published manuscripts representative of the surgical meeting abstracts? An objective appraisal. J Ped Surg 1987;22(1):11-3.
107. Soffer A. Beware the 200-word abstract! Arch Intern Med 1976;136:1232-3.
108. Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol 1977;106(3):184187
109. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient II. Analysis and examples. Br J Cancer 1977;35:1-39.
110. Clarke M. Time lag bias in publishing clinical trials [letter]. JAMA 1998;279(24):1952
111. Arida AÜ. Time lag bias in publishing clinical trials [letter]. JAMA 1998;279(24):1952

Appendix A: Literature search - comparison between data bases

Initially, for every 10th abstract, all authors were searched rather than just first and last. No additional publications were identified by this, while the number of articles for screening became unreasonably high. The strategy was therefore dropped (after 35 abstracts).

Inclusion of the year preceding the meeting identified six "precocious" publications (<1% of 692 abstracts), the earliest having been published 13 months before. Precocious publications were most often on OCR (4/6), by Germany-based authors (3/6) and in national, low impact journals (4/6). None was followed by another publication after the meeting.

Authors of 308 abstracts were searched using both MEDLINE and Embase. Of the130 publications identified, 12 were missed by MEDLINE, seven by Embase. This translated into an increase in publication rate from 38.3% for MEDLINE only, or 39.9% for Embase only, to 42.2% in the combined effort. The difference in sensitivity was statistically not significant (McNemar Test, $p = 0.4$).

Table 30 gives the names of journals that were missed. Journals missed by MEDLINE included both European and North-American based journals, and were generally of relatively low impact (median 0.3, maximum 1.1). Journals missed by Embase were more erratic, including high impact journals (median 2.1, maximum 5.4). It is of note, that in comparison to MEDLINE, Embase was particularly good for the identification of publications on clinical research, in particular clinical trials (six additional retrievals, none missed), but poorer on BSS (one additional, four missed).

Table 30: Journals missed by MEDLINE, found in Embase

	n	language; listed in MEDLINE
Eur J Gastroenterol Hepatol	4	English; listed since 1995
Can J Gastroenterol	2	English; listed since 1996
Leber Magen Darm	1	German; listed
Wien Klin Wochenschr	1	German; listed
Gullet	1	English; not listed
Med Chirurg Dig	1	French; listed
Probl Gen Surg	1	English; not listed
Int J Oncol	1	English; listed since 1998
Total	12	

Cancerlit search was performed along with Embase, and identified one more publication (Italian Journal of Gastroenterology), while generally being very incomplete on non-oncological topics (missed 118 of 131 publications). Similarly, using only HealthStar or Cinahl, publication rates were very low (10.3% and 5.1%) as compared to 56.4% for both MEDLINE and Embase in this subset. No publications were identified in addition to those found using the major data bases. Search of these databases was therefore abandoned.

The Cochrane library data base of clinical trials was searched for 1995 CCT abstracts (n = 82). In comparison to the combined Embase/MEDLINE search, three publications were missed, while two additional publications were identified (*Gastroenterology*, *Schweizer Medizinische Wochenschrift*). For this subset, the cumulative publication rates were 45.7% for MEDLINE only, 48.2 for MEDLINE and Embase, and 50.6 for MEDLINE, Embase and The Cochrane Library. The publications missed by the Cochrane Library were all in low impact journals (*Leber Magen Darm, J Am College Surg, Hepat-Gastroenterol*).

Biosis was searched for 121 1995 abstracts (64 clinical, 57 basic science). As compared to MEDLINE, no additional publication was identified, while 12 were missed out of 23 clinical, and 5 out of 19 basic science reports.

Table 31: Cumulative publication rates by database: MEDLINE + Embase

	number	MEDLINE	MEDLINE + Embase
All	308	38.3 (118)	42.2 (123)
OCR & CCT	243	39.5 (96)	44.0 (107)
BSS	65	33.8 (22)	35.4 (23)

Appendix B: Responders' comments

1. How investigators like and do not like to be surveyed:

- *"I would be very interested to know of the outcome of your survey. Could you send me a reprint, please? [...] What really hurts is when a solid-piece of work, when completed, is no longer "newsy" for fashions have changed [.. The article] was rejected because it did not focus on H. pylori..."*
- *"this is a most interesting study which you have started and I look forward to hearing the results of it in due course"*
- *"I hate questionnaires and you can fill out the rest, otherwise you have not done enough work for your research"*
- *"make your own research instead of asking others to work for you"*
- *"sorry, I am inundated with such questionnaires and simply don't have the time - most of the information should be available on MEDLINE"*
- *"I discussed your proposal with my team and we decided not to participate"*

2. Reasons for non-publication or delay:

- *"Dr. XY has all data"*
- *"funding inadequate", "project dropped by sponsor", "veto of sponsor"*
- *"disappointed that abstract was not accepted at AGA",*
- *"cavalier approach of AGA to acceptance of study"*
- *"We are in a private hospital and we're just fooling around. I have no real interest in publishing papers"*
- *"data outdated now", "results no longer relevant"*
- *"project extended", "abstract was minor part of the more comprehensive paper", "parts of the work are published in different articles", "study was essentially a pilot for ongoing work"*
- *Absolutely maddening! The analysis took even longer than the trial, writing up still longer!!!*

3. Perceived reasons for rejection:

- *"too many words and tables"*
- *"editorial biases"*
- *"reviewer did not like the results"*
- *"research contradicts previously published results by other well known author"*
- *"[project] important because it is a large study of natural history and is unlikely ever to be repeated. [However,] lack of interest because nothing to do with H. pylori"*

Appendix C: Abstract evaluation form

Reference number: Referee: Date:

Population studied: 1☐ patients 2☐ healthy volunteers
3☐ animals 4☐ tissue/cell-cultures etc. 99☐ other/n/a

Area: 1☐ therapy 2☐ diagnosis 3☐ epidemiology 4☐ physiology
5☐ basic science 6☐ economy 7☐ data-management 99☐ other/n/a

Study design

interventional, human:		*observational, human:*	
1☐ parallel controlled trial	4	6 ☐ cohort, prospective	4
2☐ cross over trial	3	7 ☐ cohort, retrospective	3
3☐ time series trial (before-after)	2	8 ☐ cross-sectional	3
4☐ non-concurrent controls	2	9 ☐ case-control	3
5☐ natural experiment	2	10☐ descript., case report/series	1
basic science study:		*other:*	
11☐ interventional, parallel	4	16☐ meta analysis	2
12☐ interventional, cross over trial	3	17☐ instrument validation	1
13☐ interventional, before-after	2	18☐ literature review	1
14☐ observational, comparative	3	19☐ other:	0
15☐ observ., non comp. (case series)	1	99☐ don't know / not sure	0

Randomization reported? 1☐ yes (+1) Placebo used? 1☐ yes

Overall sample size: Number of groups:

Study question:

Direction of results:

Statistical significance	*Direction of results*
0☐ no	0☐ negative
1☐ yes	1☐ positive
2☐ mixed	2☐ mixed
3☐ not reported, not clear	3☐ not reported, not clear
9☐ not applicable	9☐ not applicable

Source of funding: 0☐ not funded /not reported
1☐ governmental agency 3☐ private / health charity
2☐ pharmaceutical company 4☐ other or not classifiable

Therapeutic/diagnostic trials only: Intervention tested _____

1 ☐ Antacids and antiflatulents
2 ☐ H2 blockers
3 ☐ Proton pump inhibitors
4 ☐ Prostaglandins, sucralfate antimuscarinergics (pirenzipine)
5 ☐ Prokinetic/antispasmodic, antiemetic, antidiarrhoial
6 ☐ Enzymes and digestants
7 ☐ Gallstone dissolution agents
8 ☐ Foods, TPN, food formulas / supplements / exclusion
11☐ Corticosteroids
12☐ Other immunosuppressants
13☐ Antibiotics
14☐ antiviral, antimycotic antiparasitic agents
15☐ Antineoplastics
16☐ NSAIDS, aspirin
17☐ Narcotics
18☐ Other analgesics
19☐ Cardiovascular agents
20☐ Diabetes agents
21☐ (Other) hormones
22☐ Minerals, vitamins
23☐ Plasma extenders, fractions (e.g. alb, Ig)
24☐ Psychotropics

9 □ 5 ASA, SSP	25□ Endoscopic intervention	27□ Diagnostics
10□ Laxatives	26□ Surgical intervention	28□ other:

Quality assessment	yes	partial	no	n/a
1. Question / objective sufficiently described?				
2. Design appropriate to answer study question?				
3. Subject characteristics sufficiently described?				
4. Subjects appropriate to the study question?				
5. Control subjects appropriate **(if no control, check no)**				
6. Method of subject selection described and appropriate?				
7. If random allocation to treatment groups was possible, is it sufficiently described? (if not possible, check n/a)				
8. If blinding of investigators to intervention was possible, is it done/ reported? (If not possible, n/a)				
9. If blinding of subjects to intervention was possible, is it done/ reported? (If not possible, n/a)				
10. Outcome measure well defined and robust to measurement bias?				
11. Confounding accounted for?				
12. Sample size adequate?				
13. Post hoc power calculations or confidence intervals reported for statistically non significant results?				
14. Statistical analyses appropriate?				
15. Statistical tests stated?				
16. Exact p-values or confidence intervals stated?				
17. Attrition of subjects and reason for attrition recorded?				
18. Results reported in sufficient detail?				
19. Do the results support the conclusions?				
Sum				
times	2	1	0	2
Total				

Points for design (+1 if randomized) ☐
Points items 1 - 19 ☐

Total sum achieved ☐
Total possible sum = 43 - [number of n/a * 2] = ☐

Score: total / total possible ☐

Blinding of abstract evaluation: 0□ broken 1□ partial 2□ complete

Appendix D: Letter to authors

«Field1»
«Mailing_List_ID»-«Ref1»«Ref2»

Dr. «Initials» «Last_Name»
«Department»
«Organization_Name»
«Address»
«City», «Country»

Dear Dr. «Last_Name»,

We are performing a follow-up survey on GI research initially presented as abstracts. Publication rates, time to publication and, where applicable, reasons for non-publication will be studied.

You submitted the attached abstract to the «Year1» meeting of the American Gastroenterological Association. We are interested in the further progress of your research project, in particular whether it has since been published as a full report. May we, therefore, ask you or one of your co-workers to complete the included questionnaire - this takes approximately five to ten minutes. An information sheet with definitions is enclosed to ensure uniformity of data - please consult these if you are unsure of the terminology.

The information you give in the questionnaire will be treated confidentially. Details will not be published in combination with your name or with the title of your research project. By sending the completed questionnaire back to us you consent to our use of the data in the context of our project.

Please return the completed questionnaire as soon as possible. A return envelope is included. Feel free to contact us for any further questions or comments. Your help is greatly appreciated.

Sincerely,

Lloyd R. Sutherland, MDCM, MSc, FRCP, FACP
Head, Department of Community Health Sciences
Professor, Faculty of Medicine

Antje Timmer, MD

Appendix E: Author questionnaire

A. Abstract presentation

1. What was your role in the research presented?
   ☐ principal/senior investigator ☐ co- investigator ☐ other: _____

2. How was the abstract presented at the DDW-meeting?
   ☐ poster ☐ talk ☐ no presentation

3. Were you listed as an author on the abstract?
   ☐ first author ☐ last author ☐ other author ☐ no author

B. Study design

1. Please check the study design of the research project you presented (check one only – please consult information sheet)

interventional, human:
- ☐ parallel controlled trial
- ☐ cross over trial
- ☐ time series trial (before-after)
- ☐ non-concurrent / historic controls
- ☐ natural experiment

basic science study:
- ☐ interventional, parallel controlled
- ☐ interventional, cross over trial
- ☐ interventional, before-after (time series)
- ☐ observational, comparative
- ☐ descriptive (non comparative)

observational, human:
- ☐ cohort, prospective
- ☐ cohort, retrospective
- ☐ cross-sectional
- ☐ case-control
- ☐ descriptive, case report/series

other:
- ☐ meta analysis
- ☐ instrument validation
- ☐ literature review
- ☐ don't know / not sure
- ☐ other: _____

Questions 2 - 6 refer to controlled interventional studies only

2. Was randomization used for intervention allocation? ☐ yes ☐ no/not applicable

3. If randomization was used to assign a treatment: what method was used?
   ☐ random numbers (PC or tables) ☐ sequential envelopes/prepacked
   ☐ phone call /central randomization ☐ other: _____

4. Were placebo or sham controls used? ☐ yes ☐ no / not applicable

5. Were investigators blinded for intervention allocation or exposure under study?
   ☐ yes ☐ no / not applicable

6. Studies on humans only: were subjects blinded to the intervention or exposure under study?
   ☐ yes ☐ no / not applicable

All studies:

7. What was the overall sample size? ☐

8. How many groups were compared? ☐

9. Were the results statistically significant with respect to the main outcome? (i.e. p ≤.05 or 95%-confidence interval excluding reference value)
   ☐ yes ☐ no ☐ mixed ☐ not applicable

10. Irrespective of the statistical significance, how do you rate the direction of your study results?
    ☐ positive, that is the effect or association under study is (probably) present
    ☐ negative ☐ mixed ☐ not applicable

11. Irrespective of the direction of the results, how would you rate the clinical and/or scientific relevance of your research question on a scale of 1 to 10?
    (1= not important at all, 10 = extremely important)

    ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9 ☐10

C. Study organization

1. How many study centers were involved in the study? ☐ 1 ☐ 2-5 ☐ >5

2. How was the study financed? ☐ governmental agency ☐ industry
   ☐ private /health charity ☐ departm./ no funding
   ☐ other: _____

3. **Drug studies only:** was the study registered in a clinical trial registry?
   ☐ yes ☐ no If yes, please specify registry: _____

4. In how many peer-reviewed publications had you been involved as an author before the presentation of this abstract? ☐ 0 ☐ 1-5 ☐ 6-10 ☐ >10

5. If you were not the official principal (senior) investigator ("PI") of this project: in how many publications had the PI been involved as an author before the presentation of this abstract?
   ☐ 0 ☐ 1-5 ☐ 6-10 ☐ >10 ☐ n/a

D. Current status of research project

1. Please indicate the current status of the research project:

Experimental/ data collection	☐ completed ...	☐ ongoing	☐ abandoned
Data analysis.....	☐ completed ...	☐ ongoing.	☐ abandoned
Manuscript preparation.....	☐ completed ...	☐ ongoing	☐ abandoned
(First) manuscript submission...	☐ completed ...	☐ ongoing.	☐ abandoned
(First) publication.....	☐ completed ...	☐ ongoing.	☐ abandoned

2. Please state date of completion for each item, where applicable:

Data analysis complete	(month)	_____	(year)	_____
Manuscript accepted	(month)	_____	(year)	_____

3. If the project is not completed or submitted, what was the reason? (please decide on one main reason only)

Lack of time	☐ main reason	☐ partial reason
Lack of interest	☐ main reason	☐ partial reason
Limitations in methodology	☐ main reason	☐ partial reason
Sample size / recruitment problems	☐ main reason	☐ partial reason
Side effects / ethical problems	☐ main reason	☐ partial reason
Negative / null results	☐ main reason	☐ partial reason

Unimportant results	☐ main reason	☐ partial reason
Manuscript rejection anticipated	☐ main reason	☐ partial reason
No control over data/external problems	☐ main reason	☐ partial reason
Publication not aim of the study	☐ main reason	☐ partial reason
Co-investigator left	☐ main reason	☐ partial reason
Equipment/software problems	☐ main reason	☐ partial reason
Research still in progress	☐ yes	☐ no

Other: _____

E. Submission (skip this section, if no manuscript submission)

1. To how many different journals was the manuscript submitted ? ☐

2. Was the manuscript submitted to at least one journal in English language?
   ☐ yes ☐ no

3. If rejected at least once with or without option to resubmit, were reasons for rejection stated by any of the journals? (check as many as apply)
   - ☐ No reasons stated
   - ☐ Formal requirements not met
   - ☐ No interest in research topic
   - ☐ Redundancy, insufficient originality
   - ☐ Other: _____
   - ☐ Sample size problems
   - ☐ Other limitations in methodology
   - ☐ Negative / null results
   - ☐ Unimportant results

F. Publication (skip this section if no publication)

1. Please list all peer-reviewed publications that have arisen from the research presented as abstract:

Name of journal	Month /Year	Language of journal	Name of first author

2. Did any non peer-reviewed publications arise from the research, e.g. book chapters, symposium-proceedings? ☐ yes ☐ no If yes, please specify:

Name of medium	Month /Year	Language of publication	Name of first author

G. Demographics

1. Please indicate your sex and your age at the time of the research presentation:
   ☐ male ☐ female
   ☐ <35 years ☐ 35 - 50 years ☐ >50 years

2. If you were not the PI of this project: please indicate the sex and age of the PI:
   ☐ male ☐ female
   ☐ <35 years ☐ 35 - 50 years ☐ >50 years ☐ n/a, or unknown

3. In which country was the research performed (international studies: coordinated)? _____

Thank you very much for your cooperation!

Appendix F: Information sheet for authors (reduced in size)

All questions refer to the research presented in the abstract attached. If only part of a project was presented, e.g. only the regional results of a multi center study, that part is referred to. If the abstract combines work from several studies, refer to the one that seems to you the main part.

Section B, question 1:
Interventional: the study is prospective and employs an intervention. An intervention is defined as a maneuver intended to change the status the study subject/object is in. Manipulations applied to identify existing features/states, e.g. staining, are not considered an intervention.

Comparative or controlled means the study aims at comparing groups with respect to the effects of an intervention, exposure or characteristic. If several groups or interventions are studied, but the primary goal is the effect within the groups/interventions rather than between, e.g. studying in vitro and in vivo effects, the study is not considered comparative (see time series studies). A cross-over design is not considered a comparative trial, but categorized as "cross-over" (see below).

Studies on humans
Parallel controlled: the subjects receiving the intervention are compared with at least on other group. A cross-over design is not considered a controlled trial, but categorized as "cross-over" (see below).

Cross-over study: two sets of subjects receive interventions A and B (or more) respectively, then switch interventions, so that A receives B and B receives A.

Time series trial: Outcomes are measured before and after the intervention in the same subject, i.e. the subject serves as his/her/its own control. Several time series trials may be performed in a parallel set up – before-after analysis are performed on each arm separately rather than comparisons between the arms.

Non concurrent/historic controls: results are compared to results of earlier studies, using statistical analysis

Natural experiment: the intervention is not applied by the investigators, e.g. a study may examine the effects of a law requiring to wear a helmet.

Prospective cohort: groups of subjects are followed over time. Outcomes are measured which occurred after the study has begun. The groups were selected / defined based on exposure status.

Retrospective cohort: groups of subjects were assembled before the study was started. Outcomes are measured in the past or present. The groups were selected / defined based on exposure status.

Cross sectional study: measurements on outcome and exposure are made at one time for each subject, without follow up. Subject selection was not based on outcome or exposure status. This design, as case-control and cohort designs, is analytical, i.e. associations between at least two variables are studied.

Case control study: cases and controls were selected by outcome, e.g. compare patients with a disease to patients without the disease. (Most motility studies fall into this category.)

Descriptive study, case report or case series: there are no control subjects and/or only results after but not before intervention are reported. May be retrospective or prospective. Includes prevalence/incidence studies.

Meta-analysis: statistical methods are used to combine the results of several studies.

Instrument validation: A test is compared to another or some other standard. Outcome is sensitivity/specificity, maybe agreement or correlation.

Literature review: does not apply statistical methods to combine results, does not present original research not published elsewhere.

(Categorization and definitions are modified from the system by Cho and Bero, JAMA 1994)

Basic science research:
Interventional, parallel controlled: the effects of an intervention in one substrate or group are compared with the effects of at least one other intervention in another group. Does not include trials, where several interventions or substrates are studied in a parallel set up, when before-after analysis is done per arm without statistical analysis of differences between the arms (time series, see below).

Cross-over study: two sets of animals or substrates receive interventions A and B (or more) respectively, then switch interventions, so that A receives B and B receives A.

Interventional, before-after (time series trial): outcomes are measured before and after the intervention in the same animal or substrate, i.e. it serves as its own control. Includes trials, where several interventions or substrates are studied in a parallel set up, when before-after analysis is done per arm without statistical analysis of differences between the arms.

Observational, comparative - corresponds to case-control, cross-sectional and cohort studies in humans. Characteristics of different groups or subjects are compared. There may be maneuvers applied for diagnostic purposes but the state of the substrate is not altered by this.

Descriptive: no control objects, no statistical analysis. Includes case studies, case series' or the description of techniques/procedures

Section B, question 7 and 8:
Choose the size of the sample at the beginning of the study, irrespective of any drop outs. If analysis and results are based on a subsample only, choose the number of the subsample.
Choose the number of groups relevant for the primary research question. Cross over trials: one group

Section B, question 9:
Statistical significance is defined here as $p \leq .05$ or 95% confidence intervals excluding reference/null value.

Section B, question 10:
positive: e.g.
- a statistically not significant trend is considered indicative of an effect/association being present
- treatment under evaluation is considered effective
- manipulation has (expected) effect
- two groups/subjects differ
- an association seems to be present
- a test/diagnostic procedure is valid, or useful, or economic, or sensitive

•in equivalence trials: effect is equivalent
negative: The negation of the above
mixed: There is a balance between positive and negative results
not applicable: Design and study question were not aiming at assessment of effects/comparisons/associations (descriptive study)

Section C, question 1
The number of study centers refers to the number of different institutions involved in the design, case recruitment and analysis of the project. Different departments within the same institution or otherwise affiliated centers are not considered separate institutions.

Section E through F:

- ***Publication refers to full publication in peer-reviewed journals only.***
- ***Full publication means that the publication contains detailed sections on methodology and results, and a critical discussion. Abstracts and letters do not qualify. Short reports qualify, if detailed enough to allow for critical discussion of methodology and results.***
- ***Peer-reviewed means, the manuscript underwent a formal evaluation process with the option of being rejected. Pre-accepted or pre-arranged publications such as book chapters or contributions to symposium proceedings can not be considered.***