Citation:

Sauerbrei, W. & Blettner, M. (2003). Issues of traditional reviews and meta-analyses of observational studies in medical research. In R. Schulze, H. Holling & D. Böhning (Eds.), *Meta-analysis: New developments and applications in medical and social sciences* (pp. 79-98). Hogrefe & Huber.

6

Issues of Traditional Reviews and Meta-Analyses of Observational Studies in Medical Research

Wilhelm Sauerbrei
Institute of Medical Biometry and Medical Informatics
University Hospital Freiburg

Maria Blettner
Department of Epidemiology and Medical Statistics
School of Public Health
University of Bielefeld

Summary

Summarizing data from several studies is an important part in medical research. Several problems of traditional review articles are known for a long time, with the consequence to demand for more systematic reviews. We will outline the rationale for meta-analyses and describe four methods to summarize data, with the emphasis on observational studies where the association of risk or prognostic factors and certain diseases are investigated. We will compare and assess several criteria for different types of overviews such as narrative review, meta-analysis from literature, meta-analysis with individual patients data, and the prospectively planned meta-analysis. We will critically discuss some examples from the literature and will show severe problems of meta-analyses based on literature data only. We argue that a reasonable and valid meta-analysis of observational studies requires in general some re-modeling of the data.

Therefore, the use of individual data is an important requirement to reach reliable conclusions on the association between the factor and the outcome of interest.

6.1 INTRODUCTION

The serious problems and questionable recommendations from traditional review articles have been shown and the necessity of more systematic reviews in a timely fashion by using statistical techniques are well known (Antman, Lau, Kupelnick, Mosteller, & Chalmers, 1992). Therefore, much attention has been given in recent years to meta-analysis in medical research, however, numerous methodological issues particularly with respect to biases and the use of meta-analysis are still raising controversial discussions (Chalmers, 1991; Chalmers & Lau, 1993; Thompson & Pocock, 1991; Stewart & Parmar, 1993). Authors have heavily criticized the method as such ("If a medical treatment has an effect so recondite and obscure as to require meta-analysis to establish it, I would not be happy to have it used on me", Eysenck, 1994, p. 792) or identified poorly performed meta-analyses ("In my own review of selected meta-analyses, problems were to frequent and so serious, including bias on the part of the meta-analyst, that it was difficult to trust the overall 'best estimates' that the method often produces.", Bailar, 1997, p. 560), both resulting in some discredit of this method. Additionally, in many circumstances such as in medical decision making where modern techniques of health technology assessment (HTA) play a central role, often estimates of parameters are needed and produced by a "quick" meta-analysis. Deficiencies in the meta-analysis may transfer to unsound decisions.

The critique of meta-analysis should distinguish between three central aspects:

1. A major distinction should be made whether the results of randomized trials (RCT) or observational studies are summarized. Studies comparing a treatment given in one clinic with another treatment given in another clinic are here seen as observational studies. As parts of the experiment are under control of the investigators, these studies are often considered as quasi-experiments.

2. A second important feature is the measurement scale of the factors of interest. Different problems occur depending on whether binary, ordinal, categorical, or continuous factors are investigated. Additional problems occur with a summary assessment using meta-analytic techniques when different scales of measurement were used in the original studies.

3. The third and most often considered characteristic is whether individual patient data (often called MAP= Meta-Analysis of Patient) or published data (called MAL = Meta-Analysis of Literature) is used for the meta-analysis.

There is general acceptance that meta-analyses of RCT based on individual data give the most reliable results concerning the combination of studies. Obviously, conducting a MAP is a large study in its own which requires large effort and funds from the group starting such a project and the willingness of the investigators from the single studies to cooperate. For a detailed discussion, see Stewart and Clarke (1995). These difficulties are usually given as the main justification for a meta-analysis based on published data only. Concerning RCTs, well-conducted meta-analyses based on MAL can result in useful summaries of an effect of interest and they are an accepted instrument. For observational studies the situation is more complex as the studies are less homogeneous, for example, adjustment for confounder factors in a multivariate model is essential for each single study and those are often different between studies.

In this chapter, we will discuss merits, limitations, and difficulties of different types of systematic reviews for observational studies. The following is partly taken from a paper by Blettner, Sauerbrei, Schlehofer, Scheuchenpflug, and Friedenreich (1999), where the topic in epidemiology is considered and terminology from this field is often used. Most arguments given in this paper apply to other observational studies in clinical research.

Four different methods for summarizing the evidence are distinguished:

Review:	Qualitative summary, the narrative review article.
MAL:	Meta-Analysis of literature, that is, quantitative summary of published data.
MAP:	Meta-Analysis of patient data, that is, re-analysis of individual data of the original studies followed by a quantitative summary.
Prospective MA:	Prospectively planned, pooled analysis of several studies, where pooling is already a part of the protocol. Data collection procedures, definition of variables, questions, and hypotheses are as far as possible standardized for the individual studies.

In the literature, different terms are used for these four types and certainly some combinations are possible. It should be noted that the prospective meta-analysis differs in several respects from a classical multicentre clinical trial since often the single studies are analyzed and published separately. In many situations, the design of the studies is slightly different because of local or regional circumstances.

For the following, we summarize and compare the four different types for investigations where the influence of one or several factors on an outcome variable is investigated in non-randomized studies. First, we give the major

reasons and some general rules for conducting a meta-analysis. Then we describe the similarities and differences between the various types. We compare the advantages and limitations of the four types. We will discuss some examples of summaries of observational studies based on published data for risk factors, prognostic factors, and therapeutic factors mainly with the emphasis to demonstrate severe problems of meta-analyses from literature.

6.2 RATIONALE FOR META-ANALYSES

The main reason for conducting a review or a meta-analysis is to summarize the results of previously conducted studies which usually have inconsistent results. Such a situation may arise when the sample sizes of individual studies are too small to find stable results or if the results from single studies vary considerably. Meta-analyses are mainly used to assess the influence of weak risk factors, which may nevertheless have a large public health impact (such as passive smoking, use of contraceptives, or exposure to electromagnetic fields) or of treatment strategies whose small benefits can be worthwhile for a severe disease with a large incidence. For other issues in medical research, for example, prognostic factor studies or diagnostic studies, the use of meta analyses is increasing. Review articles investigate whether the available evidence is consistent and/or to which degree inconsistent results can be explained by random variation or by systematic differences between the design, the setting or the analysis of the study. In contrast to qualitative reviews, MAL or MAP are mainly performed to obtain a combined estimator of the quantitative effect of the risk factor such as the relative risk or risk difference. Some meta-analyses are also used to investigate more complex dose-response functions. The majority of meta-analyses conducted so far examined dichotomous (or categorical) factors. Briefly, the main reasons for conducting a meta-analysis or a review of observational studies are:

1. to assess qualitatively whether a factor has to be considered as a risk factor,
2. to provide more precise effect estimates and increased statistical power and to analyze dose-response relations,
3. to investigate the heterogeneity between different studies,
4. to generalize results of single studies,
5. to investigate rare exposure and interactions, and
6. to investigate risks associated with rare diseases.

6.3 CHARACTERIZATION AND LIMITATIONS OF THE FOUR TYPES

6.3.1 Review

Traditional narrative reviews provide a qualitative but not a quantitative assessment of published results. They are influenced by publication bias (Dickersin, 1990, 1997) and the file drawer problem (Rosenthal, 1979). If there is not an *a priori* strict protocol for the review, narrative reviews are only a subjective judgment of the included studies. However, if they are carefully done, they can give quite an extensive overview of the current state of the research within a short time frame and at low cost.

6.3.2 Meta-Analysis From Literature (MAL)

These studies are comparable to a narrative review with respect to time and cost, with the main difference being that the primary goal is to give a quantitative estimate of the effect of interest. They can be performed from published data without cooperation and without the agreement of other study groups. However, attempts may be made to obtain additional information from study coordinators, if necessary. So far, not many meta-analysts have tried this approach. There are some major limitations of this approach that have been pointed out by several authors (see e.g., Shapiro, 1994). One limitation is that publication bias is particularly important since some explorative analyses may be done and published selectively. Most likely, unexpected significant results may be selected for publication, yielding an overestimation of the effect. An additional problem is that studies may differ considerably in designs, data collection methods, and the precise definition of the factors of interest and the confounder variables. A special dilemma arises if different studies adjust for different confounding factors. No systematic investigation has been performed to determine whether the simple (crude) estimates or "best estimates" should be used for combining results of individual studies. Many aspects of the heterogeneity cannot be dealt with appropriately in such summaries. A combined estimate should not be calculated if the heterogeneity between studies is too high. However, in many publications, the problem of heterogeneity is not adequately handled. An estimate is often published although strong heterogeneity between study results was observed.

6.3.3 Meta-Analyses With (Individual) Patients Data (MAP)

Some of the problems that arise with MAL are avoidable if individual data from all investigations performed on the subject matter are available. Publication bias may be less prominent as it is possible that investigators are willing to contribute their data even if for a specific theme no analysis has been performed and no papers have yet been published. The cooperation between different researchers may help to identify studies, for example, if they are only

known to local investigators. With individual data, statistical re-analysis can be performed. This analysis can include a new unified definition of the available variables and new regression models. With a large number of patients the effect of rare exposures can be examined. New hypotheses for specific subgroups may also be investigated.

It is often argued that major barriers for MAP are the high cost and long duration, and that it requires close cooperation between researchers (Stewart & Parmar, 1993; Stewart & Clarke, 1995). Although an improvement of data quality is not possible, some errors in the data or in the statistical analysis can be corrected for. Furthermore, adjustment for confounding variables that have been delineated since the original studies were performed can be done if those covariables were originally collected. Differences between the study results can be actively discussed between study coordinators and reasons for these differences can be elucidated. In general, it is possible to estimate risk coefficients and their variance from the combined data.

6.3.4 Prospectively Planned Meta-Analysis

This type of analysis has not been called a meta-analysis, despite the fact that it has several aspects in common with MAL and MAP. Several large international case-control studies and occupational cohort studies have used this approach (e.g., Boffetta et al., 1997). The major difference is that joint planning of the data collection and analysis makes it possible to avoid large differences between the studies since many details can be planned in advance and standardized. The experience of many coordinators is used in the preparation of the new multicentre study to ensure comparability in design, data collection, data analysis, and reporting across all centres. In contrast to multicentre randomized clinical trials, more heterogeneity in the individual study centres may exist arising from differences in populations (e.g., race is not a confounder factor in Germany but in the United States) or in design (e.g., no methods of random population sampling exist in the U.S., no overall cancer registration in Germany). The costs for a new multicentre study are in general high. The planning phase can be substantial, even difficult, and the time incurred can be long. Alternatively, individual studies with a joint core protocol for questions of common interest may be performed. This allows individual researchers to set priorities and also permits some variation across studies. One disadvantage of a large multicentre study is that errors in the design can be multiplied. Moreover, once a meta-analysis has been performed it will be more difficult to justify a new individual study with the same topic.

6.4 METHODS FOR AN OVERVIEW

All types of overviews – whether quantitative or qualitative – have some steps in common that should be followed in planning and conducting it. Each indi-

vidual type has some aspects of the conduct that are different and that will be described later.

6.4.1 Steps in Performing a Meta-Analysis

Each type of overview needs a clear study protocol that describes the research question and the design, including how studies are identified and selected, the statistical methods to use, and how the results will be reported. This protocol should also include the exact definition of the disease of interest, the factors of interest, and the potential confounding variables that have to be considered. A main component of the protocol is the exact definition of the inclusion criteria for single studies. As described by Friedenreich (1993), the following steps are needed for a meta-analysis:

1. Define a clear and focused topic for the review.
2. Locate all studies (published and unpublished) that are relevant to the topic.
3. Select all studies that are relevant according to the explicit inclusion criteria.
4. Abstract necessary information from the published papers or obtain the primary data from the original investigators. Meta-analysis of published data may also include contacting the original project leaders to obtain data or information that have not been published in sufficient detail. For a MAL, agreement to use the original data is needed.
5. Tabulation of relevant elements of each study, including sample size, assessment procedures, available variables, study design, publication year, performing year, geographical setting, and so forth.
6. Define protocol for the analysis of all studies and estimate the study-specific effects (relative risks adjusted for relevant confounder variables).
7. Investigate the homogeneity of study-specific effects and determine whether these effects can be combined to perform a pooled analysis.
8. Presentation of published results, for example, graphically.
9. Investigate and reduce (if possible) the heterogeneity between studies.
10. Decide about remaining heterogeneity components: Coping with different designs, study types, confounder, and so forth.
11. Estimate a pooled effect with adequate statistical methods if the studies are efficiently homogeneous.
12. Conduct a sensitivity analysis.

Obviously, some variations are needed for the different forms, for example, for traditional reviews, in general, only the steps 1 to 4 together with a qualitative assessment are done. For a meta-analysis from published data, data abstraction will be done from publications, however, if required data are not given in the publications, contact with the project manager of the studies should be made.

6.4.2 Statistical Analysis

The statistical analysis of aggregated data from published studies was first developed in the fields of psychology and education (Glass, 1977; Smith & Glass, 1977). These methods have been adopted since the mid-1980s in medicine primarily for randomized clinical trials and are also used for observational studies. We will give a brief outline of some issues of the analysis. For more details we refer to several textbooks or to a recent tutorial paper (Normand, 1999).

6.4.2.1 Single Study Results A first step of the statistical analysis is the description of the characteristics and the results of each study. Tabulations and simple graphical methods should be employed to visualize the results of the single studies. Plotting the odds ratios and their confidence intervals (so-called forest plot) is a simple way to spot obvious differences between the study results. The Galbraith plot (Galbraith, 1994) is a more sophisticated way to investigate the heterogeneity and the contribution of each study to the overall estimate.

6.4.2.2 Heterogeneity The investigation of the heterogeneity between the different studies is a main task in each review or meta-analysis (Thompson, 1994). For the quantitative assessment of heterogeneity, several statistical tests are available (Petitti, 1994; Paul & Donner, 1989). A major limitation of formal heterogeneity tests is, however, their low statistical power to detect any heterogeneity present. Therefore, it is recommended to investigate the heterogeneity informally, for example, by comparing results from studies with different designs, within different geographical regions. In addition, graphical methods should be used to visualize heterogeneity, such as plots with single studies grouped or ordered according to special covariables as type of study, publication time, etc., or funnel plots to indicate publication bias, and radial plots (Galbraith, 1994).

In meta-analysis of literature (MAL), some sensitivity analysis can be performed to investigate the degree of heterogeneity. However, if individual data are available, the sources of heterogeneity can be investigated in some detail. Heterogeneity can be reduced, for example, by using the same statistical model for all single studies. In a prospective meta-analysis, the strategy for the statistical analysis and the definitions of variables can be determined *a priori* for all individual studies. Hence, an identical multiple regression analysis can be used in each centre. This avoids heterogeneity that could be introduced by different models.

6.4.2.3 Summarizing Effect Estimates Whether calculating a common estimate is appropriate should be decided after investigating the homogeneity of the study results. If the results vary substantially, no estimator should be presented or only estimators for selected subgroups should be calculated (e.g., combining results from case-control-studies only). Methods for pooling depend on the data available. In general, a two-step procedure has to be applied.

First, the risk estimates and variances from each study have to be abstracted (MAL) or calculated (MAP). Then, a combined estimate is obtained as a (variance based) weighted average of the individual estimates. The methods for pooling based on the 2×2 table include the approaches by Mantel-Haenszel and Peto (see Petitti, 1994, for details). If data are not available in a 2×2 table but as estimates from a more complex model (such as an adjusted relative risk estimate), the Woolf and DerSimonian-Laird approach can be adopted using the estimates and their (published or calculated) variance resulting from the regression model (DerSimonian & Laird, 1986). For these methods, variance estimates of the pooled estimator are available and allow the calculation of confidence intervals.

Usually two different statistical concepts are used for the combined estimator. In *the fixed effects model,* it is assumed that the underlying true exposure effect in each study is the same. The overall variation and, therefore, the confidence intervals will reflect only the random variation within each study but not any potential heterogeneity between the studies. If individual data is available, the pooled estimator and its variance can be obtained using regression models by incorporating an additional dummy variable for each centre. The *random effects model* incorporates variation between the studies. It is assumed that each study has its own (true) exposure effect and that there is a random distribution of these true exposure effects around a *central* effect. The observed effects from the different studies are used to estimate this distribution. In other words, the random effects model allows non-homogeneity between the effects of different studies.

The most common approach to combine the single estimates is the methods of moments given by DerSimonian and Laird (1986). The important difference is that for this model, study specific weights are calculated as a sum of the variance within the studies and a term for the variance between the studies, τ^2. The between-study variance τ^2 can also be interpreted as a measure for the heterogeneity between studies. Because of anticonservatism in case of the validity of a random effects model, Ziegler and Victor (1999) proposed a modification of the test based on DerSimonian and Laird (1986). The new proposal holds the nominal level asymptotically.

Comparison between fixed effects and random effects model

- Random effects methods yield (in general) larger variance and confidence intervals than fixed effects models because a between-study component τ^2 is added to the variance.
- If the heterogeneity between the studies is large, τ^2 will dominate the weights and all studies will be weighted more equally (in random effects model weight decreases for larger studies compared to the fixed effects model)
- A major critique of the random effects model is that it is not sufficient to "explain" the heterogeneity between studies, since the random effect merely quantifies unexplained variation by estimating it (e.g., Mengersen,

Tweedie, & Biggerstaff, 1995). Heterogeneity between studies should yield careful investigation of the sources of the differences. If a sufficient number of different studies are available, further analyses, such as "meta-regression", may be used to examine the sources of heterogeneity (Greenland, 1987, 1994).

If individual data is available, the fixed effects estimate can be calculated from a regression model with dummy variables. So far, there is no comparable approach available for the random effects model. Here, the two-step procedure is used even with individual data available (e.g., Lubin et al., 1995).

Several other methods have been proposed to estimate the overall effect based on maximum-likelihood methods or on Bayesian methods (DuMouchel, 1990; Smith, Spiegelhalter, & Thomas, 1995). Recent investigations have demonstrated that, for practical purposes, the differences between these methods are not very large. So far, only rather sophisticated software is available for these approaches (Spiegelhalter, Thomas, Best, & Gilks, 1996).

6.4.2.4 Sensitivity Analysis An important method for investigating heterogeneity is sensitivity analysis, for example, to calculate pooled estimators only for subgroups of studies (according to study type, quality of the study, period of publication, etc.) to investigate variations of the odds ratio. An extension of this is meta-regression as proposed by Greenland (1987), however, this method cannot be used in most meta-analyses since too few studies are available.

6.5 COMPARISON AND ASSESSMENT OF THE FOUR TYPES OF REVIEWS

The different review methods are outlined in Table 6.1 and will be discussed here in detail.

6.5.1 Design, Conduct and Literature Search

For each type of review, the hypothesis, question, and conduct should be summarized and defined in a strict protocol in which clear inclusion and exclusion criteria for the studies and the details of the literature search are described. This component of the review process is important for each study type, but is especially needed if quantitative results are required. It should also be decided whether and which data will be required from the investigators of the individual studies.

An important problem of meta-analysis is publication bias. This bias has received a lot of attention particularly in the area of clinical trials. Publication bias occurs when studies that have non-significant or negative results are published less frequently than positive studies. For randomized clinical trials, it has been shown that even with a computer-aided literature search only some of the relevant studies will be identified (Dickersin, Scherer, & Lefebre, 1994). For observational studies additional problems exist: Very often a large number

Table 6.1 Comparison of Methods for Different Literature Review Methods

Requirement for the Review Method	Review	MAL	MAP	PMA
Planning and literature search				
Protocol	+?	+	++	++
Inclusion / Exclusion criteria	+	+	++	++
Systematic literature search (incl. Abstracts, Proceedings)	+?	+	++	*
Obtaining additional information from single studies that are not published	−	+?	+	*
Evaluation of sources of errors and bias				
Investigation of sources of bias	+?	+?	++	++
Evaluation of validity of individual studies	−	+?	++	*
Control of data collection	−	−	+?	++
Adjustment of inclusion criteria for individuals	−	−	+	++
Assessment and control of statistical analysis	−	−	++	++
Estimation of publication bias	−	?	+	*
Comparability of single studies				
Standardized study design	+?	+	+	++
Standardized assessment of risk factors	−	−	−	+
Standardized definition of exposure and confounder variables (categories)	−	−	+?	++
Standardized adjustment for confounder variables	−	−	+?	++
Statistical analysis				
Quantitative estimate for the effect	−	+?	++	++
Improvement of the precision of effects measured	−	?	+	++
Estimator for dose-response relationship	−	−	+?	++
Estimator for risk in subgroups	−	?	+	+
Increase of statistical power	−	+?	++	++
Evaluation of interactions and confounder effects	−	−	+	++
Evaluation of sources of heterogeneity	?	+?	++	++
Sensitivity analyses	−	+	++	++
Reproducibility of methods	−	−	+?	+?
General aspects				
Description of state of research	+	+	+	*
New research questions	++	++	+	+
Improvement of the quality of further studies	+	+	+	+
Time and costs for the study	very low	low	high	very high

Note. MAL = Meta-analysis of literature, MAP = Meta-analysis with patient data, PMA = Prospective meta-analysis, ++ = possible in principle and (almost) always done, + = possible in principle and often done, +? = often not possible or not done, ? = only possible or useful in exceptional cases, − = never possible, * = less relevant.

of variables will be collected in questionnaires as potential confounders. If one or several of these potential confounders yield significant or important results, they may be published in additional papers, papers that have often not been planned in advance. If these confounders, however, yield expected or negative results, no publication will be made. Some regional studies may not be published in international journals and are not found in a literature search for a meta-analysis.

Inclusion criteria, data collection methods, and statistical analyses cannot be changed if published data are used for a meta-analysis. In many situations it is even difficult to determine exactly what has been done from the published literature. The methods section in many papers is often short and critical evaluation is not always possible. Errors in the original work cannot be corrected or checked and may yield to bias in the results of the meta-analysis. For MAP, the inclusion criteria for the single studies can be modified (for different age groups, tumour sites, latency times, etc.). They can also be redefined and checked. It is also possible to evaluate or adopt the statistical analysis if patient data is available. Possible sources of a systematic bias can be eliminated if a detailed statistical analysis of the single studies can be done. The evaluation of possible bias attributable to lack of control for confounding is also only possible with individual data.

6.5.2 Validation of Comparability of the Single Studies

Since many study designs are possible, it is necessary to evaluate the comparability of the single studies before conducting a review. This evaluation can be conducted partly from published data if enough detailed information is available in the papers. If individual data are available, an analysis of the single studies in one common model is possible. A major reason for different results across studies is that different statistical methods/models have been used. Hence, heterogeneity can be significantly reduced in a pooled analysis by using the same model for all studies. A pooled analysis is only possible if similar data are available from all studies and are provided to the investigator of the pooled analysis. The investigation of study-specific heterogeneity can be done, to some extent in MAL, mainly with a sensitivity analysis.

6.5.3 Quantitative Risk Estimation

Reviews are not designed to give a quantitative estimate of the effect of risk factors or to describe a dose-response relation. They only allow a descriptive comparison of the results of available research. All other types of meta-analysis allow – if the data are sufficiently homogeneous – the calculation of a pooled risk estimate. A quantitative estimate is very often considered an important goal of meta-analysis. However, calculating an overall risk estimate is not always possible because different statistical models were used in the original studies that prohibit a sensible combination of study results to be made. It should also be noted that an improvement in the precision of the risk esti-

mate could often not be achieved by pooling since only the variation caused by random error (increasing the sample size) will be decreased by pooling. Increasing the sample size cannot eliminate any bias (systematic error). Indeed in some situations, the pooled estimate is less precise than estimates of the included single studies, as was shown by Gilbert (1989) in the radiation leukaemia studies.

A less precise estimate is likely if only data for a crude categorization (e.g., 2×2 tables) can be abstracted from the publication. Bias may also be increased if different methods to control confounding have been used in the individual studies. A more precise estimate is in most circumstances only possible in a re-analysis with individual patient data. Especially, to estimate a dose-response relation, individual patient data are required at least if different categories are used. Likewise, an investigation of interaction and confounding requires individual data. Prospective multicentre meta-analyses have the advantage that the data collection procedures, the measurement methods, the exposure assessment, and the definition of all variables can be agreed upon prior to data collection. Consequently, the data can be more easily combined at a later stage. It should also be noted that subgroup analysis, which is often a goal of a planned meta-analysis, could only be performed if the data are published with sufficient detail.

To investigate whether the results are consistent across studies, published data can be used for a review as well as for MAL. However, only limited search for sources of this heterogeneity is possible. For example, whether different definitions of exposure and confounder variables or different use of confounder in a multivariate statistical model influence the results can not be determined. A valid judgement of the consistency of results in complex questions requires a new and detailed statistical analysis based on original data.

Many authors have pointed out that investigating heterogeneity is the most important aspect of meta-analysis (e.g., Thompson, 1994). Statistical methods to investigate heterogeneity can be based on aggregated data. However, statistical tests have low power and may not be able to detect heterogeneity between studies. MAP allows different strategies to be used to eliminate differences and at least to give results in a unified way. Frequently, it is difficult to compare results from different observational studies since different data presentation methods are used across publication. Even in a single study different strategies for modeling can yield rather different results (Blettner & Sauerbrei, 1993). Therefore, for a meaningful meta-analysis it is necessary to eliminate this source of heterogeneity. Such a comparison is only possible if the same (or quite similar) variables are available from all studies.

6.6 SOME EXAMPLES

Ursin, Longenecker, Haile, and Greenland (1995) report results from a meta-analysis investigating the influence of the Body-Mass-Index (BMI) on development of premenopausal breast cancer. They include 23 studies of which 19 are

case-control studies and 4 are cohort-studies. Some of these studies were designed to investigate BMI as risk factor, others measured BMI as confounders in studies investigating other risk factors. It can only be speculated that the number of unpublished studies in which BMI was mainly considered as a confounder and did not show a strong influence on premenopausal breast cancer is non-negligible and that this issue may result in some bias. As is usual practice in epidemiological studies, relative risks were provided for several categories of BMI. To overcome this problem the authors estimated a regression coefficient for the relative risk as a function of the BMI, however, several critical assumptions are necessary for this type of approach. The authors found severe heterogeneity across all studies combined (the p-value of a corresponding test was smaller than 10^{-8}). An influence of the type of study (cohort-study or case-control study) was apparent. Therefore, no overall summary is presented for case-control and cohort studies combined. However, the authors present a summary estimate for all case-control studies, although the severe heterogeneity ($p < 10^{-8}$) was still present. One reason for the heterogeneity is the difference in adjustment for confounders. Adjustment for confounders other than age was used only in 10 out of the 23 studies. Several other issues may have caused the severe heterogeneity between studies and the summary assessment of an inverse association of high BMI with risk of premenopausal breast cancer must be interpreted with caution.

White (1999) investigates the level of alcohol consumption at which all-cause mortality is least. Based on a MEDLINE search he included 20 studies in the meta-analysis; nine studies were excluded because information needed for the meta-analysis was not available. The heterogeneity of the study populations and the differences of confounding factors is obvious from the summary table in the paper. Age is the only factor used in all studies, the number of additional factors ranges from 0 to 12. The author has used a complex method to re-calculate unadjusted relative risk estimates, however, combining crude estimates may yield a severe bias if confounding plays a role. Additionally, various numbers of categories based on different cutpoints were used in the individual studies. To combine these different estimates, the author fit asymmetric quadratic functions to the results based on categorized alcohol levels in the original papers. From each function a nadir indicating the least all-cause mortality was estimated. Obviously, the nadir depends strongly on the original chosen categories and the calculation to estimate the nadir from the published data are questionable and could yield a major bias. Four studies that did not show a "typical" U-shape relationship were considered to be noninformative about the nadir and were excluded from the analysis. For different subpopulations estimates of the nadir with corresponding confidence intervals were presented. The estimate is much higher for UK men than for US men, however, because of many limitations of the MAL the results are questionable (Sauerbrei, Blettner, & Royston, 2001). The used approach gives estimates that may be wrong and may lead to possibly wrong recommendations regarding the alcohol consumption. Using the publications only, a careful investigation of the large heterogeneity between studies and countries could have been a

worthwhile exercise, but the calculation of a quantitative estimate is meaningless.

Based on published data Ben-David, Rosen, Franssen, Einarson, and Szyfer (1995) present a meta-analysis investigating the influence of dose intensity of first line chemotherapy with cis- or carboplatin alone or in combination with other chemotherapy drugs on median survival of stage III-IV ovarian cancer patients. Following some central rules for meta-analyses they identified 61 "separate units", some from randomized trials and some from observational studies, which can be seen as independent one-armed observational studies. Based on the intended dose for each study and the distribution of the prognostic factors given in the corresponding publication, they tried to identify the influence of dose intensity of platin (DI), total dose intensity of platin (TDI), and total dose intensity of all chemotherapy drugs (GDI). The amount of information about prognostic factors given in the published papers and their incorporation in adjusting the treatment effect varied substantially between the different studies. As the effect of prognostic factors on survival is much stronger than the benefit from the treatment, observational studies require their incorporation in a multivariate model for the estimation of a treatment effect. However, in the published papers this issue is handled differently, leading to estimates which cannot be combined in a sensible way. Median survival categorized as less than 20 versus more than 20 months was used as the only outcome for each study. This very simple measure is not informative as it does not use survival time per patient. This choice was certainly guided by simplicity because only published papers were used. In a comment on the paper, Sauerbrei, Blettner, and Schumacher (1996, p. 428) concluded that

> in contrast to the obvious effect of TDI on median survival presented in the paper we believe that the authors did not succeed in adding any important information to the question of dose intensity on the survival of ovarian cancer patients.

Furthermore, a large randomized trial convincingly reached a result in contrast to that of the meta-analysis presented by Ben-David et al. (1995). Hopefully, not too many clinicians read this oversimplified meta-analysis and came to wrong conclusions for the treatment of their patients.

In breast cancer more than 100 factors are discussed as being potentially prognostic, however, only the number of positive axillary lymph nodes is generally accepted to have an important influence on prognosis. For most factors, only unsystematically traditional reviews have been published. The confusion caused by the current way of summarizing the evidence will be demonstrated by citing some review papers on HER2/neu oncogene, also known as c-erbB-2, a factor of strong interest in the last years. For this factor several traditional reviews based on different included studies, using different methods and leading to different conclusions have been published.

Based on their review, Allred, Harvey, Berardo, and Clark (1998) conclude that HER2/neu is at best a weak prognostic factor in node negative patients, whereas it seems to have a more pronounced prognostic value in node positive patients. More detailed information of single studies is given by Révillion, Bonneterre, and Peyrat (1998). They state that in several studies the prognos-

tic value of HER2/neu was present only in univariate analysis, and not in multivariate analysis. Furthermore, different results are sometimes reported for the effect on disease-free survival and overall survival. The tables 7 to 9 by Révillion et al. (1998) give evidence that incorporation of classical factors varies severely between the different studies and that, despite of the strong relationship to estrogen and progesterone receptors, these factors are often not considered in the analysis. For different studies this may cause severe differences in the estimated effect of the prognostic value of HER2/neu and makes a sensible comparison of results almost impossible. Révillion et al. (1998) summarize: "In univariate analyses HER2/neu is strongly associated with poor prognosis. However, it does not retain a clinical prognostic significance in multivariate analyses since it is associated with several strong prognostic parameters" (p. 791). As this statement is only based on a listing of the classical prognostic parameters used in the individual studies and on the simple assessment of significance for each single study but without any discussion on the power in the "negative" studies or on the size of the effect if the factor had a significant influence we consider the scientific basis for these important statements as being very low. Their review clearly demonstrates the heterogeneity between the studies concerning treatment, follow-up time, and on the issue of subgroup analyses. From our point of view any summary assessment seems unjustifiable. In another review published in two papers Ross and Fletcher (1998, 1999) list 47 studies investigating the prognostic value. They do not care about mixing results from univariate and multivariate analyses of the single studies, whose results are simply given as impact on prognosis "yes" or "no". In the paper published first they conclude "The preponderance of evidence indicates that HER2/neu gene amplification and protein overexpression are associated with an adverse outcome in breast cancer" (Ross & Fletcher, 1998, p. 424). A reader gets certainly the impression of an important prognostic factor. In the latter paper they give no summary statement but concentrate more on different measurement techniques and list studies with significant and non-significant results, respectively. Obviously, a useful summary assessment of the prognostic value seems impossible with the traditional review. In their later paper they seem to have realized it, however, they did not explicitly state it. From our point of view such a statement would have added some value to their paper.

Difficulties in general associated with reviews of prognostic factor studies are discussed in Altman and Lyman (1998). Most problems are closely connected to issues of assessing the importance of a factor from the results of observational studies.

6.7 CONCLUSION

This chapter has described and critically assessed different review methods for observational studies. We strongly believe that all available data and information are needed for full assessment of weak factors and that systematic

reviews of available evidence will become increasingly important. A major impediment for meta-analysis of observational studies is the heterogeneity between studies in their design, data collection methods, and statistical modeling. Mainly because of the last aspect meta-analyses using published data are, therefore, limited and give rarely a valid quantitative estimate or dose-response function. However, a meta-analysis of published data may be more reproducible than a qualitative review. A MAL has the trivial but dangerous advantage of being less expensive and time consuming than a meta-analysis with individual data. Consequently, some authors will continue to publish results from those meta-analyses and public health regulators, and decision-makers may rely on these results, even if the scientific value is questionable. Therefore, it remains important to point to the weaknesses and flaws in meta-analyses of literature. In particular, errors and bias that can be produced when combining studies with different design, methods, and analytic models need to be addressed. Despite of the large costs in time and manpower researchers should be encouraged to aim for meta-analyses with patient data. Several successful projects have shown that it is possible to interest researches all over the world for the collaboration (Advanced ovarian cancer trialists group, 1991; Early Breast Cancer Trialists' Collaborative Group, 1992), mainly because the question was so important that the scientific community was strongly interested in scientific answers. We believe that a useful MAP does not always need to incorporate all studies conducted for the specific question of interest but that a well defined "group of studies" – for example, only new, good, and large studies – may be sufficient. Such an approach will substantially reduce the costs and largely increase the probability to receive the individual data from the studies of interest. A meta-analysis starting with a re-analysis of the individual studies would have a chance to result in valid estimates or dose-response functions. With our examples we tried to show that the more traditional ways have often failed to give a reliable assessment for a factor of interest, despite of the fact that an enormous amount of money was spent from the individual study groups all over the world and data are available on tens of thousands of patients.

Statistical methods for pooling data from different sources have to be refined and new approaches are needed. Some important work is currently in progress, for example, from members of the "Statistical Methods Working Group" of the Cochrane Collaboration. Methods for conducting and reporting of meta-analyses of published data need to consider the basic limitations. While significant progress has been made in the systematic approaches for meta-analyses of randomized clinical trials, limitations in observational studies may not be overcome by too simple statistical methods. However, an equally rigorous standard is needed as more public health decisions will be relying on the results of meta-analyses. Hence, the research community must ensure that the validity, reliability, and overall quality of these methods is improved.

REFERENCES

Advanced ovarian cancer trialists group. (1991). Chemotherapy in advanced ovarian cancer: An overview of randomized clinical trials. *British Medical Journal*, *303*, 884–893.

Allred, D. G., Harvey, J. M., Berardo, M., & Clark, G. M. (1998). Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Modern Pathology*, *11*, 155–168.

Altman, D. G., & Lyman, G. H. (1998). Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment*, *52*, 289–303.

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts – Treatment for myocardial infarction. *Journal of the American Medical Association*, *268*, 240–248.

Bailar III, J. C. (1997). The promise and problems of meta-analysis. *New English Journal of Medicine*, *337*, 559–561.

Ben-David, Y., Rosen, B., Franssen, E., Einarson, T., & Szyfer, I. (1995). Meta-analyses comparing cisplatin total dose intensity and survival. *Gynecolocic Oncology*, *59*, 93–101.

Blettner, M., & Sauerbrei, W. (1993). Influence of model-building strategies on the results of a case-control study. *Statistics in Medicine*, *12*, 1325–1338.

Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T., & Friedenreich, C. M. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology*, *28*, 1–9.

Boffetta, P., Saracci, R., Andersen, A., Bertazzi, P. A., Chang-Claude, J., Cherrie, J., Ferro, G., Frentzel–Beyme, R., Hansen, J., Plato, N., Teppo, L., Westerholm, P., Winter, P. D., & Zochetti, C. (1997). Cancer mortality among man-made vitreous fiber production workers. *Epidemiology*, *8*, 259–268.

Chalmers, T. C. (1991). Problems induced by meta-analyses. *Statistics in Medicine*, *10*, 971–980.

Chalmers, T. C., & Lau, J. (1993). Meta-analytic stimulus for changes in clinical trials. *Statistical Methods in Medical Research*, *2*, 161–172.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurence. *Journal of the American Medical Association*, *263*, 1385–1389.

Dickersin, K. (1997). How important is publication bias? A synthesis of available data. *Aids Education and Prevention*, *9 (Supplement A)*, 15–21.

Dickersin, K., Scherer, R., & Lefebre, C. (1994). Identifying relevant studies for systematic reviews. *British Medical Journal*, *309*, 1286–1291.

DuMouchel, W. (1990). Bayesian metaanalysis. In D. A. Berry (Ed.), *Statistical methodology in the pharmaceutical sciences* (pp. 509–529). New York: Dekker.

Early Breast Cancer Trialists' Collaborative Group. (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomized tri-

als involving 31,000 recurrences and 24,000 deaths among 75,000 women. *The Lancet*, *339*, 1–15.

Eysenck, H. J. (1994). Meta-analysis and its problems. *British Medical Journal*, *309*, 789–792.

Friedenreich, C. M. (1993). Methods for pooled analyses of epidemiologic studies. *Epidemiology*, *4*, 295–302.

Galbraith, R. F. (1994). Some applications of radial plots. *Journal of the American Statistical Association*, *89*, 1232–1242.

Gilbert, E. S. (1989). Analyses of combined mortality data on workers at the Hanford Site, Oak Ridge National Laboratory and Rocky Flats Nuclear Weapons Plant. *Radiation Research*, *120*, 19–35.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*, 351–379.

Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. *Epidemiological Review*, *9*, 1–30.

Greenland, S. (1994). Invited commentary: A critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, *140*, 290–296.

Lubin, J. H., Boice, J. D. J., Edling, C., Hornung, R. W., Howe, G., Kunz, E., Kusiak, R. A., Morrison, H. I., Radford, E. P., Samet, J. M., & et al. (1995). Radon-exposed underground miners and inverse dose-rate (protraction enhancement) effects. *Health Physics*, *69*, 494–500.

Mengersen, K. L., Tweedie, R. L., & Biggerstaff, B. J. (1995). The impact of method choice on meta-analysis. *Australian Journal of Statistics*, *37*, 19–44.

Normand, S.-L. T. (1999). Meta-analysis: Formulating, evaluating, combining and reporting. *Statistics in Medicine*, *18*, 321–359.

Paul, S. R., & Donner, A. (1989). A comparison of tests of homogeneity of odds ratios in K 2×2 tables. *Statistics in Medicine*, *8*, 1455–1468.

Petitti, D. B. (1994). *Meta-analysis, decision-analysis and cost-effectiveness analysis. Methods for quantitative synthesis in medicine.* Oxford: Oxford University Press.

Révillion, F., Bonneterre, J., & Peyrat, J. P. (1998). ERBB2 oncogene in human breast cancer and its clinical significance. *European Journal of Cancer*, *34*, 791–808.

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Ross, J. S., & Fletcher, J. A. (1998). The HER–2/*neu* oncogene in breast cancer: Prognostic factor, predictive factor and target for therapy. *Stem Cells*, *16*, 413–428.

Ross, J. S., & Fletcher, J. A. (1999). The HER–2/*neu* oncogene: Prognostic factor, predictive factor and target for therapy. *Cancer Biology*, *9*, 125–138.

Sauerbrei, W., Blettner, M., & Royston, P. (2001). On alcohol consumption and all-cause mortality, letter to White. *Journal of Clinical Epidemiology*, *54*, 537–538.

Sauerbrei, W., Blettner, M., & Schumacher, M. (1996). Re: Meta-analysis comparing cisplatin total dose intensity and survival: A critical reappraisal. *Gynecologic Oncology*, *62*, 427–428.

Shapiro, S. (1994). Meta-analysis/Shmeta-analysis. *American Journal of Epidemiology*, *140*, 771–778.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Smith, T. C., Spiegelhalter, D. J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, *14*, 2685–2699.

Spiegelhalter, D. J., Thomas, A., Best, N., & Gilks, W. (1996). BUGS: Bayesian inference Using Gibbs Sampling. Available from http://www.mrc-bsu.cam.ac.uk/bugs/.

Stewart, L. A., & Clarke, M. J. (1995). On behalf of the Cochrane working group on meta-analysis using individual patient data. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*, *14*, 2057–2079.

Stewart, L. A., & Parmar, M. K. B. (1993). Meta-analysis of the literature or of individual patient data: Is there a difference? *The Lancet*, *341*, 418–422.

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal*, *309*, 1351–1355.

Thompson, S. G., & Pocock, S. J. (1991). Can meta-analyses be trusted? *The Lancet*, *338*, 1127–1130.

Ursin, G., Longenecker, M. P., Haile, R. W., & Greenland, S. (1995). A meta-analysis of body mass index and risk of premenopausal breast cancer. *Epidemiology*, *6*, 137–141.

White, I. R. (1999). The level of alcohol consumption at which all-cause mortality is least. *Journal of Clinical Epidemiology*, *52*, 967–975.

Ziegler, S., & Victor, N. (1999). Gefahren der Standardmethoden für Meta-Analysen bei Vorliegen von Heterogenität [Some problems of the standard meta-analysis methods in the presence of heterogeneity]. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, *30*, 131–140.