Missing values in epidemiological studies

Werner Vach
Center for Data Analysis and Model Building
& Institute of Medical Biometry and Medical Informatics
University of Freiburg

Maria Blettner
Department of Epidemiology and Biometry
German Cancer Research Center, Heidelberg

SOURCES OF MISSING VALUES IN EPIDEMIOLOGICAL RESEARCH

In analytic epidemiologic studies, mainly case-control studies* and cohort studies* or designs derived of these two basic types (such as case-cohort studies or nested case-control studies), in general, data are collected by questionnaire, or interview (face to face, telephone, computer assisted) or are abstracted from existing records such as hospital records containing information on treatment or on diagnosis, personnel records (e.g. in occupational studies) or death certificates. In general (except in studies with a two-stage design, see below) complete information is sought on an individual base for all subjects included in the study.

In case-control studies this includes retrospective collection of data, often information is required about events or exposures very far back in the past. Adequate planning and organization of the study should insure that data are collected in an identical way for diseased persons (cases) and for healthy subjects (controls). Additionally to the main exposure of interest, data are collected on known or suspicious confounder variables in order to adjust appropriately for these variables in an multivariate analysis. In matched case-control studies, some data (e.g. sex and age) are needed to perform the correct matching. In cohort studies personal interviews are carried out infrequently, but data are abstracted from existing files or records. In occupational cohort studies one can use personnel records to abstract data on the occupational history of individuals as well as data on exposure, but also records from the office of the occupational hygienist or routinely collected data from the medical officer. The quality and completeness of such data may differ substantially between companies or even departments of the same company. Data quality may also differ for different job categories and could therefore depend on the exposure of interest. Disease information in cohort studies is sometimes abstracted from hospital records or from cancer registries files. In mortality studies, data (date and cause of death) are abstracted from official death certificates or from other sources. An important issue in planing and organizing cohort studies is to try to guarantee a non-selective retrieval of information for the personal history (occupational history, life-style, residential history). It is also important to avoid any selective follow-up, that means the date of diagnosis or date of death, the diagnosis and/or the causes of death has to be assessed in a comparable way for exposed and non-exposed subjects.

Unplanned missing values

However, despite well organized data collection and for reasons known to all researchers but not always under their control, data may contain errors, the data collection is sometimes incomplete, and missing values occur. Missing data can arise for two main different reasons: it can arise from total non-response or from item non-response. Total non-response results from refusal of subjects to participate in the study, from incapability of find the selected subjects (e.g. in population based case-control studies, controls may have been selected but are not accessible because they have just moved). Total non-response is a frequent source of selection bias*. In this paper we restrict ourselves to item non-response.

Item non-response may arise because a person refuses to answer to certain questions, e.g. if the question is too sensitive or is regarded as too private (e.g. alcohol consumption, sexual behavior, income, health related questions). What is regarded as sensitive may differ rather substantially between persons and it may vary with personal behavior and/or depend on the answers to these or other questions. Older people may be more willing to answer to certain question then younger people. Persons with a very high or very low answer may not be willing to report their income. Another reason for missing values is that subjects do not know the answer because they are unable to recall certain events in their past. It also happens that a given answer is inconsistent with other answers and can therefore not be used in the analysis (e.g. if a persons says at one part of the questionnaire that she never smoked but reports a daily consumption of 20 cigarettes). Missing values can also occur if the interviewer fails to ask all questions, mainly if the interview was interrupted before all questions were asked. It also can happen that parts of the questionnaire are not readable or destroyed during the process of data editing. If data is abstracted from records, theses records may be incomplete for some persons, not readable or just missing. Different rules in some departments of an industrial setting or a hospital may have caused that records have been destroyed for some employees or for some patients. In many situations, records may include gaps, insufficient or controversial information, resulting in missing values. Similar, measures based on chemical or physical procedures may fail to produce a value, e.g. because this requires a certain amount of blood or tissue not always available, or just due to a lab accident where the material, the experiments or the results are destroyed and yield missing values. All these sources mentioned so far have in common, that the missing values are unplanned, so that we know the reasons usually only up to some vague degree. This makes this type of missing values so unpleasant for an analysis.

Planned missing values

Epidemiologic studies require collecting of data for many variables for many subjects.

Some sample strategies have been developed, where less data collection is required. A two stage design may be performed so that in a first stage data on the disease and exposure status is collected for many subjects, but additional information on detailed exposure or on confounding variables in a second stage only for a subsample. The second stage may include a fixed (similar) numbers of exposed and unexposed subjects. In a two stage design, a large amount of data can be missing values, but the reasons for the missing values are known. The probability that a value is missing is known or can be calculated easily and can be used for the analysis. Simple and efficient procedures to estimate exposure effects for such designs have been proposed by White [60] already in 1982. The idea of planned missing values is often propagated within the context of measurement error* and validation studies. Here an easy to measure surrogate variable is collected for all subjects, and exact measurements are made only for a subsample.

MISSING VALUE MECHANISMS

Whenever we want to handle a data set with missing values appropriately, the probability law generating the missing values will be of importance. Formally, this law, usually called the missing value mechanism, is the conditional distribution of the missing indicators, given all variables considered. To facilitate the discussion, it is now time to introduce some notation. We will consider in this contribution only the situation with one exposure and one confounder variable, where the confounder variable may suffer from missing values. Hence we consider for each subject four variables. The disease status D, the exposure E, the confounder C and the response indicator R, such that we actually observe C if and only if $R = 1$. This situation is complex enough to explain most problems and the basic ideas of solutions. Some solutions are, however, more or less restricted to this situation and lack generalizations to constellations with several exposures and/or confounders, especially in the case of arbitrary missing patterns; we will point this out where it is necessary. Also, one can exchange the role of E and C.

Now the missing value mechanism is given by the conditional probabilities to observe C, i.e. by

$$q(d, e, c) := P(R = 1|D = d, E = e, C = c).$$

To understand the possible dependencies of the observability of C on D, E and C, we shall discuss some specific situations. In case-control studies, missingness depends often on the disease status as cases and controls may differ in their behavior and willingness to participate in the investigation and to respond to specific questions. For example, Schlehofer et al. (1992) report results of a case-control study on risk factors for brain tumor investigating among other factors also the blood type group. For controls, only interview data were available, but for cases additionally hospital records could be used. This results in missing rates of 9 % for cases, but of 46 % for controls. Contrary, in a prospective cohort study, one can usually exclude a dependence of the response probabilities on the disease status, if all covariate data is collected at start of the study. Retrospective cohort studies and most hybrids between cohort and case-control studies suffer often by a dependence of missing probabilities on the disease status.

Also, the exposure variable may have an influence on observability of the confounder. Investigating the risk of radiation therapy, a given therapy may be associated with hospital records containing a detailed anamnesis including information on potential confounders. Investigating exposure levels in nuclear plant workers, higher exposure levels may be associated with frequent medical examinations increasing the chance to assess information on confounders.

There exist a variety of constellations where the probability to observe a variable depends on the value of the variable itself. Collecting data by a questionnaire or interview, heavy drinkers or smokers may refuse to admit this, very poor or very rich people may refuse to give information on their income and long term unemployed subjects may refuse to give information on their working history. Often the value of a variable may influence the probability to know or to remember it, for example if we ask subjects for cases of a disease within their first and second degree relatives and if there is no such case, he or she will often answer "I don't know", because he or she does not know all the relatives, but if there is one case, it suffices to know this one to give an answer. Also "objective" sources like hospital records are no guarantee to exclude a dependence on the true value. Looking for information on a special therapy, it is easy to detect it if it has been given, but the opposite can be only assessed if the hospital records cover completely the possible time period, such that a definite negative answer is possible

Especially in epidemiology we may often have a rather complicate mixture of these constellations. For example in case-control studies, cases may refuse more often to admit an unhealthy lifestyle than controls, because they feel guilty. On the other side they may better remember exposures in their life time, because they have sought for reasons of their illness. Similar, the willingness to admit specific sexual behaviors may differ between sex and age groups. As another example the availability of information on confounder variables may depend both on the disease status and the exposure level: If we have good sources for exposed subjects and for cases, only unexposed controls may suffer from missing values. These possible interactions make handling of incomplete data especially difficult.

So far, we have described possible constellations. Some of them are more dangerous than others, which, however, depends on the type of analysis. If one wants to make efficient use of subjects with incomplete confounder information, the missing at random (MAR) assumption is of central importance. It reads in our context

$$q(d, e, c) = q(d, e),$$

and it forbids that the true value of C has an influence on its observability. This assumption allows to estimate the conditional distribution of C, given D, E and $R = 0$ from those subjects with $R = 1$, which is the key to make efficient use of all data. Note that the MAR assumption allows a dependence on D and E. In two stage designs we can exclude a dependence on C, because the missing values are planned in advance, but sampling fractions typically depend on D and E. In the literature on missing values you can also find the missing completely at random (MCAR) assumption $q(d, e, c) = q$, but this is realistic in epidemiology only very seldom.

If one wants to ignore the subjects with incomplete covariate data, it is essential to assume that the selection of subjects introduces no selection bias, which leads to different requirements; this is further discussed later. We should finally mention that in a case-control study the definition of $q(d, c, e)$ refers to the selected subjects but it coincides with the values in the total population, provided that selection probabilities really depend only on the case-control status and not on the availability of information which is a requirement for any well-conducted case-control study.

FITTING LOGISTIC REGRESSION MODELS WITH INCOMPLETE COVARIATE DATA

For epidemiological investigations logistic regression* is an important tool to analyze the joint effect of one or several exposure variables on the disease risk adjusted for one or several confounding variables. In the case of one exposure and one confounder variable it is based on the assumption that the conditional probability to be diseased given the exposure value e and the confounder value c can be described by

$$P(D = 1|E = e, C = c) = \Lambda(\beta_0 + \beta_E e + \beta_C c) =: p_\beta(e, c)$$

with $\Lambda(t) = \frac{1}{1+exp(-t)}$. This way of writing suggests that E and C are binary or continuous variables, extensions to categorical variables are straightforward and most statements of this paper are valid for any type of covariates.

In the case of complete data we can estimate the parameters β_0, β_E and β_C by the maximum likelihood principle. In the case of incomplete data, there exist a lot of proposals of different quality. To understand the behavior of most simple methods to handle incomplete covariate data it is worth to look at the conditional probabilities of the disease status given the actual information we observed. Considering subjects with complete data we have

$$P(D=1|E=e,C=c,R=1) = \Lambda(\beta_0 + log\frac{q(1,e,c)}{q(0,e,c)} + \beta_E e + \beta_C c), \qquad (1)$$

which can be easily shown in analogue to the justification of logistic regression models for case control data as given by Breslow and Day [5] (p. 203), if we note that $q(d,e,c)$ are nothing else but the probability to select these subjects. (1) implies, that fitting a logistic regression model to these subjects alone will give valid estimates for β_E and β_C, if $q(d,e,c)$ can be decomposed into $q(d) \cdot q(e,c)$. Considering subjects with a missing value we have

$$P(D=1|E=e,R=0) = \int \Lambda(\beta_0 + log\frac{1-q(1,e,c)}{1-q(0,e,c)} + \beta_E e + \beta_C c) dF^{C|E=e,R=0}(c) \quad (2)$$

Most simple methods to handle incomplete covariate data try to approximate (1) and (2) by simple logistic models and the resulting misspecification can cause serious bias. Contrary, methods relying on the likelihood or on appropriately chosen estimation equations have the potential to produce consistent estimates. Hence we have now to consider the likelihood in the incomplete data case. Considering the joint distribution of the observed variables subjects without a missing value contribute with

$$q(d,e,c) \cdot p_\beta(e,c)^d \cdot (1-p_\beta(e,c))^{1-d} \cdot P(C=c|E=e) \cdot P(E=e)$$

and subjects with a missing value contribute with

$$\int (1-q(d,e,c)) \cdot p_\beta(e,c)^d \cdot (1-p_\beta(e,c))^{1-d} \cdot P(C=c|E=e) \cdot P(E=e) dc\ .$$

If the MAR assumption $q(d,e,c) = q(d,e)$ holds, not only $P(E=e)$ but also the terms involving q can be removed from the likelihood. However, the likelihood depends still on

$P(C = c|E = e)$, hence the classical maximum likelihood principle requires to specify the distribution of the covariates at least in part, which is a fundamental difference to the complete data case. Trying to avoid these difficulties leads to semiparametric approaches. Of course, the likelihood presented above is based on a prospective sampling scheme. In the case of complete data it is well known that nevertheless such a likelihood is allowed to be used in the analysis of case-control studies (Prentice & Pyke [32]). This is also true in the case of incomplete data as shown by Carroll et al. [9].

In the following we try to give an overview of the major simple and sophisticated methods to handle incomplete covariate data.

Complete Case Analysis

In a complete case analysis all subjects with a missing value are omitted from the analysis. The validity of this approach is based on the implicit assumption, that the regression model within the subjects with complete data is identical to the model for all subjects, i.e. that

$$P(D = 1|E = e, C = c, R = 1) = P(D = 1|E = e, C = c)$$

holds. With (1), this is true, if $q(d, e, c) = q(e, c)$, i.e. if missing probabilities do not depend on the disease status. This is also intuitively clear; if missing probabilities depend only on the covariate values, restriction to subjects without missing values changes only the population, but not the regression model, whereas missing probabilities depending additionally on the outcome introduce some type of selection bias*. A sole difference between the missing probabilities of cases and controls affects only the estimation of the intercept, but does not affect the estimation of β_E and β_C; in general consistent estimation of the latter is guaranteed if $q(d, e, c) = q(d) \cdot q(e, c)$, which follows directly from (1) (Glynn & Laird [19]).

So a complete case analysis has the favorable property to result in consistent estimates of the regression parameters, even if the MAR assumption is violated. Contrary it has the unfavorable property that consistency of parameter estimates depends on the assumption that missing probabilities do not depend jointly on the disease status and the covariate values. The latter is however often typically for case-control studies (cf. last section). The bias of the odds ratio based on a complete case analysis can be easily computed (Vach & Blettner [55]), and it can be shown that realistic differences in the missing probabilities can lead to substantial bias. For example if exposed cases are better documented than unexposed cases and controls such that the missing probability for the exposed cases is

10% and 40% for the other groups, then the odds ratio for exposure is overestimated by a factor of 1.5.

Additional Category or Missing Indicator Method

Since in epidemiology it is widespread to work with categorical variables, it is also widespread to work with the value "missing" as an additional category. This implies, that we analyze the data under the implicit assumption that

$$\begin{aligned} P(D=1|E=e, C=c, R=1) &= \Lambda(\beta_0+\beta_E e+\beta_C c) \quad \text{and} \\ P(D=1|E=e, R=0) &= \Lambda(\beta_0+\beta_E e+\beta^*)\ . \end{aligned}$$

Equivalently we can impute for the missing values of C the value 0 and add the missing indicator $M = 1 - R$ to the regression model; i.e. this "Missing Indicator Method" — applicable also for continuous covariates — results in the same specification and hence the same estimates. This approach is rather inappropriate, as one cannot expect to achieve good estimates for the adjusted risk β_E if the adjustment for the unobserved values of the confounding variable is tried to be managed by introducing the additional parameter β^*. To see this, let us assume, that $q(d,e,c) \equiv q$, i.e. MCAR, such that the subjects with and without missing values form two random subsamples. Then in the first line above β_E corresponds to the adjusted log-OR of the exposure, whereas in the second line β_E corresponds to the unadjusted log-OR, because $\beta_0 + \beta^*$ can be regarded as one intercept. Consequently, the estimate $exp(\hat{\beta}_E)$ arrived tends to estimate a quantity somewhere between the adjusted and unadjusted odds ratio. Hence the aim to achieve more realistic odds ratios describing the effect of exposure by adjusting for confounding variables cannot be achieved if missing values in the confounding variables are regarded as an additional category. Moreover, if the missing probabilities are allowed to depend on the disease status and/or exposure status, then $exp(\hat{\beta}_E)$ can tend to values outside the range between the adjusted and unadjusted odds ratio. The bias is often accompanied by underestimation of the variability; Greenland & Finkle [20] report the results of a simulation study with two Gaussian covariates, where the missing indicator method results in true coverage probabilities of 55% for nominal 95% confidence intervals.

So far we have considered the effect of coding missing values as an additional category on the estimation of β_E. In the epidemiological literature the estimate of β^* is often reported, too, and compared to the value of $\hat{\beta}_C$. Often there is an implicit assumption that $\hat{\beta}^*$ has to be between 0 and $\hat{\beta}_C$, or, in the case of several categories, within the range of the effect estimates (including 0 for the baseline category). If missing probabilities depend only on the exposure, and the degree of correlation between confounder and exposure is small, this is approximatively true, which can be shown using the approximation discussed in the next section. However, if missing probabilities depend on the disease status, the relative disease frequency within subjects with complete data differs from the relative disease frequency within subjects with incomplete data, and β^* mainly reflects this difference.

Although regarding missing values as an additional category cannot be recommended in general, it can be appropriate in special settings, where missing values characterize a meaningful subset of all individuals. For example Commenges et al. [11] report a study comparing different procedures to diagnose dementia in a screening setting. They found missing values in those variables corresponding to the results of two tests to be highly predictive, because here the missing values reflect a subject's failure to comprehend the test.

Single-imputation methods

This class of methods is characterized by imputing for each missing value a single value and to analyze the completed data set. If the confounder C is continuous, the most simple choice is to replace each missing value by the overall mean $\bar{C}$ of the observed values of the confounding variable. Instead of using an estimate for the overall expectation of C, one may use estimates of the conditional expectations: If E is categorical, we can impute the mean of the observed values of C within each category of E; if E is continuous, we can compute a regression of the observed values of C on E. If C is binary, relative frequencies replace the means, and Schemper & Smith [46] proposed the term probability imputation. The imputation of estimates for the conditional expectations yields an approximatively valid inference, if missing probabilities do not depend on the disease state and the true, unobserved value, i.e. if $q(d, e, c) = q(e)$. In this situation, we have

$$\begin{array}{lrcl} \text{by (1)} & P(D=1|E=e, C=c, R=1) & = & p_\beta(e,c) \quad \text{and} \\ \text{by (2)} & P(D=1|E=e, R=0) & = & \int \Lambda(\beta_0 + \beta_E e + \beta_C c) dF^{C|E=e}(c). \end{array}$$

If we regard Λ as an approximatively linear function, we have

$$P(D=1|E=e, R=0) \approx \Lambda(\beta_0 + \beta_E e + \beta_C \cdot E[C|E=e]).$$

Hence imputing estimates for the conditional expectation results in an approximatively correct specification of the conditional disease probabilities, and hence the resulting bias of the parameter estimates is often small. In general one has to expect additionally, that variance estimates tend to be too small, because the imputed values are treated as true ones and no adjustment is made for the additional variability introduced by imputing estimates. Results of simulation studies (Schemper & Smith [46], Vach & Schumacher [58], Vach [53], Schemper & Heinze [45]) suggest, that both bias and underestimation of the variance become only a problem for extreme parameter constellations with high missing rates and very influential confounding variables.

The justification so far depends on the assumption that missing probabilities do not depend on the disease status. This is not necessary, because imputation of conditional expectations can be regarded always as an approximation to simple semiparametric approaches (Vach & Schumacher [58]). However, some care is necessary: If missing probabilities depend on the disease status, then naive estimates for conditional expectations are wrong; it is necessary to estimate the conditional expectations separately within diseased and undiseased subjects and then to form a weighted average (Vach & Schumacher [58]). Moreover, for extreme parameter constellations the bias can be still substantial (Vach [53]).

Generalizations to several covariates with arbitrary missing patterns are straight forward, as far as there are enough subjects with complete information. But there may be many auxiliary regression models to be fitted to compute all predictions to be imputed. In general, misspecification of these auxiliary regression models can be a source of additional bias of the parameter estimates, but little is known on the relevance of this problem.

Modifying the complete case estimates

Under the MAR assumption the response probabilities $q(d, e)$ can be easily estimated by the observed data, for example by fitting a logistic regression model with outcome variable R and covariates D and E. The bias of the complete case estimates can be expressed as a function of q, and hence we can correct the bias (Vach & Blettner [55], Vach [53]). Alternatively, one may fit a logistic regression model with estimated offsets according to (1) to the subjects with complete covariate data (Breslow & Cain [4]). If E is categorical and a saturated model is used in estimating q, both approaches coincide and are identical to maximum likelihood estimates (Vach & Illi [57]). As also simple expressions for the asymptotic variances can be provided (Cain & Breslow [7]), this is a simple method to achieve consistent and efficient estimates in this special setting if the MAR assumption can be maintained. Unfortunately there exists no simple generalization to the situation of arbitrary missing patterns.

Estimation of the score function: Weighting, Filling and the mean score method

In the complete data case maximization of the likelihood is equivalent to finding a root of the score function

$$S_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} S_\beta(D_i, E_i, C_i) \quad \text{with} \quad S_\beta(d, e, s) = \frac{d}{d\beta} p_\beta(e, c)^d (1 - p_\beta(e, c))^{1-d}.$$

In the incomplete data case the contribution to the score function is unknown for subjects with a missing value. Nevertheless, one can try to estimate $S_n(\beta)$. A first approach is to regard the subjects with complete covariate information as a subsample with selection probabilities $q(d, e, c)$ and to try to estimate the "population average" $ES_\beta(D, E, C)$. The classical Horvitz-Thompson estimator* satisfies this task by weighting each contribution of the subsample with $q(d, e, c)^{-1}$. However, $q(d, e, c)$ is unknown, and only under the MAR assumption we can arrive at estimates $\hat{q}(d, e)$ and at a weighted score function

$$\tilde{S}_n(\beta) = \frac{1}{n}\sum_{\substack{i=1 \\ R_i=1}}^{n} S_\beta(D_i, E_i, C_i)/\hat{q}(D_i, E_i)$$

and solving $\tilde{S}_n(\beta) = 0$ results in consistent estimates of β. Solving $\tilde{S}_n(\beta) = 0$ can be done by any software package for logistic regression, if it allows arbitrary weights. However, variance estimates obtained this way are invalid, and can be much too small (Vach [53], Section 5.11). If a parametric model $q_\alpha(d, e)$ is used in estimating the response probabilities, explicit estimates of the variance can be provided (Pugh et al. [33], Vach [53], p. 17), but they cannot be computed with standard software. If E and C are both categorical, the approach is equivalent to distributing subjects with a missing value to the cells of the contingency table of subjects without a missing value proportional to an estimate of the conditional probability for the true value. This intuitive method was called "Filling" by Vach & Blettner [55]. The idea to weight contributions to the score function reciprocally to the response probabilities is also used by Flanders & Greenland [15] and Zhao & Lipsitz [61]. However, they consider the analysis of designs, where the response probabilities are known.

An alternative idea to estimate S_β is to replace each unknown contribution $S_\beta(D_i, E_i, C_i)$ for subjects with unknown C_i by an estimate for $E[S_\beta(D_i, E_i, C_i)|D_i, E_i]$, i.e. an estimate for the conditional expectation of the score function given the observed variables. Reilly & Pepe [34] investigate this approach in detail for the special case where E is categorical. Then estimates of the conditional expectations are simple averages within the subjects without missing values, and the approach is equivalent to weighting. However, whereas the weighting approach is difficult to be generalized to the case of several covariates with arbitrary missing patterns, this is in principle possible for the individual estimation of the conditional expectations by using methods of nonparametric regression.

Finally, estimates based on the weighting or the mean score approach are consistent under the MAR assumption, but not always efficient. Especially if missing rates are larger, there can be a substantial loss in comparison to efficient approaches (Zhao & Lipsitz [61], Robins et al. [38], Vach [53], Section 5.2).

Maximum Likelihood Estimation

Application of the maximum likelihood (ML) principle requires a parametric specification $f_\alpha(c|e)$ for the conditional distributions $P(C = c|E = e)$ (cf. above). Then under the MAR assumption the contributions to the likelihood are given by

$$\begin{array}{rll} p_\beta(e,c)^d(1-p_\beta(e,c))^{1-d} f_\alpha(c|e) & \text{if} & R=1 \\ \int p_\beta(e,c)(1-p_\beta(e,c))^{1-d} f_\alpha(c|e)dc & \text{if} & R=0\ . \end{array}$$

The integral in the likelihood makes maximization a little bit cumbersome. The EM-algorithm* (Dempster, Laird & Rubin [12]) is a standard tool to maximize the likelihood in incomplete data problems. However, if C is continuous, also the EM-algorithm may require numerical integration. If C is categorical, integration reduces to summation, and both the EM-algorithm (Ibrahim [24]) or a direct Newton-Raphson method* are feasible. The latter has the advantage to compute automatically the quantities necessary to estimate the variance of the parameter estimates, whereas use of the EM-algorithm requires additional efforts (Louis [30], Tanner [52]). The ML principle is applicable in the same manner also in the general setting with several covariates and arbitrary missing patterns, so far we are able to specify a parametric family for the conditional distribution of the covariates affected by missing values given the covariates unaffected.

The ML estimates are consistent and efficient as long as the MAR assumption is valid and the true distribution of the covariates is within the specified family. This specification is one crucial point of the ML approach, because this requirement is not necessary in the complete data case and our knowledge about the distributions of and dependencies between the covariates is usually limited. A misspecification of the distribution of the covariates, however, can imply a bias of the regression parameter estimates, so we have the situation that large efforts are necessary with respect to nuisance parameters. If all covariates are categorical, log-linear models may serve as a simple framework to describe the joint distribution (Vach & Blettner [56]), but if continuous covariates are involved, parametric classes flexible enough seem to be out of reach in general.

If all covariates are categorical, one can also fit a log-linear model to the joint distribution of all variables (Fuchs [16], Williamson & Haber [59]) and can use relationships between log-linear and logistic models.

Semiparametric Maximum Likelihood Estimation

We have seen in the last section that maximum likelihood estimation requires to specify a parametric family for the conditional distribution of C given E. It is a straightforward idea to avoid this unpleasant task by replacing $f(c|e)$ by a nonparametric estimate. Pepe & Fleming [31] consider the case of a categorical exposure, such that the empirical distribution within each exposure stratum can be used, Carroll & Wand [8] consider a continuous exposure and use kernel estimates. Both approaches rely on the assumption that missing probabilities do not depend on the disease status, but they can be generalized to this setting (Vach & Schumacher [58]). Computations of the resulting estimates of β require special software, and estimation of the variance, too. The resulting estimates are not fully efficient in comparison to the estimates of the next section. It is also difficult to generalize these approaches to settings with several covariates with arbitrary missing patterns, because this requires non-parametric estimation of high-dimensional multivariate conditional distributions.

Semiparametric Efficient Estimation

The last two sections have shown, that the handling of incomplete covariate data is basically a semiparametric problem: We are interested in the parameters of the regression model describing the conditional distribution of disease status given all covariates reflecting exposure and confounding variables, but the distribution of the covariates, in spite of being essential for the likelihood, should be left unspecified. In recent years there has been substantial progress in the general field of efficient semiparametric estimation* (e.g. Bickel et al. [3]), and Robins et al. [38] succeeded in making this progress fruitful for the problem of fitting generalized linear models to incomplete covariate data. They showed that roughly any consistent estimator for β is asymptotically equivalent to one defined as the solution of an estimating equation $\sum_{i=1}^{n} S_\beta(D_i, E_i, C_i) = 0$, where

$$S_\beta(D, E, C) = R\frac{h(E, C)(D - p_\beta(E, C))}{q(D, E)} - \frac{\varphi(D, E)(R - q(D, E))}{q(D, E)}$$

They were also able to characterize functions h_{opt} and φ_{opt} which lead to a semiparametric efficient estimate, i.e. the asymptotic variance of this estimate is exactly the supremum of the asymptotic variances of all maximum likelihood estimators based on parametric families $f_\alpha(c|e)$ covering the true $f(c|e)$. Of course, this is the best we can expect without imposing parametric assumptions. Unfortunately h_{opt} and φ_{opt} depend on the true values of β and the true distribution of C given E and are moreover not available in closed form.

However, an adaptive procedure is possible which starts with a parametric assumption on the distribution of the covariates, then estimates all parameters, uses an iterative procedure to compute $\hat{h}_{opt}$ and $\hat{\varphi}_{opt}$ based on the assumption that the estimates correspond to the true parameters, and finally solve the estimation equations with h and φ replaced by $\hat{h}_{opt}$ and $\hat{\varphi}_{opt}$, and q replaced by an appropriate estimate. Contrary to ML estimation a misspecification of the covariate distribution does not result in inconsistent estimates, and in spite of the adaptive steps the estimates are efficient, if the specification of the covariate distribution was correct. Details of this adaptive procedure can be found in Robins et al. [38] and Rotnitzky & Robins [40]. The approach can be also generalized to several covariates with arbitrary missing patterns; however, here the computation of $\hat{h}_{opt}$ and $\hat{\phi}_{opt}$ is more difficult.

Multiple Imputation

Multiple imputation is a general technique for statistical inference with incomplete data. The basic idea is to create several data sets with different values imputed for the missing values, and to analyze each data set by standard software, here some software for logistic regression. If the imputations are generated in an appropriate manner, the average of the parameter estimates provides a consistent estimate. Furthermore, the average of the variance estimates and the empirical variance of the multiple parameter estimates can be combined to a variance estimate, and confidence intervals and p-values can be computed, too. Rubin & Schenker [44] present an overview of the basic techniques.

For generating imputations a straightforward idea is to draw from estimates of the conditional distribution of the unobserved values. However, this is an improper method in the sense, that variance estimates can be too small, because they do not take into account the variance due to estimating the conditional distributions; proper methods can be defined by additionally estimating the conditional distributions in each imputation step based on a random sample with replacement of the subjects without missing values (Rubin [42,43], Efron [14]). Of course, any attempt to estimate the conditional distribution of the missing values from the observed values depends on the MAR assumption.

With respect to our setting Reilly & Pepe [34,35] have considered the special case where E is categorical. Values to be imputed for missing values in C are drawn from the empirical distributions of C within the strata defined by D and E. This hot-check imputation method is of course improper, however, Reilly & Pepe [35] provide a valid variance estimator. Moreover they showed that hot-check multiple imputation with infinite imputations is asymptotically equivalent to the mean-score method. This especially implies, that we have the same deficiencies with respect to efficiency. Greenland & Finkle [20] report results of a simulation study with E and C both continuous and affected by missing values. Imputations are drawn from estimated conditional distributions resulting from fitting bivariate Gaussian distributions within the diseased and undiseased subjects. Although this is an improper method they observed that confidence intervals keep their nominal level. They also observe a loss of efficiency in comparison to maximum likelihood estimation.

Multiple imputation can be also applied in general settings with arbitrary missing patterns. The crucial point is the choice of the procedure to estimate the necessary conditional distribution. If we rely on parametric assumptions on the distribution of the covariates, we have the same unpleasant situation as with ML estimation. However, one can alternatively draw imputations from a set of nearest neighbors, i.e. subjects with complete information and similar values with respect to the observed variables. The choice of an appropriate distance measure requires of course some knowledge on the distribution of the covariates, but not necessarily an explicit model. Heitjan & Little [22] give here an illuminating example.

Methods Based on the Retrospective Likelihood

The methods considered so far rely on a prospective sampling scheme implying independence of the disease status among different subjects. In case-control studies this assumption is violated. However, also in incomplete data problems the use of the prospective likelihood can be justified (Carroll et al. [9]): The resulting estimates are consistent, the estimated standard errors are never too small and correct, if we make no assumptions on the distribution of the covariates. Nevertheless, methods based on the retrospective likelihood are of interest, especially for the analysis of two-stage designs. In such a design, the number of subjects with complete data is fixed in advance, and hence missing indicators are not independent, so we have further violations of the prospective sampling scheme.

Maximum likelihood estimation with respect to the retrospective likelihood is considered by Scott & Wild [51] and Breslow & Holubkov [6]. Pseudo maximum likelihood estimates, where some parameters are preestimated in a naive manner, are considered by Breslow & Cain [4] and Schill et al. [47]. A weighting approach is due to Flanders & Greenland [15]. Comparisons with respect to the asymptotic relative efficiency and simulation studies (Zhao & Lipsitz [61], Breslow & Holubkov [6], Schill & Drescher [48]) reveal often large deficiencies of the weighting approach and some deficiencies of the two pseudo maximum likelihood approaches, which give usually similar results.

Handling of a Questionable MAR Assumption

All sophisticated, and especially all efficient approaches to handle incomplete covariate data rely on the MAR assumption. In many applications this assumption is questionable, but one may still want to use methods relying on the MAR assumption. Then it is necessary to think about or investigate the possible impact of a violation. One may argue that if there is a pure violation in the sense, that missingness depends only on the true value of the covariate, the impact must be small, because the association between the covariates and the outcome is not changed. Schemper & Smith [45] provide an informal argument for this conjecture. Investigations for the special case of both C and E being categorical (Vach & Illi [57]) corroborate the conjecture and further demonstrate that the impact on the exposure effect estimate can be substantial large, if there are small differences in the degree of violation between diseased and undiseased or between exposed and unexposed subjects, which is also intuitively clear, because such differences change the observed association.

If one does not want to rely on such general, theoretical considerations, one may try to investigate the impact of an invalid MAR assumption for a particular data set. This can be easily done within the multiple imputation framework, for example by drawing more frequently larger values for a variable or more frequently a specific category (cf. Rubin & Schenker [44]). Vach & Blettner [56] present a framework to specify violations within the framework of ML estimation and perform a sensitivity analysis for two case-control studies. Baker [2] makes an additional step and does not specify, but tries to estimate the parameters of the non-MAR mechanism. Rotnitzky & Robins [40] consider this step within the framework of semiparametric efficient estimation. However, a (saturated) logistic model and a (saturated) non-MAR model are in general not jointly identifiable, hence any attempt to estimate non-MAR mechanisms relies on restrictions of the two models allowing identifiability. This alone, however, is not enough, as identifiability does not imply reasonable properties of resulting estimates in this setting: Rotnitzky & Robins [40] show in the semiparametric setting, that in spite of identifiability there need not exist a $\sqrt{n}$-consistent estimator. Hence, the usefulness of these approaches has to be investigated further, before recommendations can be made.

Robins & Gill [37] point out, that in settings with arbitrary missing patterns the MAR assumption as defined by Rubin [41] allows some constellations of no practical relevance. This can be used to change this assumption allowing some special non-MAR mechanisms to be estimated without problems of identifiability. Robins & Gill [37] and Robins [36] present two examples of this kind.

HANDLING OF INCOMPLETE DATA IN OTHER STATISTICAL METHODS RELEVANT FOR ANALYTIC EPIDEMIOLOGY

Poisson regression, Gaussian regression and generalized linear models

Nearly anything we have said in the last paragraph with respect to logistic regression is also valid for other regression models where parameters are estimated by maximum likelihood. Especially the difficulties with maximum likelihood estimation in the incomplete data case are the same, and the semiparametric approaches work in the general setting of generalized linear models*. With respect to the simple methods, there are two differences. First, there is no general analogon to the modifications of the complete case estimates. Second, the single imputation methods need more care. We can expect nearly unbiased estimates of the regression parameters after imputation of conditional means, as this implies a roughly correct specification of the conditional expectation of the outcome variable. Indeed, in the case of Gaussian regression one can prove consistency (Gill [18]). However, only in binary regression models correct specification of the conditional mean implies correct specification of the conditional variance. In general, the conditional variance of the outcome increases, if some covariate values are missing, hence after the imputation of conditional means a further analysis should be based on a heteroscedastic model. For this reason in Gaussian regression the use of weighted least squares estimates is advocated after imputation of conditional means. An overview for this and other techniques suitable for Gaussian regression models is given by Little [27]. Note that some of the proposals depend on the assumption of a multivariate normal distribution of all variables and hence are not very suitable for epidemiology. The impact of the variance heterogeneity for other types of regression models, especially Poisson regression, has not been investigated until now, so we can give only the recommendation to use single imputation methods here with care.

Cox regression with incomplete covariate data

For the analysis of (censored) survival times the use of the proportional hazard model* (Cox [10]) has become widespread also in epidemiology. Simple methods to handle incomplete covariate data are subject to the same criticism as for logistic regression, with the additional difficulty, that, especially in retrospective studies, censoring may be associated with missingness in covariates, such that in a complete case analysis the assumption of non-informative censoring can be violated. With respect to more sophisticated approaches, it is more difficult to generalize the partial likelihood approach here than for logistic regression, as the nuisance parameter involves the baseline hazard, although a semiparametric partial maximum likelihood approach is possible (Zhou & Pepe [62]). A weighting approach has been proposed by Pugh et al. [33], and Lin & Ying [26] consider an appropriately modified score function, but their approach requires MCAR. None of these approaches can be easily generalized to situations with general missing patterns and hence are only useful in particular situations. Robins et al. [38] also point out the difficulty to obtain a feasible solution from the theory of semiparametric efficient estimation. In face of this problem one may be willing to use alternative fully parametric regression models for survival data, such that, especially in the case of categorical covariates, the ML principle can be used. In this spirit, Schluchter & Jackson [50], Baker [1] and Vach [54] suggests to approximate the Cox model by a logistic model for grouped survival data, and Lipsitz & Ibrahim [29] considers Weibull models. The use of single imputation methods has been considered by Schemper & Smith [46].

Analysis of matched case-control studies

The handling of incomplete covariate data in matched case-control studies has been paid little attention. Haber & Chen [21] consider the case of a single exposure variable as the only covariate and compare the matched and unmatched odds ratio estimator. They conclude, that in the case of missing exposure information for some cases and controls, the advantages of the unmatched estimator increase in comparison to the complete data case. If we want additionally to adjust for confounding variables, conditional logistic regression* is a standard tool in analytic epidemiology for the analysis of matched case-control studies. Missing values in the covariates constitute here a problem even greater than in ordinary logistic regression, as a complete case analysis would imply in the case of one-to-one-matching, that a missing value in either a case or a control causes loss of the complete pair. Nevertheless, a systematic investigation of the problem is still missing, we know only a report on a small simulation study of limited value (Gibbons & Hosmer [17]).

Regression models for longitudinal or multivariate data

Regression models for longitudinal or clustered data, especially marginal models*, have been paid increasing interest in epidemiology, especially for the analysis of family aggregated data or in environmental studies. With respect to incomplete covariate data, there is little to add to what we have said in the last sections. However, in these settings we have also to handle missing values in the outcome variables, especially with drop outs in longitudinal data. There exists a fast growing literature on this topic, and we want to restrict us here to some basic comments, especially on the differences to the incomplete covariate problem.

First, the MAR assumption is again of central importance. In the case of drop outs it requires that the reason is only associated with observed variables. Hence the crucial question is, whether we are able to observe the crucial event before the drop out, or whether the drop out hides the event. Second, if the MAR assumption can be maintained, and if we consider regression models specifying the joint conditional distribution of the outcome variable and allowing to use the ML principle in the complete data case, then the ML principle can be used also in the presence of missing values in the outcome variables and reduces usually to an analysis of all units with measured outcome. Third, the popular marginal models (Liang & Zeger [25]) do not belong to this class, and the MAR assumption is here not sufficient to exclude a bias due to missing values, if only the available units are used; a solution has been provided by Robins et al. [39]. Fourth, if the MAR assumption is violated, we have often some rather precise ideas on the drop out mechanism, which allow to adjust for its effect by choosing an appropriate model (Diggle & Kenward [13], Little [28], Hogan & Laird [23]).

STRATEGIES TO COPE WITH INCOMPLETE DATA

The best advice with respect to missing values is to avoid them. Here we have great opportunities in planning appropriate data collection procedures and in the design of interviews and questionnaires, such that subjects have little reason to refuse an answer. An adequate planning can also help to avoid differential missingness or dependence of missingness on other important factors. Basically, the same data collection procedure should be used for cases and controls, and exposed and unexposed subjects should be paid the same care in completing the follow up. A second good advice is to keep the occurrence of missing values under the control of the investigator. Usually one knows in advance, which variables will suffer from missing values. Then a strategy to make the problem feasible is to collect data on a surrogate variable less affected by missing values and to collect the variable of interest only in a small subsample with additional efforts. This way the problem has been transformed to a measurement error* problem with a validation sample, but now this is an incomplete data problem where the MAR assumption holds, because the occurrence of missing values is planned in advance. Hence it is possible to use statistical methods very similar to the sophisticated methods discussed earlier. The only difference is that the surrogate variable is not considered in the regression model. If this solution is not possible, a third good advice is to collect additional data, such that the occurrence of missing values becomes reproducible. For example we can collect data on variables with a high predictive value on the occurrence of missing values like level of education, and by incorporating these variables in the analysis, the MAR assumption may be more reliable. A fourth rigorous strategy is to draw a sample from the nonresponders and to try to collect the missing data in a second stage. If this succeeds, a valid analysis becomes possible in principle.

If all these attempts are either impossible or unsuccessful, and we have no other chance than to analyze data as it is, one should try to discuss the possible impact of the missing values on the results of the analysis. For this the first step is to report the missing rates for all variables, stratified by the disease status and the exposure, and a summary of the major associations with other variables. The second step is a justification of the chosen methods; if a complete case analysis is applied in a case-control study, one has to give arguments to exclude a qualitative difference of the missing value mechanism between cases and controls. If one uses methods relying on the MAR assumption, the latter must be justified or a sensitivity analysis should be conducted.

CONCLUSIONS

Missing values are a common problem in the analysis of epidemiologic studies. Similar as for the problem of measurement error we can expect solutions only, if the problem is addressed already in the planning of a study. Then we can find either ways to avoid missing values, or to plan them in advance or to monitor their appearance, such that their probability law is under the control of the investigator or at least understandable to a degree, which allows valid inference. If this prerequisites are fulfilled, there exists a statistical methodology promising to make efficient use of all data, although today there are still some deficiencies with respect to practical experience and availability of software. However, we can here expect a parallel development producing better studies as well as better software. Contrary, the occurrence of unplanned missing values will always prevent an efficient analysis of an epidemiological study, and in the case of case-control studies it may even prevent to draw any valid conclusion. It is not within the power of statistics to solve this problem, and partial solutions can only be given if some knowledge on the mechanism generating the missing values can be assumed.

References

[1] Baker, S.G. (1994). Regression analysis of grouped survival data with incomplete covariates: Nonignorable missing-data and censoring mechanisms. *Biometrics* 50, 821-826.

[2] Baker, S.G. (1996). Reader reaction: The analysis of categorical case-control data subject to nonignorable nonresponse. *Biometrics* 52, 362-369.

[3] Bickel, P.J., Klaassen, C.A., Ritov, Y., and Wellner, J.A. (1993). *Efficient and adaptive estimation for semiparametric models*, Baltimore: John Hopkins University Press.

[4] Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika* 75, 11-20.

[5] Breslow, N.E. and Day, N.E. (1980). *Statistical methods in cancer research, vol. 1 - The analysis of case-control studies*, IARC Scientific Publications No. 32: Lyon.

[6] Breslow, N.E. and Holubkov, R. (1997). Weighted likelihood, pseudolikelihood and maximum likelihood methods for logistic regression two-stage data. *Statistics in Medicine* (to appear)

[7] Cain, K.C. and Breslow, N.E. (1988). Logistic regression analysis and efficient design for two-stage studies. *American Journal of Epidemiology* 128, 1198-1206.

[8] Carroll, R.J. and Wand, M.P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society B* 53, 573-585.

[9] Carroll, R.J., Wang, S., and Wang, C.Y. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association* 90, 157-169.

[10] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* B 34, 187-220.

[11] Commenges, D., Gagnon M., Letenneur, L., Dartigues, J.F., Barbarger-Gateau, P., and Salamon R. (1992). Improving screening for dementia in the elderly using mini-mental state examination subscores, Benton's visual retention test, and Isaacs' set test. *Epidemiology* 3, 185-188.

[12] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39, 1-38.

[13] Diggle, P. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* 43, 49-93.

[14] Efron, B. (1994). Missing data, imputation and the bootstrap (with discussion). *Journal*

of the American Statistical Association 89, 463-479.

[15] Flanders, W.D. and Greenland, S. (1991). Analytical methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* 10, 739-747.

[16] Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association* 77, 270-278.

[17] Gibbons, L.E. and Hosmer, D.W. (1991). Conditional logistic regression with missing data. *Communications in Statistics, B – Simulation and Computation* 20, 109-119.

[18] Gill, R.D. (1986). A note on some methods for regression analysis with incomplete observations. *Sankhya B* 48, 19-30.

[19] Glynn, R.J. and Laird, N.M. (1983). Regression estimates and missing data: Complete case analysis. Unpublished manuscript, Department of Biostatistics, Harvard University

[20] Greenland, S. and Finkle, W.D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analysis. *American Journal of Epidemiology* 142, 1255-1264.

[21] Haber, M. and Chen, C.C.H. (1991). Estimation of odds ratios from matched case-control studies with incomplete data. *Biometrical Journal* 33, 673-682.

[22] Heitjan, D.F. and Little, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics* 40, 13-29.

[23] Hogan, J.W. and Laird, N.M. (1997). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine* (to appear)

[24] Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* 85, 765-769.

[25] Liang, K.Y and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

[26] Lin, D.Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* 88, 1341-1349.

[27] Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association* 87, 1227-1237.

[28] Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90, 1112-1121.

[29] Lipsitz, S.R. and Ibrahim, J.G. (1996). Using the EM-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis* 2, 5-14.

[30] Louis, T.A. (1982). Finding the observed information when using the EM algorithm. *Journal of the Royal Statistical Society B* 44, 226-233.

[31] Pepe, M.S. and Fleming, T.R. (1991). A nonparametric method for dealing with missing covariate data. *Journal of the American Statistical Association* 86, 108-113.

[32] Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* 66, 403- 412.

[33] Pugh, M., Robins, J., Lipsitz, S., and Harrington, D. (1993). Inference in the Cox proportional hazards model with missing covariate data. Technical report 758Z, Division of Biostatistics, Dana-Farber Cancer Institute, Boston

[34] Reilly, M. and Pepe, M. (1995). A mean score method for missing and auxiliary covariate

data in regression models. *Biometrika* 82, 299-314.

[35] Reilly, M. and Pepe, M. (1997). The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine* (to appear)

[36] Robins, J.M. (1997). Non-response models for the analysis of non-ignorable missing data. *Statistics in Medicine* (to appear)

[37] Robins, J.M. and Gill, R. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* (to appear)

[38] Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846-866.

[39] Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 90, 106-121.

[40] Rotnitzky, A. and Robins, J.M (1997). Analysis of semiparametric regression models with non-ignorable non-response. *Statistics in Medicine* (to appear)

[41] Rubin, D.B. (1976). Inference and missing data. *Biometrika* 63, 581-592.

[42] Rubin, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics* 9, 130-134.

[43] Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*, Wiley: New York.

[44] Rubin, D.B. and Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine* 10, 585-598.

[45] Schemper, M. and Heinze, G. (1997). Probability imputation revisited for prognostic factor studies. *Statistics in Medicine* (to appear)

[46] Schemper, M. and Smith, T.L. (1990). Efficient evaluation of treatment effects in the presence of missing covariate values. *Statistics in Medicine* 9, 777-784.

[47] Schill, W., Jöckel, K.H., Drescher, K., and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika* 80, 339-352.

[48] Schill, W. and Drescher, K. (1997). Logistic analysis of studies with two-stage sampling: A comparison of four approaches. *Statistics in Medicine* (to appear)

[49] Schlehofer, B., Blettner, M., Becker, N., Martinsohn, C., and Wahrendorf, J. (1992). Medical risk factors and the development of brain tumor. *Cancer* 69, 2541-2547.

[50] Schluchter, M.D. and Jackson, K.L. (1989). Log-linear analysis of survival data with partially observed covariates. *Journal of the American Statistical Association* 79, 772-780.

[51] Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics* 47, 497-510.

[52] Tanner, M. (1994). *Tools for statistical inference. Methods for the exploration of posterior distributions and likelihood functions*, New York: Springer.

[53] Vach, W. (1994). *Logistic regression with missing values in the covariates*, Lecture Notes in Statistics 86: New York, Springer.

[54] Vach, W. (1997). Some issues in estimating the effect of prognostic factors from incomplete covariate data. *Statistics in Medicine* (to appear)

[55] Vach, W. and Blettner, M. (1991). Biased estimation of the odds ratio in case-control

studies due to the use of ad-hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology* 134, 895-907.

[56] Vach, W. and Blettner, M. (1995). Logistic regression with incompletely observed categorical covariates – Investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine* 14, 1315-1329.

[57] Vach, W. and Illi, S. (1997). Biased estimation of adjusted odds ratios from incomplete covariate data due to violation of the MAR assumption. *Biometrical Journal* (to appear)

[58] Vach, W. and Schumacher, M. (1993). Logistic regression with incompletely observed categorical covariates – A comparison of three approaches. *Biometrika* 80, 353-362.

[59] Williamson, G.D. and Haber, M. (1994). Models for three-dimensional contingency tables with completely and partially cross-classified data. *Biometrics* 50, 194-203.

[60] White, J.E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* 115, 119-128.

[61] Zhao, L.P. and Lipsitz, S. (1992). Designs and analysis of two-stage designs. *Statistics in Medicine* 11, 769-782.

[62] Zhou, H. and Pepe, M.S. (1995). Auxiliary covariate data in failure time regression. *Biometrika* 82, 139-149.