Stockholm School of Economics
Department of Economics
Master´s Thesis in Economics

Missing Data in Health Economic Evaluations

- handling the associated uncertainty

ABSTRACT
The purpose of this thesis is to explore the problems associated with missing data in studies included in health economic evaluations in order to promote validity and improve the base for decision making. The literature study focus on adjustment methods from a realized dataset with individual data, but also covers methods to prepare and prevent the problem. Missing data, especially nonresponse in survey, is common and threatens validity for two reasons. First, it can bias results, for example when occurring more frequently in the right tail of cost distributions that are skewed due to rare expensive events. Second, variance estimation can be distorted. A battery of methods should preferably be applied to make the reason for missing data observed within the gathered data, since otherwise crucial modelling assumptions are needed. Robustness of models is therefore a subject for sensitivity analysis. Most efficient methods are based on maximum likelihood and multiple imputation, but are primarily large sample tools unless applied with Bayesian estimation. Their potential in the examined dataset were low due to the limited information. This thesis will hopefully provide help in reducing uncertainty accompanied with missing data, which in the end would lead to improved decision making.

Author: **Nicklas Pettersson**
Tutors: **Mattias Ekman**
Niklas Zethraues

Opponent(s):
Examiner:
Presentation:

LIST OF ABBREVIATIONS		first appearance on page:
ACA	Available-Case-Analysis	26
CBA	Cost-Benefit Analysis	8
CCA	Complete-Case-Analysis	26
CEA	Cost-Effectiveness Analysis	6
CMA	Cost-Minimization Analysis	6
CUA	Cost-Utility Analysis	7
EM	Expectation-Maximization	31
HYE	Healthy-Year Equivalents	7
ICER	Incremental Cost-Effectiveness Ratio	6
LR	Logged Odds-Ratio	40
MAR	Missing At Random	21
MCAR	Missing Completely at Random	21
MCI	Mild Cognitive Impairment	35
MCMC	Markov Chain Monte Carlo	14
MI	Multiple Imputation	31
ML	Maximum Likelihood	30
MSE	Mean Squared Error	17
NMAR	Not Missing at Random	22
QALY	Quality Adjusted Life Year	7
RCT	Randomized Control Trial	10
SPAR	Statens Person- och AdressRegister	36
TTO	Time-Trade-Off	7
WTP	Willingness-To-Pay	8

1 Introduction

1.1 Background

Missing data is a common problem in quantitative studies in economic and social research, and is often a source of frustration. Instead of taking care of missing data in a proper way, the problem is often either swept under the carpet or dealt with in a rather primitive way, for example by excluding incomplete observations or imputing the missing data with average values. If missing data is not treated in a proper way, biases might be introduced, and accuracy reduced. This could cast doubt on and flaw any otherwise complete analysis, and thus worsen the foundation for rational decision making.

Although the applications in this study are to health economic evaluations, the general principles reviewed here are of much wider interest and can be applied to all sort of surveys[1] in economics and business, such as income and consumption surveys and market surveys. Since health economic evaluations often are intertwined with clinical or medical studies, terms like patients and treatments are often used for ease of commonness. Still, they are generally applicable to non-clinical health related settings as well. The typical unit of study in economic evaluations is human patients, and much of the theory is gathered from survey studies, where missing data usually appear as nonresponse. The described principles usually equate apply to all kinds of missing data situations.

Johnston et al [1] and Briggs et al [2] call attention to the problem of missing data within cost-analysis with patient-level data, commonly used in economic evaluations. They provide several reasons why the treatment of missing data is important, most of them always relevant in research, but some are special to or accentuated in economic evaluations. Briggs points out (on page 378) that: *"data on resource use for all patients in a trial are unlikely to be complete. However, it is rare to find any discussion of how missing data were handled in economic evaluations alongside clinical trials."*.

Many studies are also carried out on unhealthy patients. There is a considerable risk that they will drop out, withdraw conscience or deliver less complete data due to their condition or potential side effects from treatment, and that the problem is exaggerated among the most severely affected. High-cost events are often rare, but also more common among the most severe patients, leading to skewed cost distributions [3]. This combination of skewed cost distributions and incomplete information could be devastating because of the increased risk of underestimating the costs and getting biased result.

Further, economic evaluations are often piggybacked to patient surveys or clinical trials, which means that they are added to planned or sometimes even ongoing studies and carried through together with them. Surveys are in general accompanied by nonresponse. There is also an obvious risk that less attention is paid to considerations of a piggybacked economic evaluation, and that of the clinical trial will govern design and data collection issues. This could result in higher rates of missing data on cost data. Also, clinical outcomes are usually less variable than cost outcomes. Therefore, when sample size is determined by requirements of showing significant clinical effects, the potential of finding significant economic effects are smaller. The popularity of having economic evaluations piggybacked on randomized control trials can thus make data a scarce resource, which calls for optimizing use of all available data.[4]

[1] Survey is a method of gathering information from a sample drawn from a population that is being studied.

When missing data is present it always has to be handled in some way; there is no non-treatment solution. If the problem is ignored it will usually be handled by the default of the statistical software, which usually is exclusion of cases with incomplete data. This implies a further loss of data, leading to a reduction in statistical power and potentially biased results. Briggs [2] is pessimistic about this and says (on page 378); *"we suspect that in most cases health economists when confronted with missing data will use very simple methods (complete case, available case or unconditional mean imputation) to overcome the problem."*.

Different methods are available to handle the problem of missing data. Most of them are based on an idea about the mechanism causing the missing data, followed by some procedure to adjust for the missing data. Methods could be quite simple, such as applying weights to cases, depending on the missing data. Others are far more sophisticated, involve attentive modelling, and may require consultation with statistical expertise before implementation. The toolbox of methods is also still expanding, and there is not always consensus on which are most appropriate. Some standard statistical software packages have either included methods or provided extra modules, but the fauna is still flourishing. Therefore it might not be surprising that in practice, available methods are not always incorporated in scientific studies, and economic evaluations are no exception. This calls for strengthening awareness of the problem on how to handle missing data in a proper way. When confronted with missing data, the problem should not be ignored as is often done. Rather, it should be assumed that missing cases might differ in analytically important ways from cases where values are present, unless there are good reasons for believing otherwise.[10]

1.2 Purpose

The purpose of this thesis is to explore the problems associated with missing data in studies included in health economic evaluations in order to promote validity and improve the base for decision making. The main focus is on adjustment methods with a realized dataset available, and the description of data collection and other methods to prepare for and prevent the problem is brief.

1.3 Methods

A literature study is carried out on economic evaluations and methods available to cope with missing data. One case where missing data was found to be a potential problem is reviewed and different methods are tested to correct for missing data.

1.4 Delimitations

The study is limited to health economic evaluation studies that involve data collected from individuals. In general, the methods described apply to continuous random variable or variables that are approximately continuous, unless else is stated.

1.5 Disposition

In chapter 2 an overview is given on the concepts of health economic evaluations. Different aspects of an economic evaluation such as perspectives, analysis method, modelling, data collection and data sources are presented here. Chapter 3 focuses on missing data in economic evaluations, and gives an introduction to quality aspects, and presents the common reasons for patterns of missing data and the underlying mechanisms. Other practical issues on preventing and preparing for missing data are also raised. The next two sections present methods used to adjust for missing data. Chapter 4 presents simple methods, and chapter 5 more advanced methods. Chapter 6 presents an earlier study, where missing data was found to be a potential problem, and different methods are applied to this study. In chapter 7 the results are presented, and chapter 8 includes a discussion and some concluding remarks.

2 Health economic evaluations

The main focus of health economic evaluations is how to allocate scarce health care resources among alternative interventions with restricted budgets and many unsatisfied goals. Thus, there is always an opportunity cost to a selected option, since other lines of action are foregone. Here, economic evaluations provide guidance to rational decision-making in distributing resources, among alternative interventions on the margin, through ranking of different interventions. The most common perspective taken is that of optimizing total health care for society as a whole, where all costs are included irrespective of the payer. Other perspectives that do not include all costs may be prone to suboptimization.

Economic evaluations use statistical methods, modelling and sensitivity analysis to calculate scenarios and uncertainty in order to establish the ranking within and between different categories of interventions and programs according to their costs and outcomes. The tasks of economic evaluations are thus to identify, measure, value and compare costs and outcomes of selected relevant alternatives.

The chapter is mainly based on [6] except for section 2.3 that is mainly based on [7]. Section 2.1 presents common analysis methods, and section 2.2 focus on costs and costing. In section 2.3 data sources and data collection is presented. Section 2.4 then focus on uncertainty and section 2.5 on modelling as used in health economic evaluations.

2.1 Analysis methods

There are several types of economic evaluations where the most common are described below; cost-minimization analysis, cost-effectiveness analysis, cost-utility analysis, and cost-benefit analysis. All types include some measure of cost and outcome where the difference between them mainly lie in the way outcomes are measured.

The simplest form of economic evaluation is the *cost-minimization analysis (CMA)*. It can be performed when there is no (significant) difference between the outcomes, and it is only costs that differ among the investigated interventions. The decision rule is then simply to choose the least costly intervention. It could be argued that this is not a full economic evaluation since no measure of outcomes is concerned. However, CMA is seldom used since very few interventions have the same outcomes. A non-significant difference in outcomes may also simply be due to a small sample size.

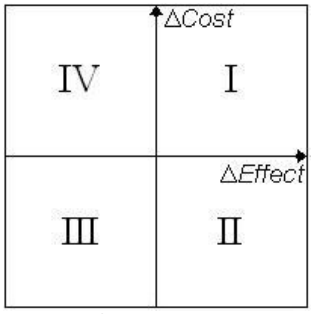
When there are differences in outcomes among the interventions, *cost-effectiveness analysis (CEA)* is more appropriate then CMA. Here costs are related to the number of units of effect, typically measured as physical health, for example number of life-years gained, number of heart attacks prevented or scores on a severity scale. The term effectiveness in this setting thus reflects the impact of an intervention of health in real practice settings and should not be mixed up with efficacy, which refers to an impact under ideal conditions, or appropriateness, which reflects a broader range of issues in deciding whether an intervention should or should not be carried out including acceptability, feasibility and cost-effectiveness.[7]

In the simplest example an alternative treatment A is compared to a standard treatment B. To decide if treatment A should be used instead of B, an incremental cost-effectiveness ratio (ICER) is used as support. The ICER is defined as the ratio of the difference in costs between the two interventions to the difference in effects between the two interventions. In this way the ratio also takes into account the alternative use of resources instead of simply relating cost to effect.

$$ICER = \frac{Cost_A - Cost_B}{Effect_A - Effect_B} = \frac{\Delta Cost}{\Delta Effect}$$

The ICER can be placed in a cost-effectiveness plane. The cost-effectiveness plane, see figure 2.1, is split into four quadrants. If the ICER ends up in quadrant II, treatment A dominates treatment B and should therefore be selected. In quadrant IV it is the other way around and treatment B dominates treatment A. In quadrant I and III it is not clear which one to select since A will be more effective but also more costly than B in quadrant I and less costly but also less effective than B in quadrant III. Which one to adopt then depends on the size of the budget, and what the preferences are to pay for extra units of effectiveness.

Figure 2.1 Cost-effectiveness plane



Adapted from [6].

A *cost-utility analysis (CUA)* can be seen as a special case of the cost-effectiveness analysis where the effect is measured as utility. The logic for this is that an intervention might affect many events, and a single measurement as used in CEA will not capture all the effects of the intervention, both in quantitative and qualitative terms. In CUA the same ICER and decision rule is used as in CEA, but utility is substituted for the outcome effect. A common measure of utility is quality-adjusted-life-years (QALYs).[2] QALYs could be measured from patients in a specific health state, or on members of the general population reflecting potential patients. The rational for using patients is that they have experience of the health state and may adapt to the illness as not foreseen by the public. Sometimes it may though be argued that patients are too sick to judge fairly about themselves or that they will tend to overstate their problems to have more attention towards it.

QALYs are calculated as number of years of life multiplied by a weight reflecting the quality of life, which is measured on a scale between *0* (reflecting death or health status equal to death) and *1* (reflecting full health). There are three ways of measuring these weights. The simplest method is the rating scale, also known as a visual-analog-scale. Here the respondent is asked to mark how the health state is valued on a single line where the endpoints are death and full health. Then the scale is normalized to *1* and the QALY weight is the marked point. The standard gamble and time-trade-off (TTO) method both involves a trade-off between two alternatives. In the standard gamble method the respondent is asked to vary a probability *p* for full health, implying probability *1-p* for immediate death, until this gamble would equal the choice of the health state at question. The choice between the gamble and the health state is limited to a specific period, usually ten years. The probability *p* will then represent the QALY weight. In TTO the respondent chooses between *T* years of being in the health state and *X* years of full health. When the respondent is indifferent between the two alternatives the QALY weight is calculated as *T/X*. The advantage of using

[2] Alternatives to QALY:s are for example healthy-years equivalents (HYEs) or disability-adjusted life years (DALYs).

QALYs is that both the qualitative aspects of reduced morbidity and the quantitative of reduced mortality are taken into account in a single measure.[8]

The theoretically superior method is *cost-benefit analysis (CBA)*, where both costs and outcomes are valued in monetary units, through monetizing all effects. The outcome measure of an intervention is then the net of costs and benefits, and can be compared to any other cost-benefit intervention irrespective of the outcome measure used. The problem with CBA is how to value outcomes in monetary terms in a valid way. In practice, the valuation is often carried out through measuring the willingness- to-pay (WTP) for different outcomes.

WTP can either be measured through revealed preferences in the market or stated preferences through direct elicitation with contingent valuation surveys. Revealed preferences could be observed through actual choices involving health risks. The contingent valuation method uses a single question to measure the respondent's maximum willingness to pay for a specified health change. The question is either open-ended or binary. Binary questions are stated as bids that the respondents have to accept or reject. To eliminate the risk of starting-point bias different bids have to be offered to different groups of the population, limiting the information from each respondent and therefore demanding a larger sample then the open-ended question. Open-ended questions generally have lower response rates, since it is considered more difficult to specify a maximum willingness to pay then to accept or reject a prespecified offer. The difference in expressed WTP has been found to merely reflect differences in the marginal utility of income than actual difference in WTP. There is also a risk that stated preferences are biased towards the respondents own interests, for example the possibility of receiving health care in the future. Therefore the revealed preferences approach is preferable. The problem is that is seldom possible to observe undistorted preferences in the market for health.[8]

Under the assumption of a constant cost per QALY irrespective of individual preferences or the size of change in QALY, a CUA can be transformed into a CBA. This is simply through multiplying the change in QALYs with the WTP per QALY. CEA and CUA are more commonly used even though the obvious advantage of CBA. One reason is that when an effectiveness measure only captures part of all of the benefits, and the rest are difficult to monetize, then CBA might involve too large of a burden to perform. Another reason might be that the benefits of intermediate effects are not always clear. A summary on type of study and associated measurement of cost and outcome is given in table 2.1.[9]

Table 2.1 Measurement of costs and outcomes in economic analyses

Type of study	Valuation of cost	Type of outcome for alternative A and B	Valuation of outcome
CMA	Monetary units	None	None
CEA	Monetary units	Single effect common to A and B	Natural units, e.g. life-years gained, disability-days saved
CUA	Monetary units	Single or multiple effects, sometimes common to A and B	Healthy years measured e.g. QALYs, HYEs
CBA	Monetary units	Single or multiple effects, sometimes common to A and B	Monetary units

Adapted from [6].

2.2 Costs and costing

In general, interest is on cumulative costs of interventions over a specified period of time, and not at a certain point in time. It is seldom feasible to estimate every single cost in detail, so focus lies on large costs. The underlying resource use is therefore measured on a patient specific level, for example days of stay in hospital or amount of drugs. For small costs, resource use could be derived from larger groups, even from population based parameters.

The process of monetizing all resource use is referred to as costing. Different types of costs are then calculated by multiplying measured resource use with assigned unit costs, within a specific period of time. They can therefore be seen as monetary units per units of time.[3] From a theoretical viewpoint, it could be argued that unit costs should be equivalent to the opportunity cost of the resource. But mostly they are derived from local hospital finance departments, or gathered from published costs in previous economic evaluations, even if these do not reflect the true market prices. In a CBA, where the WTP is part of the costs, all other costs that have to be estimated are those that are not included in the WTP of the patients who receive the treatment. It is sometimes uncertain, however, which costs the respondents actually include when they state their WTP for a treatment.

There are different ways of grouping costs. Johannesson [8] makes a distinction between programme costs, morbidity costs and mortality costs. Programme, morbidity and mortality costs can further be divided into change in health inputs, change in market production of care-receiver and change of leisure for care-receiver. It can also be practical to divide costs in direct and indirect[4], since indirect costs usually refer to those outside of the health-care sector. Direct costs include the costs of medical resource use and non-medical resource use in relation to a certain disease. Indirect costs include costs that are incurred when the affected patient is absent from productive labour due to illness or treatment. It is also possible to divide direct and indirect costs further into tangible, health care and non health-care costs. The distinction between direct and indirect cost is not always clear, and might therefore lead to confusion [6].

For several reasons the distributions of costs are often positively skewed, and individual cost data tend to vary more widely within one patient over time, compared to quality of life or clinical data. Costs can naturally never take on negative values, and in many situations especially when dealing with patient data, the more severe cases require substantially more resources then less severe cases. This is because certain expensive events only occur among a few patients. A good example here is costs due to hospitalization, which in general occur less frequent but involve large costs.[10] It could also turn out from a selective patient withdrawal when one treatment is considerably more effective than another and/or because adverse events occur more frequently in one treatment group than in another. The result is small proportions of patients being responsible for a high proportion of health care costs, which means that in a statistical model a single case can be very influential.[11] It is also expected that missing data will be more frequent among severer cases. The unhealthier a person is the less willing and able he or she will be in participating in surveys and other research related activities. The distributions of costs (and other odd distributed variables) are therefore important to consider in modelling the missing data, see also Appendix A.

[3] Costs incurred in the future can be discounted using a proper social discount rate reflecting intertemporal preferences.
[4] Direct and indirect costs should not be mixed up with specific and overhead costs as often used in cost accounting.

2.3 Sources of data and data collection

Economic evaluations often combine data collected from different sources, either as part of a research protocol (primary data), or abstracted or extrapolated from published research (secondary data). Secondary data may though not always be available on an individual basis. There are several factors that will influence the selection of data sources. Secondary data are probably less costly but also less targeted on the subject, while primary data collection can be customized but at a higher cost. Piggybacking could probably be cheaper then performing a full study on its own, and allows access to patients of interest, but may have its limitations with different and possibly conflicting interest and goals. Availability of targeted secondary data is usually limited and primary data will therefore be the only feasible way. Ethical issue will sometimes also put up constraints.[1]

Except for clinical measures, there are several modes[5] that can be used to collect both health effects and resource use data. Mixed modes, where more than one mode is used, can also be used in order to raise response rates. Data can for example be collected directly from a patient through interviews, performed face-to-face or by telephone, or through questionnaires handed out by post or hand-to-hand. Questionnaires can be used to collect data for estimation of travel, time and productivity costs. In clinical trials it is common to use case report forms, which might include both clinical information and quality of life measurements. Another form of data collection is diary cards, where a person is asked to record received care on a weekly basis. Proxy respondents are sometimes used when patients are not able to answer themselves, possibly related to their health status, and are often a relative or the hospital staff. Estimation of WTP and QALYs from a general public is typically undertaken with mail questionnaires.[1]

The following hierarchy (with relevance in decreasing order) in collecting data for CEAs is proposed by Mandelblatt et al [7]: well-conducted randomized control trials; observational data including cohort, case-control and cross-sectional studies; uncontrolled experiments; descriptive series; and expert opinions.

In a typical *randomized controlled trial (RCT)* two groups, a treatment group and a control (placebo) group are followed for a specified period. Participants are randomly allocated to the two groups, and baseline data is recorded. Under double-blind situations neither the patients nor the research staffs is aware of who is receiving treatment or placebo. The groups are then analyzed in terms of outcome defined at the outset. If the groups are similar at outset, controlled for baseline, any difference may then be assumed to be due to treatment. The use of randomization when comparing two or more groups in RCTs is an advantage to other study designs, since it then can control for known and unknown confounders and potentially remove sample selection bias[6], even if the randomization is seldom perfect since all eligible patients are not randomised. The results in a well-performed RCT are characterized by high internal validity. On the one hand, this might be a problem in an economic evaluation. The reason is that efficacy under idealized experimental conditions can differ from efficiency in real-world settings, which is usually the subject of economic evaluations. On the other hand, if the patients, the clinical practice, and other settings in a RCT are representative of the real-world settings, external validity will be high, and the results of the RCT can be directly used to assess the desired measure. However, if the settings differ from the real-world settings, it might be necessary to model the difference. RCTs are themselves in general expensive and time consuming. They also often involve too few patients or have a too short follow-up period to be satisfactory for an economic evaluation, since clinical outcome measures often require less time and sample size [4]. Other practical problems are that it can be difficult to blind

[5] The mode of data collection refers to the medium that is used to obtain the data.
[6] For an explanation of selection bias, see section 3.2.

assessors to randomisation status of patients, which might influence the outcome. There might also be a contamination effect where the control group is affected by the same treatment as the trial group, if the intervention that is studied is already widespread in clinical practice.

Cohort studies are longitudinal studies used to compare groups exposed to different factors, usually one group that has been exposed to the treatment of interest and one who is not. The groups are followed up to see whether there is a difference in outcome(s) between them. Cohort studies are usually used to study disease aetiology[7], but also to assess disease prognosis and to establish timing and direction of events. Compared to RCTs, observational studies can be better to show how something works in real-world settings. In this sense, it can be assumed to test effectiveness, and its potential for broad population inclusions. However, since treatment assignment is predetermined and not in control of a researcher, it may be subject to unknown selective problems and only known confounders can be controlled for. A cohort study often also requires long periods of follow-up. Cohort studies may also be uncontrolled.

Case-control studies will have the benefit of being made on real-life settings, and like cohort studies test effectiveness. They are usually performed on low prevalence or rare diseases where only a few subjects are available. As with other observational studies, treatment assignment is out of the researchers' control. In a case-control study, every single patient in a treatment group is matched up with a single person in a control group. The matching is done on baseline characteristics that are believed to relate to the outcome. Data is then collected retrospectively to find a difference between the treatment and the control group. It will therefore not require long observation periods, opposite to a cohort study where the groups are decided first and outcome observed retrospectively. The effort is then to find the exposure(s) that differ among the groups given the outcome. Case-control studies may suffer from selection bias in the same way as cohort studies, since it only can control for known confounders. They may also involve difficulties in selecting a proper control group. With no other data, uncontrolled case-series may be used.

Administrative data comes from registers, for example the inpatient registry ("slutenvårds-registret")[8], patient records and other databases. This can be quite economical to an economic evaluation, since no new data needs to be collected, and will generally only put a small burden on patients. But records for many characteristics are not readily accessible or do not exist, and when they do they might be collected for other purposes and standards than used in economic evaluations. Also the time period for record data and study data may not coincide [12]. Therefore, they might not provide the wanted data or may be incomplete. Confidentiality issues can also be a problem here since access to medical records often requires informed consent from study subjects [4]. Public registers are often full of information, especially in the Nordic countries, and in Sweden they contain variables such as gender, age, nationality, income, education, field of occupation, and geographical variables [13].

Expert opinions are naturally not available on an individual basis, but can sometimes be useful when few data exists. The process of combining data is then left to the judgement of the experts. This is merely quasi-scientific, and subject to bias from cognitive heuristics, but is sometimes the only way to go if there are not any other relevant data available. Techniques such as the Delphi method, which is a way of structuring a group communication process to deal with a complex problem, have been used to try to improve the quality of expert groups.

[7] Aetiology is the science of investigation of the cause or origin of diseases.
[8] "Slutenvårdsregistret" is administered by the Swedish national board of health and welfare.

Data from multiple published (and sometimes also unpublished) studies are sometimes synthesized in *systematic reviews*, in order to summarize and provide information on similarities from pooling them. A systematic review that uses quantitative methods to summarize the results is called a meta-analysis. Meta-analyses can help in evaluate effectiveness, plan new studies and can address questions not previously posed. The quality of a systematic review depends utterly on the quality of the original studies included. There is a risk that only certain studies will be part of a review because of publication bias, where only those studies that show a positive result are included.

2.4 Uncertainty and statistical inference

There are several sources of uncertainty in economic evaluations, including data inputs, methods, extrapolations, and generalizability of the results. Uncertainty about data inputs might concern the chosen discount rate, collected data on resource-use, unit costs, and estimated parameters, and uncertainty about methods include statistical models, data collection techniques, and assumptions made by the researcher. This also includes uncertainty related to missing data since this is part of the collected data.[1]

Table 2.2 Methods for handling uncertainty in economic evaluations

Type of uncertainty	Patient-level analysis (stochastic)	Decision-analytical modelling study (deterministic)
Methodological	Methodological standard/ sensitivity analysis	Methodological standard/ sensitivity analysis
Sampling variation	Statistical analysis	Not possible
Parameter uncertainty	Statistical analysis	Probabilistic sensitivity analysis
Modelling uncertainty	Sensitivity analysis	Sensitivity analysis
Generalizability/ Transferability	Sensitivity analysis	Sensitivity analysis

Adapted from [6].

Table 2.2 gives an overview of accessible methods used to handle different types of uncertainties in patient-level analyses and decision analytical modelling studies. Focus in this thesis is naturally on patient-level analyses, since decision analytical models are usually not based on a readily available dataset, even though some of the parametric information might stem from previous patient-level studies.

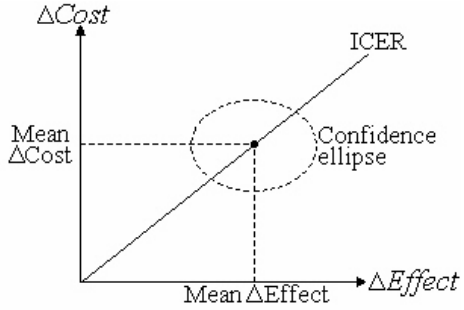
Statistical analysis with patient level data in economic evaluations is usually based on the classical school of inference, using hypothesis testing or confidence intervals to draw inference and handle uncertainty. Most of the inferential statistics used belong to the general group of general linear model, such as regression, analysis of variance, and analysis of covariance, where population parameter estimates of means, and covariances or correlations, are needed. The estimation is usually carried through with least-squares, maximum likelihood[9]. Bayesian estimation[10] is sometimes also applied.

One example of how to reflect parameter uncertainty about an ICER is confidence limits, see figure 2.2. If it is first assumed that the incremental cost and incremental effect are correlated and follow a joint normal distribution, then an elliptic confidence area can be calculated for the joint ICER based on the joint variance. This is known as the Fieller's method [14]. It is also possible to estimate the confidence area non-parametrically through bootstrapping[11], which is a simulation method whereby uncertainty is estimated with repeated resampling with replacement from the observed data. This

[9] See appendix B.
[10] See appendix C.
[11] See appendix D.

method works for the simple reason that in the absence of any better information, the data itself is the best possible estimate of the probability distribution from which it came.

Figure 2.2 Confidence area and ICER on cost-effectiveness plane



Adapted from [6].

The sensitivity analysis consists of identifying crucial parameters, and varying them over possible outcomes. Different scenarios can then be calculated, and results and conclusions can be tested for consistency and robustness. The selection of parameters and what ranges to test will ultimately depend on the researcher. A range may be based on what is meaningful, and may for example be the confidence interval, or the minimum and maximum of the parameters. Sensitivity analysis might be performed univariately, examining single parameters at a time, or multivariately, where parameters are varied simultaneously, if they are believed to interact. An extreme analysis can also be carried out, where the set of parameters that provides the worst and the best scenarios are explored. Threshold values for parameters can be found where different alternative interventions become cost-efficient. A probabilistic sensitivity analysis sets up probability distributions, as linked to the ranges of the key parameters. Random draws are then made from these distributions, in order to generate a distribution of the outcome ratio of interest. This is usually performed through Monte Carlo simulation[12]. Methodological standards usually take the form of a reference case.

2.5 Modelling in economic evaluations

There are often limitations to which scenarios can be tested directly through a study. Since medical interventions studies are more focused on intermediate rather than final outcomes,[13] economic evaluations may need to extrapolate effects to account for long-term outcomes. In piggybacked studies economic evaluation requirements may not be prioritized, or simply a small budget might put limitations to what extent different scenarios can be represented. Ethical problems might also be involved, where a patient would risk extensive suffering from specific scenarios.[1]

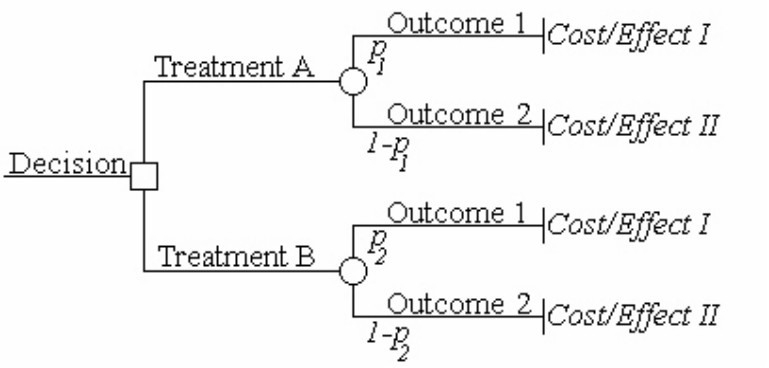
Here is where modelling takes on. Modelling allows manipulating interventions to extrapolate to different groups, points in time, disease endpoints, and putting up ranges of scenarios. It is also crucial in performing synthesis with different data sources. Modelling is entangled to the sensitivity and threshold analyses in asking what data parameters that would be needed to be considered cost-effective. All numeral information in a model should originate from the collected data, even though decision tree and Markov chain models need not be directly related to patient-level data, but survival models does.

In a *decision tree model*, nodes and branches form different paths which ends with values of a cost and an effect, see figure 2.3. The model builds on calculating expected values from rolling back the

[12] See appendix D.
[13] An example of an intermediate outcome could be blood pressure in clinical trials. The blood pressure level is in turn related to the long term risk of cardiovascular events like heart failure and stroke, which are final outcomes.

decision tree, and choosing the alternative associated with the branch showing the highest expected value. The outcome of a chance node depends on the branches associated probabilities based on earlier experience, while the outcome of a decision node depends on the expected value in that node. The model is mainly applicable for a few events that occur over time and are often used for life expectancy or quality-of-life expectancy. A decision tree can also end in a Markov chain.

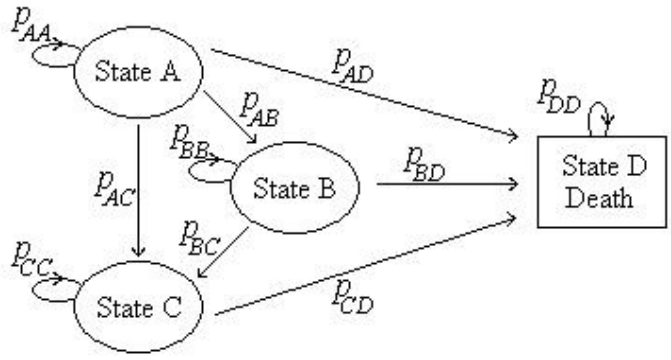
Figure 2.3 Decision-tree



Adapted from [6].

A *Markov chain model* can be viewed as a discrete-time stochastic process, also known as a random walk, dividing patients into states and time into cycles, see figure 2.4. Each state is usually associated with a cost and an outcome, which each patient accumulates during their stay in the model. At the end of the simulation, the total time that the patients in the cohort have stayed in different states are summed and weighted in order to receive expected costs and outcomes. The set of states $S=(s_1, s_2, \ldots s_n)$ could for example be different states of progression of a disease, starting with a state defined as no disease followed by different levels of progression, and with the final state defined as death. The initial probabilities and transition probabilities govern which state a patient starts, stays, and ends up in, and each patient's path among the states forms a chain. A patient that is in state s_i will move to a state s_j with probability p_{ij} in the next cycle, or remain in state s_i with probability p_{ii}. It is assumed that probabilities p_{ij} do not depend on earlier states of the chain, so that the states are statistically independent. When this assumption is released and p_{ij} are allowed to depend on earlier states in the chain, the model is called a *state transition* or *Markov chain Monte Carlo*[14] *(MCMC) model.*

Figure 2.4 Markov chain model

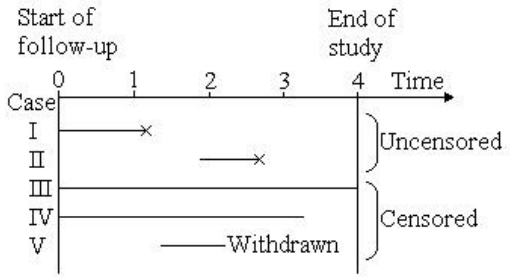


Adapted from [6].

[14] Monte Carlo simulation is further described in appendix D.

Survival analysis models calculate survival curves of survival over time, and can be used to extrapolate survival beyond a trial. Focus is on measuring time from the start of a study until the endpoint(s), where death is a natural endpoint [15]. Survival analysis often plays an important role in economic evaluations, since survival (or gained life years), measured as the area under the survival curve, can be compared between different treatments. When the exact survival time is not known for every subject, data is said to be censored, see figure 2.5.

Figure 2.5 Censored cases in survival analysis



Adapted from [16].

The three main reasons for censoring is that the study ends before the event takes place, the patient is lost due to follow-up, or withdrawn due to adverse events. The last two reasons will show up as right censoring[15].[16] Censoring is a special form of missing data, where the only thing that is known about them is that their survival time is at least as long as the recorded time until the censoring point. Several models, such as Kaplan-Meier and Cox proportional hazards model, are developed to handle censored data. These will not be covered here, and the interested reader is referred to references [10] or [17].

Drummond and Sculpher [18] are critical to the extensive use of modelling and the way sensitivity analysis is used. They argue that the choice of parameters to vary, and the range to vary them within, are often not adequately justified. Second, they also believe that sensitivity analyses are too often performed univariately. Buxton et al [19] are also critical, and propose more pragmatic clinical trials aimed at evaluating effectiveness in real-world settings rather then efficacy, enhancing both internal and external validity of economic evaluations. But they still believe that modelling is an unavoidable fact of life, and give two examples of legitimized use of models (on page 223): *"in the early stages of the development of a new health technology or intervention, when few data are available.'* and states that *'modelling should be used as a last resort, when there is no more reliable way to provide appropriate information for decision makers."*.

[15] Left censoring can also take place if a patient enters late into a study.

3 Quality concepts and missing data in economic evaluation studies

Barber et al [20] found that in the presentation of statistical methods in economic evaluations on RCTs with cost values suitable for statistical analysis, only 24 out of 45 articles identified from the database Medline in 1995, gave information on completeness of data. Three reported that data were complete, and 21 stated some missing data up to 35% of the sample. Eleven evaluations excluded subjects with MD, five checked for biases by comparing with those who had complete data, one imputed values, another one used a sensitivity analysis, and two used statistical methods for longitudinal data not requiring data to be complete. This made the authors call for revised guidelines on the presentation of missing data. In a more recent article [2] the same problem is addressed, stating that it is rare that details are given on how problems of missing data are overcome, even though the problem is believed to arise in most reported economic evaluations.

Not only is missing data common in economic evaluations studies, the amount of missing data is sometimes considerable. Some examples are contingent valuation studies estimating WTP, which often show up to 50% nonresponse, [21] and longitudinal clinical trials, which might have even higher rates of dropout [22]. There are also often reasons to believe that unobserved data differ from the observed data. It is therefore important to investigate what impact missing data will have on the inferential statistics used in economic evaluations.

Therefore, first some appropriate terminology about estimators is given in section 3.1. Then in section 3.2, error and data quality in survey data is described. In section 3.3 concepts of validity are presented, and it is shown how these relate to missing data. Further, the reasons for missing and patterns of missing data are described in sections 3.4 and 3.5. An introduction to the missing data mechanism is then given in section 3.6, with a more theoretical view in appendix E. Focus is then on how to reduce nonresponse bias through ignoring the missing data mechanism in section 3.7, and through reducing the rate of missing data in section 3.8. Section 3.1 is mainly based on [23], section 3.2 and 3.8 on [12], section 3.3 on [24], and section 3.5 on [25].

3.1 Properties of estimators

A statistic is the result of applying a statistical algorithm to a random dataset, for example a mean, median, or variance. When a statistic is used to estimate an unknown population parameter, using a random sample, it is called an estimator. The relative frequency, namely the probability distribution, of all values of a statistic that can be calculated through drawing all possible samples from a population is called a sampling distribution. Given that a random sample is drawn from a population, an estimator therefore can be treated as a random variable, with mean and variance that can be estimated from the sampling distribution. Two important properties of estimators are consistency and efficiency. Consistency means that an estimator is approximately unbiased in large samples, while efficiency is a relative measure of variance between different estimators. An estimator that is both consistent and has a variance that is at least as small as that of any other consistent estimator may be referred to as a minimum-variance-unbiased-estimator.

Bias (B) is the expected value of the difference between an estimator and the parameter θ that it is estimating. A biased estimator is thus an estimator that either overestimates or underestimates what is being measured. If a parameter is being estimated by an estimator, then the bias is the expected value of the estimator minus the true population parameter value, or equivalently the difference between the average value of the repeated estimate and the true population parameter value.[16]

[16] Bias may also refer to a sample, when members of the population it is drawn from are not equally likely to be chosen.

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Variance (Var) of an estimator of the parameter θ is the expected value of the squared difference between the value of the estimator and the expected value of the estimator. This is equivalent to the squared average value of the difference between the repeated estimates and the average value of the estimate.[17]

$$Var(\hat{\theta}) = E\left[\hat{\theta} - E(\hat{\theta})\right]^2$$

A measure of error that is used universally in statistics is the mean-squared-error (MSE). MSE can be used to decide on the quality of an estimator, and if more than one estimator is available, which one to apply. An estimator with a small MSE is thus preferable to an estimator with a large MSE.

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + B(\hat{\theta})^2$$

While variance measures the dispersion of the distribution around the expected value of an estimator, MSE measures the dispersion around the true value of the parameter, or similarly the average squared difference between the repeated estimate and the true population parameter value. It can also be shown that MSE can be computed as the squared bias plus the variance of the estimator. In order to compute the bias part of the MSE, knowledge is required about the true population parameter, while the variance part can be computed without that knowledge.

3.2 Error and data quality in survey

Naturally, the smaller the amount of error in the data that is gathered within a survey, the higher the data quality will be. Essentially the same reasoning applies to an estimate. If a proper estimator for a population parameter is based on data with a large amount of error, the estimate will also be poor.

All stages of a survey can potentially affect the data quality through nonsampling error. Nonsampling error can thus be divided into five major sources; specification error, frame error, measurement error, processing error, and nonresponse error. What is also important for the quality of an estimate is the sample size. Even if an estimate is based on high quality data, with a small sample size it can still be unreliable since sampling error will be high. Sampling error is the error in a sample estimate that is due to selecting only a random subset of a population rather than the whole population. If all members of a population were selected in a sample, assuming there are no other errors involved, the sample estimate would actually be the true parameter. But since samples are usually not perfect and smaller than the whole population, it is highly unlikely that an estimate from a sample will perfectly equal the true population parameter. The quality of an estimator will thus be a function of the total survey error, made up by the sampling error and the nonsampling error. Both these types of error can also be divided into variable error, which effects the variance of an estimator, and systematic error, which biases the estimator. Some of them will also in general pose a higher risk to the quality of an estimator, see table 3.1.

[17] Similarly, in a sample, the estimated variance can be calculated as the expected value of the squared difference between a random variable and its mean.

Table 3.1 Risk of variable and systematic error by source of error

Source of error	Risk of variable error	Risk of systematic error
Sampling error (sa)	High	Low
Specification error (sp)	Low	High
Frame error (fr)	Low	High
Nonresponse error (nr)	Low	High
Measurement error (me)	High	High
Data processing error (dp)	High	High

Adapted from [12].

Variable error appears when the values of an observed variable tend to deviate from the true values, both in a positive and negative direction, but on the average the positive and negative values cancel out. This will introduce noise into the observed variable, which limits the ability to understand what the data is telling. Variable errors can both affect (and in most cases increase) the variance of an estimator, but sometimes also bias the results. Systematic error, on the other hand, appears when the values of an observed variable tend to deviate from the true values more prevalently, in either a positive or negative direction. The values will not cancel out in the same way as with variable error, but instead produce biased estimates of the parameters of interest. For linear estimates, such as population means and proportions, systematic error is probably a worse problem then variable error, since systematic errors will bias the estimate while variable errors almost always can be reduced by increasing the sample size. For nonlinear estimates, such as correlation coefficients and standard errors, both systematic and variable errors can lead to bias. Variable error can bias these estimates towards zero, while systematic error can work in both directions.

In order to improve survey quality, the general goal is to minimize total survey error, which is analogous to minimizing the MSE of each estimator that is used. Subject to the constraints imposed by study budget and other resources available, this will thus involve finding a balance between sampling and nonsampling error, so that overall error becomes as small as possible. In general, nonsampling errors contribute more to the total error than the sampling error and can also be many times larger. All sources of nonsampling errors and the sampling error will also contribute to the MSE through the variance and bias components. Systematic errors causing bias is usually larger concern than reducing variable error. Using all components presented in table 3.1, the formula for MSE can be expanded to consider all components of variance and bias.

$$MSE = (B_{sp} + B_{fr} + B_{nr} + B_{me} + B_{dp} + B_{sa})^2 + (Var_{sp} + Var_{fr} + Var_{nr} + Var_{me} + Var_{dp} + Var_{sa})$$

Missing data is mainly related to the nonresponse error source, which thus includes nonresponse variance and nonresponse bias. Nonresponse bias is sometimes mixed up with selection bias, which is a more general measure of bias. Selection bias refers to how randomly drawn samples tend to differ systematically in characteristics from the population under study because of unequal sampling probabilities. Nonresponse bias instead refers to how nonrespondents tend to differ systematically in characteristics from respondents.[18] If a sample is a random draw from a population, nonresponse bias then will act as a selection bias.

Even though nonresponse variance in general is a problem of smaller magnitude, it might be crucial in economic evaluations, for example when piggybacked clinical studies uses sample sizes designed for clinical efficacy endpoints and outcomes, while cost comparisons in general require larger sample size because of larger variability in the underlying measures [4] The decrease in

[18] Nonresponse bias is also a function of the nonresponse rate.

sample size, and increase in nonresponse variance, can be further augmented if incomplete cases are discarded, which is often the standard procedure in much statistical software.

3.3 Concepts of validity

With high quality data from surveys, it is more likely that validity of the inference drawn, and that of the corresponding economic evaluation as a whole, will be high. Validity usually refers to the degree of truth concerning conclusions about causal relationships [24]. It can further be subdivided into four main types building on each other stepwise, with precedents as requisites for the latter.

First assume that a proposition is made about a relationship in a population that is studied. This could for example be that a certain treatment program increases the quality of life among a specific group in the population. The first type of validity is then *conclusion validity*, which refers to whether there actually is a relationship between a cause and an effect, and whether this relationship is reasonable. Now, if it is reasonable to assume that participation in the program is positively correlated to the increase in quality of life, the statement is said to have conclusion validity. The two errors that can be made about conclusion validity are either to conclude that there is no relationship when in fact there is, or conclude that there is a relationship when in fact there is not. This will depend on the statistical power and reliability of the measures. Statistical power in term depends on the sample size, the level of significance (α) and power (β), as well as the salience of the effect measured. With conclusion validity, the next question to ask is if the relationship is a causal one. Could it really be the effects of the treatment program that increases quality of life, and nothing else? If this is reasonable then the statement is said to have internal validity. The main threat to *internal validity*, concerning a causal relationship in a study used in an economic evaluation, is selection bias. The third type of validity is *construct validity*, which refers to whether the program was performed the way it was intended, and whether what was measured was what was intended to be measured. Construct validity is thus the degree to which inference can be made from the operationalization of the study to the theoretical constructs on which the operationalizations were based. For example, it could be asked if the treatment program only included persons from the specific group, or whether the measure of quality of life actually measures quality of life. If there is no doubt on this part, the proposition has construct validity. Finally, if the statement is valid for the studied group, the last question to be asked is if the relationship is true under other conditions as well, for example other populations, times, places or other settings. If it is, then the proposition is said to be *externally valid* and generalizable to those settings. External validity thus refers to in what degree conclusions drawn from the study will hold in other settings.

Nonresponse variance from missing data will affect the ability to detect relationships using estimates of the parameters of interest, and thus threatens conclusion validity. And without conclusion validity, the question is no longer relevant about internal and external validity. Nonresponse bias prohibits internal validity, because if only a selective part of the population is studied rather then the whole intended to be included, inferences will only be valid for that subgroup. And when there is uncertainty about the population under study, it also becomes difficult to tell whether the results are generalizable to other populations, and thus undermines the external validity as well.

3.4 Reasons for nonresponse and other missing data

Nonresponse is the most common reason for missing data, but there are other reasons as well. The most trivial is perhaps that data is not collected at all. Data might also be unavailable for confidentially reasons, journals might be incomplete, and medical test-results lost or unregistered. Missing data can also appear due to technical reasons, such as data acquisition failures.

Nonresponse is either full or partial. The first is referred to as unit nonresponse, where no measurement is obtained from the unit, that is, the individual or patient. There are two major reasons for unit nonresponse in surveys: refusal to respond and unavailability to respond [9]. Unavailability simply means that respondents can not be reached. This could be due to vacation, illness, work, or if they have moved. Refusal is usually more prevalent. Some common reasons for refusal are: lack of motivation (tired, bored, uninterested, rebellious); lack of time or too large survey; fear of being registered by authorities; sensitive and personal questions (privacy); illness or impaired (lack of physical ability); incomprehension or illiteracy (language problems).[26] But even if people refuse to participate, they may be more willing to provide the reason why they refuse [12]. A special type of unit nonresponse is dropout in longitudinal studies, which appears when more and more patients either withdraw or are being excluded during a study. The reason for dropout is often associated with the study, such as recovery, lack of improvement, treatment-related side-effects, unpleasant study procedures, death, but also external factors unrelated to the trial [22]. Patients might also dropout because of early recovery, but perhaps afterwards suffer from a relapse.

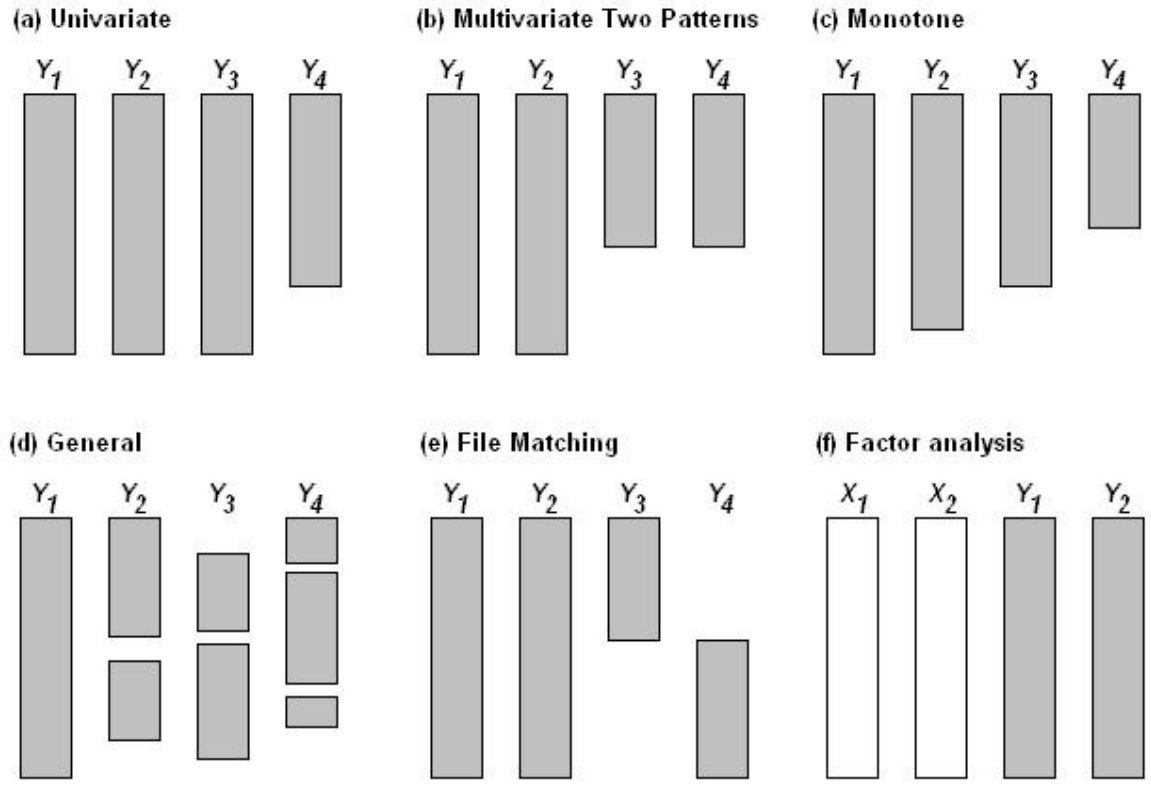
Partial nonresponse is also called item nonresponse. It refers to when measurements are obtained on only some variables of a unit. Typical reasons for item nonresponse in a questionnaire are: sensitive or difficult questions, for example open-ended questions; the respondent skip the instructions; boring and frustrating questions; time consuming and too long questionnaire; technical errors of interviewer or respondent so that answer should be deleted; by accident missed questions or failure to return to a skipped question. Sometimes answers such as “don’t know” or “no opinion” on sensitive question are treated as item nonresponse as well.[12]

3.5 Missing data patterns

A complete dataset consist of n rows and k columns, forming a matrix Y of $n*k$ cells, where each cell contains a value. Columns usually represent variables, such as resource use, and rows represent observational units, such as patients or respondents in a survey. A dataset without missing data is a complete dataset, while one with missing data is an incomplete dataset. An incomplete dataset can also be seen as a hypothetically complete dataset, consisting of observed Y_{obs} and unobserved cells Y_{mis}. By sorting the data according to the observed and unobserved cells in a hypothetically complete dataset, different patterns may emerge, see figure 3.1.

The simplest case is univariate missing data (a) where data are missing on one variable Y_4 for one or more units. In the multivariate two pattern missing data (b) data for more than one variable, Y_3 and Y_4, are missing on the same units. This pattern may emerge with unit nonresponse on a subset of individuals, and only design variables, Y_1 and Y_2, known for every unit. Monotone missing data (c) is usually dropout and withdrawal in longitudinal studies, where the monotonic wave pattern is due to more and more dropout over time. The general case (d) emerges with item nonresponse, where missing data seems to be haphazardly. In the file matching case (e), at least two variables are not observed jointly. This is common when combining datasets where only a few of the variables are common to all datasets. The factor analysis (f) is primarily a conceptual way of illustrating when missing data is represented of some latent variables, X_1 and X_2, that are not observed at all.

Figure 3.1 Typical patterns of missing data



Adapted from [25]. Rows correspond to observations, columns to variables.

3.6 Introduction to the missing data mechanism

Assumptions about the occurrence of missing data are usually formulated as a missing data mechanism, relating the observed values Y_{obs} to the unobserved values Y_{mis} of the hypothetically complete dataset. [19] There are three different types of missing data mechanisms.

When the occurrence of missing data depends on values of the data that are observed but not on unobserved data, the missing data mechanism is said to be *missing at random (MAR)*. This implies that the mechanism leading to missing data can actually be observed within the available data. Therefore, when taking account of the information in the data, nonresponse bias can be dealt with.

If the missing data mechanism is independent of both unobserved and observed data, then the missing data mechanism is said to be missing *completely at random (MCAR)*. This means that the probability for a single value to be missing is independent of the values of the hypothetically complete dataset Y, and that nonresponse bias will be absent. If all missing values in an incomplete dataset are MCAR then the dataset can be seen as a random subset of the complete data. MCAR can also be seen as a special case of MAR, without the relation to the observed data.

If the parameters that are to be estimated and the parameters that govern the missing data mechanism are distinct[20], then MAR (including MCAR) are said to be ignorable. The missing data

[19] A more theoretical explanation of the missing data mechanism is given in appendix E.
[20] See appendix E for an explanation.

mechanism is then ignored by the observed data. As a result, the missing data mechanism needs not to be modelled, in the sense that no assumptions have to be made about the missing values. This condition of ignorability is also rarely violated. And even if it was, estimation would still be valid, though not fully efficient. MAR and ignorability is therefore used interchangeably from here.[27]

Finally, if the missing data mechanism depends on unobserved values, the missing data mechanism is said to be *not missing at random (NMAR)*[21], and data will thus suffer from nonresponse bias. Then it is typically also nonignorable, which implies that both the parameters of interest and the parameters that govern the missing data mechanism have to be modelled jointly. Given a model for the data, there is only one possible ignorable mechanism, but infinitely many nonignorable.[28]

A problem is that it can not be detected from the observed data alone whether the missing data mechanism is ignorable or nonignorable. Also, when modelling a nonignorable mechanism, there is really no way to test the necessary assumptions. Ignorable mechanisms are therefore preferred, since they do not need to make the same assumptions. They are also computationally convenient since inference can be drawn without actually having to specify a correct missing data mechanism [29]. Even if it is not possible to know the true nature of a missing data mechanism, under certain assumptions there a few tests might give some guidance. To test for MCAR against MAR for example the Ridout, or Park and Davis tests might be used [30]. And under really strong assumptions, it is also possible to test MAR against NMAR with the Lagrange Multiplier test, or Heckman's test for sample selection [21].

Making the missing data ignorable through observing the difference between respondents and non-respondents in the observed data is one way of dealing with nonresponse bias. In order to truly reduce nonresponse bias the difference between observed and unobserved cases would have to be reduced, thus making the missing data mechanism MCAR. Another way of reducing nonresponse bias, which is typically easier, is to increase the response rate. However, if focus is only one either reducing the difference or reducing the nonresponse rate, nonresponse bias could actually worsen, for example if the difference between observed and unobserved cases is raised through the increase in response rate. Both means should be included within a study, and are therefore explored here in section 3.7 and 3.8.[12]

3.7 Reducing nonresponse bias through ignoring the missing data mechanism

Ignorability is usually a plausible assumption with a general missing data pattern, and a questionnaire with missed or skipped questions can even be believed to be MCAR [22]. Data that are missing by design, for example if missing data is planned, will generally be MAR [31]. Missing data due to dropout are in general more likely to be NMAR, and it can often be expected that the reason for dropout is related to the last observed values prior to the dropout [22]. The missing data mechanism may also vary due to the disease setting, and it is more likely that missing data are ignorable in adjuvant[22] settings and NMAR in advanced diseases [30]. An example is when patients with an unchanged or a worsened state stays in a study, while recovered patients dropout [28]. However, missing data is typically a mixture of both MAR and NMAR, and may differ among variables and subgroups of data [1]. Fitzmaurice et al [32] give one example where the missing data mechanism operating in the treatment group in a clinical trial is quite different from the one operating in the placebo group, owing to side-effects associated with the treatment under study.

[21] MAR and NMAR are disjunct and thus constitute all possible missing data mechanisms.
[22] The term adjuvant refers to when several therapies are combined or added to another therapy in order to enhance effectiveness, for example when surgery is combined with radiation- and chemotherapy in cancer treatment.

The missing data mechanisms are also relative, and the crucial assumption about ignorability is not that the probability of response is unrelated to the missing data, but rather that this relationship can be explained by the observed data. Thus, when observed data is rich and contains a lot of information and the data model is sufficiently complex, the dependence of the missing data mechanism upon the unobserved data should be small after conditioning on the observed data. Therefore, even when the MAR assumption is not precisely true, it will still protect from the nonresponse bias that is explained by the observed data.[29]

An effective way to make the mechanism ignorable is to gather auxiliary information from the sample frame on the sampled cases, since lower response rates usually are expected from people with certain characteristics. In many countries, and certainly in Sweden, public registers are well up to date and can often provide information on factors such as gender, age, marital status, level of education, income and geographies, without having to ask the respondents directly [13]. If a pilot study is used, effort could also be put on finding out potential reasons for missing data, and then try to measure these as part of the study [26]. In a questionnaire one might add a question requesting for the reason for nonresponse, although it is unlikely to get much information from this since someone who will not answer any other questions probably would dismiss this one as well [30]. In longitudinal clinical studies, baseline values that might help in predicting dropout could also be recorded. All reasonable attempts should also be made to retain individuals in a study, and otherwise record reasons for missing data. Through scheduling frequent measurements times it might also be possible to arrange for dropout times to coincide with measurement times.[22]

Extensive follow-up and mixed modes can certainly improve the assumption of ignorability. Respondents may for example be grouped into different waves depending on when data is collected. If initial respondents, respondents in different waves of a follow-up, and hardcore non-respondents differ systematically, this information can make ignorability a more plausible assumption.[33] Also, with intensive follow-up of a random sample of nonrespondents, the remaining nonrespondents could be assumed to be ignorable given the information from the follow-up group [29]. Unless there is some prior knowledge, the only information that is helpful in practice when making assumptions about non-ignorable missing data mechanisms is the one that is gathered about the nonrespondents [33]. Uncertainty about the missing data mechanism should also be considered as a subject for sensitivity analysis, in order to test robustness. If MAR is assumed, a good start is to check different plausible assumptions about NMAR, and whether these lead to different conclusions. It is also prudent to calculate estimates on a variety of models rather than to rely exclusively on one model, especially when the amount of missing data is considerable.[30]

3.8 Reducing nonresponse bias through increasing response rates

Several measures can be taken to increase response rates in a survey. A single measure is seldom enough to reach reasonable response rates, and rather a battery of measures is usually required. What measures to be used should naturally be based on the cause of nonresponse. This information is usually gathered from previous knowledge or pilot studies. Pilot studies are often used to identify potential problems in study design, such as evaluation of how to phrase questions in questionnaires, often in relation to response rates [26].

Nonresponse, and in particular refusal, is usually more common among people with certain characteristics. In general the experiences from many studies are that low response rates are usually expected from: metropolitan residents, single people, persons with low income, members of childless households, older people, divorced/widowed people, persons with lower educational attainment and self-employed people.[13]

Cialdini [34] found six psychological factors that had a large influence on refusal rates in surveys. The first factor is r*eciprocation*, which means that respondents will be more inclined to answer if there is some repayment or incentive, for example gifts such as lottery tickets or donations to charities. Prepaid incentives are then often preferred, since promised incentives even might reduce response rates compared to having no incentive at all. Using incentives have been found to possibly raise response rates by 10 percent. The second factor, *consistency*, suggests that compliance should be consistent with the respondents announced position, beliefs, attitudes or values. *Social validation* suggests that respondents are more willing to comply if they believe that others will. The forth factor is *authority*, which suggests that response rates will be higher if the request comes from a legitimate authority. This would explain why governmental surveys in general show higher response rates then marketing surveys. The fifth factor is *scarcity*, which suggests that people will be more willing to comply if they think it is a rare opportunity to take advantage of. Advance letters might for example be used to point out the scarcity and legitimacy of a study. The sixth factor is *liking*, which refers to when a sample member finds an interviewer appealing or similar to themselves. Liking is primarily apparent in face-to-face interviews.

Some general advice on designing a questionnaire in order to reduce nonresponse is to keep it easy for the respondent to complete it and to introduce it in an interesting and professional way. Questionnaires should also be kept as short as possible, and if possible avoid personal or sensitive questions [30]. It has also been proven in longitudinal studies to raise incentives by asking ”How likely is it that you will remain in this study through the next measurement period?” since this might both encourage future participation and give valuable information on future behaviour [35]. Item nonresponse can often be reduced through preventive work on wording and placement of question, but a large item nonresponse rate could also be an indication of a branching error in the questionnaire instrument, a flaw in interviewing instructions, unanticipated respondents sensitivity to a question, or some other system defect. In clinical studies, if the only alternative is nonresponse, proxy respondents may sometimes be used, for example if a patient is too sick or cognitively unable to answer themselves. But when proxy respondents are mixed with ordinary respondents, there is an obvious risk of selection bias.

Data collection and follow-up procedures often constitute a large part of research cost. Biemer and Lyberg [12] describe the advantages and disadvantages of seven modes of data collection. *Face-to-face interviewing* gives a maximum degree of communication and interaction, and allows direct observation, but is a high cost alternative. It also involves a risk of social desirability bias where true answers on sensitive questions might be masked, as well as interviewer variance due to personal influence from different interviewers. *Telephone interviewing* is less costly than face-to-face interviews, and can be set up more quickly. Usually the social desirability bias and interview variance are also smaller. But this is at the expense of flexibility. Telephone interviews often have to be shorter than face-to-face, and rates of “Don’t know” and “no answer” tend to be larger than in face-to-face. It is often used as a supplement to other modes in a mixed-mode. *Mail surveys* are inexpensive, and suitable for sensitive information, because there is even less social desirability bias and no interviewer effect. Visualizing is also possible here, but a certain literacy level is also needed, and long field periods are often needed to obtain acceptable response rates (usually eight weeks or more from first mail to final return). In general, low response rates are to be expected, with a great risk of considerable item nonresponse. Research suggests that 75% response rates are possible with general mailing and up to 90% to a targeted group.

Diary cards can be used to collect data on events retrospectively. The respondent is usually asked to record directly when an event has taken place, which could be on a daily basis or even more frequently. Diary cards will demand a high level of commitment, and has a high risk of the

respondent to become fatigued and less motivated. Therefore they are in general fairly short. They also involve a risk that a respondent will start conditioning on the survey, and alter behaviour. Other self-administered modes are mainly *computerized systems*. In general there is no social desirability bias or interviewer variance involved. Internet-based modes will have almost unlimited design choices, but still everybody does not have access to computers or internet access. A system with computerised record linkage can be an inexpensive follow-up method if it is available. *Administrative records* are a quite inexpensive way to collect information since they already exists, and will be no extra burden to the cases concerned. However, some problems are that supply of data may be limited or imperfectly collected or coded since the quality of administrative data are seldom accessed. Data may for example be an artefact of how data is collected, and include measurement errors, or be out of date. *Direct observation* is the recording of events that can be observed directly or using physical measurement devices, such as hospital or laboratory equipment. Trusting an observers' sense may produce similar effects as the interviewer variance, since information can be misperceived. There is also a risk that uncalibrated mechanical devices will produce systematic errors in observations.

A general accepted ordering of the modes by their expected response rates, with the highest rate first is: face-to-face interviews, telephone interviews, mail, and other self-administered modes such as Web surveys and diary surveys. A single measure is though rarely sufficient to get a reasonable response rate, and mixed modes can certainly improve response rates. A main mode is then combined with a secondary, and perhaps even a third mode. An example is to start with mail, and use it to its full potential. Then telephone interviews are added and perhaps also face-to-face interviews. Different methods can attract population groups differently, and address one or more of the many factors that are believed to contribute to nonresponse. Mixed modes can often effectively increase response rates and nonresponse bias, though theoretically, adding a mode that raise the response rates could actually increase nonresponse bias if it causes the difference in characteristics between respondents and nonrespondents to increase.

Categorizing nonresponse is useful for data collection, since follow-up methods can differ depending on the type of nonresponse. Units that are difficult to contact can be traced and contacted several times, while refusals might be handled by persuasion, and by making the participation as unburdensome as possible. Accurate documentation of the cause of dropout is also recommended. Categorization can also be helpful for making plausible assumptions about the missing data mechanism, in order to make data ignorable.

The next two chapters describe different missing data adjustment methods with a coarse division into simple (chapter 4) and advanced methods (chapter 5), similar to that used by Brand [36]. The most common assumption is that data are distributed according to a multivariate normal model. Depending on the data, assumptions usually have to be made about the distribution of the data, and also transformations of data to correct for lack of fulfilling the assumptions. Distributional assumptions and transformations are covered in Appendix A.

4 Simple missing data methods

Simple missing data methods are here divided into listwise deletion (4.1), pairwise deletion (4.2), weighting procedures (4.3) and single imputation (4.4). All four types are ignorable models. The two first are very simple methods and rests on MCAR assumptions, while weighting and imputation can be extend to MAR.

4.1 Complete-case-analysis

To perform a complete-case-analysis (CCA), sometimes called listwise deletion, all cases where at least one value is missing is simply deleted and discarded. This is often the default option in statistical programs. The main advantage of this method is that it is very easy to perform, since it immediately gives a rectangular dataset applicable for standard statistical analytical methods. All analysis will also be made with the same dataset. The main disadvantage of the method is the loss of statistical power by throwing away data. Also, if data is not MCAR and incomplete cases thus differ from complete cases, this method offers no protection against nonresponse bias.[26]

4.2 Available-case-analysis

Available-case-analysis (ACA), also called pairwise deletion, can be used to estimate models that only require input of the means and covariances (or standard errors and correlations), for example linear regression models. While CCA always exclude all cases with at least one value missing on any of the variables in the analysis-dataset, ACA exclude only those cases where the value is missing for each variable (or pair of variables) that is investigated at a time. Means are thus estimated with all available cases on each variable, and covariances are estimated with available values for each pair of variables. The estimated means and covariances can then be used to estimate the parameters of interest, typically using some statistical software. For example, with a missing data pattern similar to e) in figure 3.1 in section 3.5, means are simply calculated using all available cases on each variable, and thus only Y^1 and Y^2 would be based on all cases. Covariance between Y^1 and Y^2 would also be based on all cases. Since pair of values including missing values would be excluded from the computations, the covariance between Y^2 and Y^3, and Y^2 and Y^4 respectively, would only be based on different halves of the observations, and the covariance of Y^3 and Y^4 would not exist at all (unless there had been some overlap).[28]

There are some problems associated with ACA. First, different sample bases will be used when computing means for different variables. It is therefore not clear which sample size that should be used in calculations of the covariance with different number of observations on the two variables. For the same reason there is no guarantee (in small samples) that the covariance or correlation matrices will be positive-definite, since each element of the covariance matrix is computed from a different subset of cases [2]. It has also been shown that estimates can be seriously biased with departures from the MCAR assumption [28]. As showed by Little and Su [37], given that the missing data mechanism is MCAR, ACA can be more efficient then CCA when correlations among variables are low. But when correlations are high, CCA is more efficient.

4.3 Weighting methods

This section is mainly based on [38]. Weighting was originally used to account for unequal probabilities of survey sampling, where each sampled case was given the weight corresponding to the inverse probability of that case being selected into the sample. This logic has then been applied analogously to missing data, in order to weight for differential sample selection. The method is usually applied to unit missing data and can be extended to monotone missing data, but does not generalize readily to an arbitrary pattern of nonresponse and then may become very complex [39].

Respondents and non-respondents are first grouped together into a relatively small number of classes, based on the auxiliary variables that are predictive of the nonresponse. The nonresponders are then assigned weights zero and the weights of the responders are proportionally inflated to compensate for the nonrespondents with similar characteristics, so that total weights of the auxiliary variables are preserved from the original sample. A generalization of this method is to regress a binary variable with value *1* indicating a unit-response and value *0* indicating unit-nonresponse, on the auxiliary variables. Weights are then defined as proportional to the inverse response propensities, or by forming adjustment cells based on the propensity score. If there are many auxiliary variables, a well-judged selection is needed, based on prior knowledge or preliminary analysis. Even more sophisticated models can also be achieved using calibration [13].

When the missing data mechanism is MAR with respect to the auxiliary variables, thus they are MCAR within each adjustment cell, weighting can effectively reduce nonresponse bias. It might though be problematic to use in some cases since it is less concerned with efficiency. One example is when dealing with longitudinal censored data, where a few cases can receive dominant weights at the end of a study, or when adjustments otherwise are complex. Then, if the weights are not restricted, they may lead to estimates with very large variances. Also, conventional standard errors are inappropriate with weighting, and many complete-data analyses do not allow weighting because intervals and p-values are not immediately accounted for. If cases with some missing values are discarded, weighting will also suffer from the same disadvantages as CCA and ACA [40].

4.4 Single imputation methods

Imputation is based on filling in missing values with predicted values, in order to create an imaginary complete dataset. The predictions can be attained in numerous ways, and the immediate advantage of the method is that statistical analyses can be performed as if the dataset was complete. There are though some juridical limitations of imputation. In Sweden, as in all other States of the European Union, there are restrictions on records with data identified with a unique person, saying that an insertion of any value that is not a “true observation” is disallowed [41]. But when the unique information is suppressed, this will not be a restriction to imputation.

The first goal of imputation is to propagate uncertainty of the imputations properly, in order to estimate correct variance measures and thus minimize nonresponse variance. The second goal is to preserve important relationships between variables and of the distributions of the data, in order to minimize nonresponse bias. This must not be mistaken with predicting the missing values with the highest accuracy, or describe data in a causal manner, since imputation is a predictive and not a causal model.[39]

There are two sources of uncertainty involved in imputing missing values. The first source stems from that imputed values are not the true values, but rather predictions of the true values, and thus the parameters of the predictive model are themselves estimated and not known. When the inserted values are solely based on predictions, this will tend to inflate the correlations. This is easily seen with regression imputation, since the inserted values then will be perfect linear combinations of the predictor variables. Only some of the described methods can handle this in an acceptable way.

The second source of uncertainty stems from the fact that imaginary complete datasets are used as if they actually were complete. When the imputed values are treated as if they were the true values, sample size will be overestimated, and thus variance underestimated. This can not be solved by single imputation, but is utterly the motivation for multiple imputation, see section 5.2.[36]

Mean imputation assumes that the missing data mechanism is MCAR, and is simply applied through substituting the missing values with the mean of the observed values on the same variable. It is easy to see that this method will underestimate variability, and corrupt covariances and correlations, since it shifts all possible extreme values to the middle of the distribution of the variable and nothing is done to represent the imputation uncertainty. An analogy to mean imputation for longitudinal data with monotone missing data is called *last observation carried forward*. All missing values due to dropout are simply substituted with the last observed value from the same case. Since drop out is seldom MCAR the method will almost certainly create biases and also underestimate variances and covariances. *Substitution* can only be used with unit missing data, and is usually done in an early stage of the data collection phase. Under the assumption of MCAR, non-respondents are replaced with alternative units not previously selected into the sample, but with the same values on design variables. There is a considerable risk that the MCAR assumption is unplausible, and that the substituted units will differ from the originally selected, hence causing nonresponse bias and distorting variances. A somewhat similar method is c*old deck imputation*, which is based on substituting missing values with data from an external source, assuming that missing values are MCAR. Item nonresponse could for example be replaced with population means gathered from an external record. This method will underestimate variance the same way as mean imputation, and involves a risk of introducing bias if the MCAR assumption would not be true. Since all these methods rely on the MCAR assumption they should not be used in practice, especially not if they also distort variances measures.[25]

A more appealing method is *hot deck imputation*, which also assumes MCAR but in general also will perform well under MAR. There are several variations of hot deck imputation, but in principle the missing values are filled in with values gathered from cases that match the incomplete cases on some key variables, so called donors. The procedure for finding donors can also be refined further, in order to reflect imputation uncertainty. Hot deck imputation is primarily used with categorical data, or variables that can be treated as categorical such as ordinal data or count data that are restricted to a few set of values. When the matching is based on more than one key variable, associations between and distribution of the variables will also be preserved since the values are derived from the observed units. Another attraction of hot deck imputation is also that it does not allow "impossible" or out-of range values. Though, with limited datasets, a single unit may serve as a donor for imputed values more than once, so there is also a risk that certain units may receive abnormal weights and that estimation of variance therefore is disturbed. It could also happen that for some units there are no other unit that makes a good match.[25]

For continuous variables the preferred choice is *regression imputation*, which assumes that the missing data mechanism is MAR. With a univariate missing data pattern, a regression model is formed for the variable with missing values as dependent variable.[23] The predicted values of the regression are then imputed for the missing values. With a monotone missing data pattern, the regression can be performed with one variable at a time acting as the dependent variable. An extension to a general missing data pattern is Buck's method [42], which applies a so called sweep operator [43]. Regression imputation can also offer a solution to part of the imputation uncertainty. This is through adding an error term to the predicted value, which is intended to reflect that the true value is not known. This error term could for example be drawn from a normal distribution with variance equal to the MSE, or from the residuals of the predictive regression [2].

It is also possible to combine both hot deck and regression imputation in the same dataset, using hot deck imputation on categorical variables, and regression imputation on quantitative variables.

[23] Selection of variables for the prediction model is usually a bit different from ordinary stepwise processes or similar statistical criteria, as described in Appendix F.

Regression and hot deck imputation might also be united in a predictive mean matching imputation model, where missing values are substituted from the unit with the closest predicted value. The complete case with the closest predicted value to the incomplete case is then used as a donor for the missing values. Predictive mean matching can also be extended to bootstrap, but a problem is then to decide what cut-off point should be used to make a case contribute to a donor pool.[28]

Sometimes it might also be demanded to make the imputations a bit ad hoc, for example due to external knowledge about the missing data mechanism. One example in health economics is consumption of pharmaceuticals, where drugs may act as substitutes, or where other dependencies between different types of drugs and doses are common. In this case it might be better to concentrate on the cumulative cost of pharmaceutical consumption, instead of trying to model each drug and dose. It is also important to remember that imputation models need not to be scientifically meaningful, but the goal is to preserve important aspects of the data of interest for the analysis.[29]

5 Advanced missing data methods

The advantages of the simple missing data methods are that most of them are quite easy to perform, but they lack efficiency and are can often be biased when the assumption of MCAR does not hold. The advanced methods presented are either based on maximum likelihood (section 5.1) or multiple imputation (section 5.2), and are more efficient then the simple methods. They also assume that the missing data mechanism is MAR. The division between the two methods is not clear-cut, since maximum likelihood may be used to produce imputations, the same simulation methods may be used to propagate uncertainty, and the logarithms applied can be very similar. Both models can also be used in case of non-ignorable mechanisms, given that the missing data mechanism is modelled (5.3). This is either through selection models or pattern-mixture models.

5.1 Maximum likelihood based methods

The maximum likelihood (ML) approach to incomplete data can be used in estimation of linear models, such as regression, factor analysis and structural equation models. When the missing data mechanism is ignorable and data can be ordered according to a monotone missing data pattern, the ML estimates for a set of parameters can be found through factorizing the likelihood into conditional and marginal distributions, which are then maximized separately. It is though difficult to get good estimates of standard errors this way.[24] A method that can obtain ML estimates with general patterns of missing data is expectation maximization (EM), which bears a strong resemblance to data augmentation[25].[28]

EM is an iterative procedure which can be divided into an expectation step (E) and a maximization step (M). The E-step is simply regression imputation. Starting values are first chosen for the unknown parameters mean and covariances, using listwise or pairwise deletion, and imputations are then made using the estimated parameters. This can be though of as if means and covariances were known the missing data could be estimated. After all the missing data have been imputed, the M-step consists of calculating new values for the means and the covariances, using the imputed data along with the observed data. This can be though of as if the missing values were known, the means and covariances could be estimated. Once the M-step is finished, the E-step starts over again, and the two steps are then repeated until the parameter estimates of means and covariances converge. Convergence is reached when there are only small changes in the parameter estimates between two adjacent iterations. The final output is then the estimates of means and covariance matrix, which can be used to estimate linear models.[27]

Convergence can sometimes be slow, usually because of small samples, high rates of missing data, or models that have many parameters in relation to the amount of observed data. It is also prudent to run EM with different starting values to ensure that it does not convergence to a local maximum [2]. Another disadvantage of EM, which it shares with single imputation, is that it does not take into account that the dataset is not completely observed, and therefore usually bias variances downwards. Asymptotic covariance matrices can however be achieved using a supplemented EM algorithm[26], but with many parameters this is computationally prohibitive [36]. Also, when it is necessary to impose restrictions on the covariance matrix because the linear model that is being estimated has fewer parameters than the number of means and covariances, EM estimates will not be true ML estimates.[27]

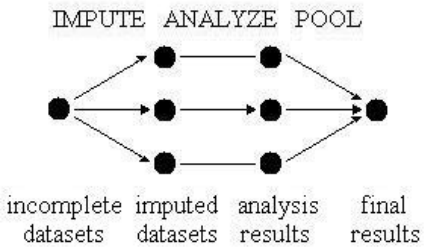
[24] See appendix B for a further explanation of maximum likelihood estimation.
[25] See appendix D.
[26] There are also other extensions of EM, such as EMis described in section 5.2.

A model that allows restrictions on the covariance matrix is full information ML also known as direct or raw ML. The estimation then takes place as part of the model of interest, the likelihood function is maximized directly, and variances are computed efficiently. There are several computer programs that can perform this and the statistical properties are also well-known. Still the method assumes that the fitted model is not false, and will therefore be sensitive to both the assumption of MAR, and that of a multivariate normal distribution.[44] As with EM, full information ML is also a large sample tool. But the behaviour of likelihood methods in small samples can be improved through extending the likelihood by a prior component for the parameters, and base inference on the resulting posterior distribution[27]. With missing data, the posterior may become extremely complex, which is why Monte Carlo techniques[28] usually are required.[36]

5.2 Multiple imputation

As described in section 4.1, single regression imputation can dampen the correlations between the imputed and predictor variables through adding an error term to the predicted values, and hot deck imputation does the same through a selecting from a pool of donors. Still, single imputation will not take account of that the dataset is used as if it were complete. Multiple imputation (MI) offers a solution to this problem through repeating the imputation process $M>1$ times. Each dataset is then analyzed, and then the results are pooled together to attain final results, see figure 5.1. In this way, each missing value in a dataset will be imputed with a set of values that differ between the datasets, in order to better reflect the uncertainty about the true value.[27] MI is therefore superior to single imputation in propagating imputation uncertainty.

Figure 5.1 The multiple imputation process



Adapted from [45].

The assumptions of MI is that the missing data mechanism is MAR, that the model used to generate the imputations is correct in some sense, and match up with the model used for the analysis in some sense [46]. The predictive model should thus be rich enough to preserve associations and relationships among the variables and the distributional assumptions should be reasonable.[29] Even though data seldom conform exactly to the assumed distributions, experience shows that MI is quite forgiving to these departures.[30]

When selecting variables in an analysis model, stepwise regression or similar statistical criteria is often used. When these criteria are applied to different datasets, this might result in different analysis models, since the imputed values in each datasets naturally will differ somewhat. To decide on the analysis model, one approach is then to look for a common model among the different datasets, or to use further imputed datasets that are not used in the pooling. Yang et al [47]

[27] See appendix C.
[28] See appendix D.
[29] See appendix F.
[30] See appendix A.

also propose two alternative methods for stepwise variable selection using MI. The first is called *impute then select* and apply a Bayesian model selection after the imputation, and the second is called *simultaneously impute and select* and integrates imputation and analysis model selection.

Ideally, the estimated parameters used in an imputation process should come from their Bayesian posterior distributions[31]. There are several algorithms based on MCMC technique that can generate these randomly, for example Data Augmentation and Gibbs sampling. An alternative method to represent imputation uncertainty is resampling methods, where for example the Approximate Bayesian Bootstrap algorithm may be suitable to hot deck imputation. Resampling methods make less modelling assumptions then MCMC, and are therefore less vulnerable to misspecification of the model. But they also demand many more imputed datasets, and their behaviour in small samples is questionable. A further description of simulation methods is given in appendix D.[25]

The pooling of results is quite simple for estimates of population means and variances. Theses can be computed directly from the estimates resulting from the analysis, while non-additive estimates such as standard deviations and correlations have to be derived from the MI:s covariances. To pool the estimates of a population mean θ, the mean of the estimates is simply computed as the average of the means in each imputed dataset.

$$\hat{\theta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i$$

To compute a variance estimate across the datasets, this involves combining the within datasets variance with the between datasets variance with a bias correction factor.

$$\hat{Var}(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M}\hat{\text{var}}(\hat{\theta}_i) + (1+\frac{1}{M})(\frac{1}{M-1})\sum_{i=1}^{M}(\hat{\theta}_i - \hat{\theta})^2$$

The approximate relative efficiency of a variance estimate, in terms of how much more precise it would have been using an infinite number of datasets, will depend on the number of imputed datasets M and the fraction of missing data γ.

$$\left(1+\frac{\gamma}{M}\right)^{-1}$$

Table 5.1 shows that even if γ is as large as 50%, approximately $M=10$ datasets is enough to reach an efficiency of 95%. With smaller amounts of missing data there is seldom need for more then about $M=5$ imputed datasets to reach this efficiency level.

Table 5.1 Approximate efficiency of MI

Number of imputed datasets (M)	Fraction of missing data (γ)				
	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Adapted from [31].

[31] See appendix C.

MI has statistical properties that closely approach that of ML, and will produce estimates that are efficient and unbiased if the underlying assumptions hold [27]. ML might also be used to produce MI:s. One example is EMis, which is an extension of the EM algorithm with importance resampling. As described in section 5.1, the resulting covariances from EM are biased downwards, so that the parameter estimates that are produced also are biased downwards. In the EMis algorithm, the importance sampling part is aimed at propagating this imputation uncertainty. An EM algorithm is first run, and the estimated vectors of means and covariance matrix are stored. A normal approximation is then made to the stored vector and matrix using an importance ratio. Further, a simulated draw is made, which is used to generate an imputed dataset. The entire importance resampling part can then be repeated to calculate *M* datasets. This algorithm is considerably faster than that of data augmentation, even though it can be shown that they produces MI:s from the same posterior. EMis might therefore be useful if convergence is slow. The algorithm may though perform poorly with small samples, many parameters that have to be estimated, or when the normal approximation is poor.[48]

The main advantage of MI is that it can be used with almost any kind of data and any kind of analysis, without needing any specialized software [46]. It is also fairly robust, because the imputation model is only applied to handle the missing part of the data and not to estimate the parameters, as compared to ML. Still, the statistical properties of MI approach that of ML. When a new kind of analysis is performed on the same data, there is usually no need to reimpute with or reformulate a prediction model.[28] MI is though primarily a large sample tool, but when combined with Bayesian methodology it can perform well in smaller samples as well.[25] A drawback of MI is that it is not exactly replicable, since it produces different datasets every time it is used [28]. Its flexibility might also look scary to novice users, which probably will need statistical expertise.

5.3 Methods with non-ignorable missing data

The basic strategy with ignorable data; to adjust for all observable differences in the data between missing and nonmissing cases and assume that all remaining differences are unsystematic, are not feasible with non-ignorable missing data, since the differences are not observable. Therefore either distributional assumptions or assumptions on associations are needed. Since the assumptions in principle will be untestable, and the information they are based on is never exact, they should be used with great caution and always be accompanied by a sensitivity analysis.[27] The methods used for ignorable missing data, preferably ML and MI, can be adapted to non-ignorable missing data as well. Given that the chosen model is correctly specified, they will also have the same optimal properties as with ignorable missing data.[28] Even when an ignorable model is assumed, but the true model is non-ignorable, ML or MI can still help to reduce bias by making use of incomplete cases [37].

Two approaches for modelling non-ignorable missing data are selection models and pattern-mixture models. The main difference between these is the way that the data and the missing data mechanism are modelled. In selection models, the joint distribution of the data and the missing data mechanism is specified in terms of the marginal distribution of the data, and the conditional distribution of the missing data mechanism given the data. Pattern-mixture models reverse this order. The joint distribution is then specified as the marginal distribution of the missing data mechanism and the conditional distribution of the data given the missing data mechanism. This means that in a selection model, data is first modelled as if there would be no missing data. Then, given the data, it is modelled whether the data is missing or not. This is similar to saying that the data first takes on a distinct values, and then depending on the values and the ways of collecting the data, the values are either observed or not observed. Pattern-mixture models may seem less

theoretically appealing, since they first model whether the data is missing, and then let this information govern the distribution of the data. Pattern-mixture models are though well suited to sensitivity analyses, since the parameters that can not be estimated from the observed data are readily determined.[39]

A well known selection model to econometricians is Heckman´s selectivity bias model. The model is designed for missing data on the dependent variable in a linear regression, and was used to estimate wages for women. A woman not being in labour is assumed to be related to a low expected wage if she would enter the job market. Therefore the reason for missing value is related to the value of the variable of interest, and thus NMAR. The probability of missing data is therefore assumed to follow a probit model, which can be maximized through ML. This model will be very sensitive to non-normal distribution of *Y*. Heckman therefore proposed a two-stage estimator which is less sensitive to non-normality.[28]

Pattern-mixture models are often used to model dropout. Generally separate distributions are assumed for responders and nonresponders, or different kinds of missing data. A datasets with the two variables Y_1 and Y_2 with missing data can lead to four possible patterns. A pattern indicator variable is denoted with *Q*. Then first both Y_1 and Y_2 can be observed ($Q=1$), none of them observed ($Q=2$), or either Y_1 ($Q=3$) or Y_2 ($Q=4$) is observed. The marginal probability of the missing data mechanism can then be assumed to be equal to what is observed, and the conditional probability of the data given the missing data mechanism assumed to follow a bivariate normal distribution. Since either Y_1 or Y_2 is missing for three of the patterns, parameters such as means and covariances can not be computed in these cases. A solution is then to impose restrictions on these unestimable parameters[32], which can be different for all patterns. The model can then be solved with ML.[28]

MI can also be used in a pattern-mixture model. The dataset is then first imputed under an ignorable model, and afterwards the imputed missing values are manipulated according to their pattern. Using the two-variable example above, assume that the missing data mechanism of Y_1 is ignorable while missing values of Y_2 are believed to be about 30% higher than that of observed values of Y_2. The imputations in the two patterns with Y_2 missing can then be (linearly) transformed to constitute different scenarios, for example with 10-50% higher values. The transformation could of course be more complicated, if knowledge suggests so, but when little is known it is probably better to use a simple and comprehensible transformation.[33]

[32] See appendix E.

6 Material

In order to exemplify the described methods, a cross-sectional study is reviewed where missing data was a potential problem. The reviewed study [49] was aimed at collecting primary data on community-based health utilities for four stages of mild cognitive impairment (MCI) and dementia, from a general population sample, using the TTO method. First an introduction is given in section 6.1 and in section 6.2 the study design and data is examined. Section 6.3 then discusses applicable adjustment methods, wherefrom an imputation model is built in section 6.4. Finally in section 6.5 (and appendix G) the results are presented.

6.1 Introduction

Dementia is a progressive brain dysfunction, which results in restrictions of daily activities and in the long term often leads to large need of care. The most common type of dementia is Alzheimer's disease. MCI is characterized by cognitive deficits, mainly short-term (episodic) memory losses, which are significant but not severe enough to hinder the functional activities of daily living. It can be hard to distinguish MCI from questionable or mild Alzheimer's disease, and there is also a considerably increased risk of conversion from MCI to Alzheimer's disease, but for the purpose of the study there was no meaning in estimating different utilites for these health states.

The Clinical Dementia Rating is a tested and validated scale that describes different stages of dementia. It is usually used as a measure of cognitive function and severity, based on a five point scale covering six domains of cognitive and functional performance: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. Each domain is accompanied by a description characterizing each stage of dementia, which can be used to rate the health status of different patients according to the scale, and then be combined into an overall score with the same steps as the individual score.

Five possible stages (with accompanied overall score) were defined to be: No dementia (0), Questionable dementia (0.5), Mild dementia (1), Moderate dementia (2), and Severe dementia (3). MCI (0.5) is defined to have the same overall score as Questionable dementia. The four stages of dementia were translated into vignettes, and were then used as scenarios in the TTO questions. For each scenario, the TTO method gave a hypothetical choice between remaining for T years (T equal to ten), in each of the stages and then die, or spending X years *(*equal to or less than ten) in full health and then die. Utility weights are then computed as X/T, so that quantity of life is traded for quality of life. These quality of life weights can be used in economic evaluations of new therapies aimed at slowing down the disease progression of dementia, and are of great importance since no medical treatment have showed any significant effects on life expectancy. The main outcome was the estimates of the health utilities for the four different stages of dementia, also cross-classified with gender and agegroups, and data collection was carried out through a postal questionnaire to the general public.

6.2 Study design, data collection and data processing

The sample frame for the survey was the Swedish public aged 45-85 years. The TTO questions were accompanied with background questions on demographics (age, gender, marital status, education and occupation), and a couple of self-assessment questions regarding health status and memory, see table 6.1. Two pilot samples (with sizes of 100 and 199 cases) were used in advance to investigate potential problems with the phrasing of the TTO-questions, and the impact of reminders and incentives on the response rate. Due to the limited budget, convenience samples based on geographical region, gender, and initial letter of surname were drawn from the internet

phonebook Eniro[33]. Neither of the samples gave information that led to any major changes in questionnaire phrasing. The first sample had no reminder, and showed very low response (25%). The other sample had both a reminder after two weeks, and incentive in form of a lottery ticket, a "trisslott", to all respondents. The resulting response rate was then 38%. Since the higher response rate was believed to be due to the reminder, the final study used two reminders and no incentive.

The main sample was then drawn from the database SPAR (Statens Person- och Adressregister)[34] in year 2004, and were stratified by age (45-54: 55-64: 65-74: 75-84), 400 from the first three groups and 600 from the oldest group. In order to correct for the unequal probabilities of sampling, weights based on the difference between the population and sampling distributions, adjusted to the amount of respondents, were used in the analysis. All 1800 questionnaires were then sent out in the beginning of April in 2004, accompanied with an advance letter stating the importance of the study. In total 749 questionnaires were returned, which implies an overall response rate of 42% or equivalently 58% unit nonresponse. There was also some item nonresponse, especially on the TTO-questions. The most common known reason for nonresponse of the scenario questions was identified to be that the respondents found it impossible to imagine how to live with a disease like dementia without any personal experience. The psychological factor authority probably had some impact on the nonresponse, since about 10-15 contacts were made from respondents on this issue. When confronted with the fact that the survey did not come from any (governmental) authorities, they decided not to participate.

Table 6.1 Rate of observed values on each variable (after exclusion of 15 cases)

Variable (Categories); VARIABLE NAME	Percentage of whole sample (n=1800)	Percentage of respondents (n=734)
Age (45-54; 55-64; 65-74; 75-84); AGE	41	100
Gender (Male; Female); GENDER	41	100
Marital status (Married/Cohabiting; Divorced; Widow/Widower; Single); MARITAL	39	96
Education (Compulsory school; High school; University/college; Other); EDUCATION	38	94
Employment (Employed; Unemployed; Studying; Retired; Early retirement/disability pension; Other); EMPLOY*	38	94
Own experience of working in health care (Yes; No) OWNEXP	39	96
Close relatives' experience of working in health care (Yes;No) RELEXP	39	96
Experience of close relative with dementia (Yes; No); EXPDEM	40	97
Assessment of one's own health (Very good; Good; Neither good nor poor; Poor; Very poor); HEALTH**	40	99
Problems with one's own memory (Not at all; Sometimes; Quite often; Very often; All the time); MEMORY***	40	98
Scenario 2 – Mild cognitive impairment; MCI	31	75
Scenario 3 - Mild dementia; MILD	27	67
Scenario 4 - Moderate dementia; MODERATE	28	69
Scenario 5 - Severe dementia; SEVERE	27	66
Mean observed response	36	89

*EMPLOY was coded as a single dummy for the categories Retired and Early retired/disability pension.
**The categories Poor(1%), and Very poor(1%), were small and coded into the category Neither good nor poor(16%).
***The categories Very often(1%), and All the time(1%), were small and coded into the category Quite often(7%).

[33] The phonebook Eniro has the web-address: http://www.eniro.se
[34] SPAR covers all people living in Sweden except those who have sent a request that their address should not be disclosed to third parties. Those people are therefore not included in the sample frame.

In the final dataset the age-stratification variable was not available, but had to be constructed from the data. Therefore it could not be controlled for consistency. Four cases though gave proof of some inconsistency, since there ages were out of the sampled range. Another nine cases had no answer to age at all. These cases were therefore excluded. To get complete observation of gender only two more cases had to be excluded. All these 15 excluded cases were also otherwise more or less incomplete. The mean response rate of all variables after this exclusion was 36%. The response rate also differed between the four stratified agegroups, where only 33% of the questionnaires where retrieved from the group 45-54, while the group 65-74 showed a response rate of 52%, see table 6.2.

Table 6.2 Distributions and size of population, sample, and respondents

	Population		Sample		Respondents		
Agegroup	N	Distribution	N	Distribution	N	Response rate in strata	Distribution
45-54	1180090	32%	400	22%	133	33%	18%
55-64	1154502	32%	400	22%	157	39%	21%
65-74	746074	20%	400	22%	206	52%	28%
75-84	582435	16%	600	33%	238	40%	32%
45-84	3663101	100%	1800	100%	734	41%	100%

Respondents seemed to be moderately representative when compared with population registers, see table 6.3. Females were only slightly overrepresented. Those belonging to the category with Married/cohabiting were overrepresented among respondents, while divorced people were only about half as many as in the general population. Respondents with compulsory school were also overrepresented, and those with high school underrepresented in the sample. Some of the difference in education is probably due to that part of the category other is classified as high school for the general population.

Table 6.3 Distribution of respondents and population on demographics

Variable	Category	Respondents	General population*
Gender	Male	46%	48%
	Female	54%	52%
Marital status	Married/Cohabiting	67%	58%
	Divorced	8%	17%
	Widow/widower	13%	10%
	Single	8%	14%
	Missing	*4%*	-
Education (45-74 years)**	Compulsory school	38%	30%
	High school	22%	42%
	University/College	27%	26%
	Other/Not classified	9%	2%
	Missing	*6%*	-

*Source [50].
**Data on education from the general population were only available for agegroup 45-74.

6.3 Adjustment methods

In order to be able to compare results of different adjustment methods, several methods were applied to the dataset. The first was a simple CCA, based on cases where all four scenario questions were answered. The methods thus assumes that respondents do not differ from nonrespondents and thus that both unit and item nonresponse is MCAR. But as seen in table 6.2, the response probabilities do actually differ between the four agegroup strata. Therefore a weighted CCA was applied, with weights based on the response rate in the different agegroups rather then only on sampling probabilities. Thus, the assumption is still that respondents do not differ from nonrespondents and that missing values are MCAR, but incorporates the difference in response rate between the four strata (and in this sense is MAR). This type of nonresponse weights were then applied, instead of the sampling weights, in all forthcoming approaches. An ACA approach with all observed values on each of the four TTO-questions was also applied. The assumption is still that data is MCAR.

Then a general MI model was built, including all four TTO-questions at the same time. Since no auxiliary variables were available for the nonrespondents (except for the stratum variable age coded from the answers), and weighting was applied to correct for unit non-response, the MI:s were restricted to imputing the item nonresponse of the 734 respondents. Two different standalone software were used, both applying regression imputation based on multivariate normal models. The first is called AMELIA, described in reference [48], and applies an EMis algorithm. The second is called NORM, described in reference [31], and use data augmentation to produce imputations. Five imputed datasets where always used.

An alternative to regression imputation would have been to use a hot-deck or predictive mean matching imputation method. Still, applying a hot-deck imputation based on integers would have meant that the TTO-questions needed to be rounded of to the nearest integer. Even though more than 90 percentages already were integers, some information would still have been lost. This alternative was therefore discarded. A predictive mean matching would perhaps have been a more interesting alternative, but since this model would have been quite complex without any readily available software, this alternative was found to be impracticable. A further alternative would have been to produce a monotone-pattern from the data, see figure 6.3, and estimate this with ML. It is though not obvious how to produce such a pattern. The most straightforward would probably have been to singly impute all missing demographic values, but only TTO-questions that are observed on MCI. Still this alternative is discarded. Since the EM algorithm is already involved in the starting phase of both MI algorithms, it will not be presented separately either. Since no linear model is estimated full information ML is not either undertaken.

In order to test robustness of the results, an NMAR assumption was also tested. This is through manipulating the imputed values of the TTO-questions in a pattern-mixture model. Two simple scenarios were considered. Either the imputed values were assumed to be 30% higher or 30% lower than the imputed values under the MAR assumption.[35] Still, it is assumed that the unit-nonresponse is simply MAR with respect to age.[36]

6.4 The imputation model

The predictor variables for the MI model were selected through the four step strategy proposed by Van Buuren et al [51] described in appendix F. AGE and GENDER were included in the first step, since these were fully observed and also used for cross-classification of the health utilities. There

[35] Thus the applied computations was: $Y_{imp}(NMAR+30\%)=Y_{imp}(MAR)*1.3$, and: $Y_{imp}(NMAR-30\%)=Y_{imp}(MAR)*0.7$.
[36] These could also have been manipulated, resulting in an even larger leverage due to the large nonresponse.

were no other variables that were known to influence the nonresponse, except for the already included variable age. The distributions of the available cases on each variable was hence compared through using binary response indicator variables on each TTO-question, with fully observed cases coded as *1* and cases with item nonresponse coded as *0*. It was found that the distributions differed between the categories of several variables, see table 6.4. Through regressing the indicator variables on the different variables, using stepwise selection, it was found that the variables AGE, EDUCATION (one dummy each for compulsory school and university/college), and MEMORY (one dummy for not at all problem) explained about one fourth of the variation in the response variables[37]. The variable EMPLOY was highly correlated with AGE, and therefore excluded. Thus, in the second step the two dummies for EDUCATION and one dummy for MEMORY were included in the imputation model.

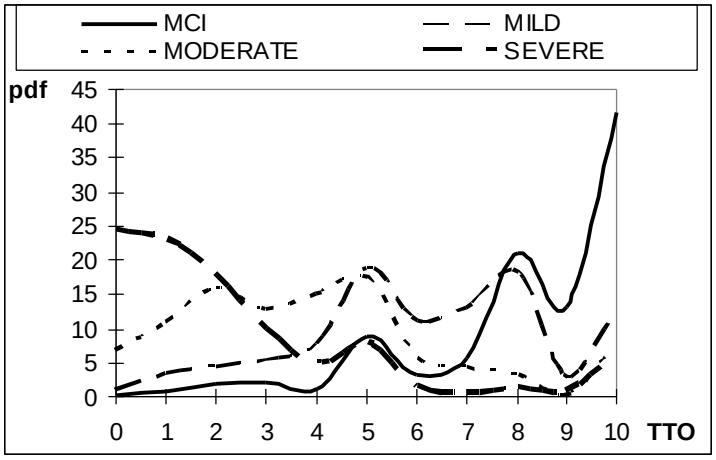
Table 6.4 Significance from chi-square-tests of association

Variable	R (MCI)	R (MILD)	R (MODERATE)	R (SEVERE)
AGE	***	***	***	***
GENDER				
MARITAL	**	**	*	*
EDUCATION	***	***	***	***
EMPLOY	***	***	***	***
OWNEXP			*	
RELEXP				
EXPDEM	**	*	*	**
HEALTH				
MEMORY	*	*		

*=significant at 0.05 level; **=significant at 0.01 level; ***=significant at 0.001 level.

In the third step, variables that explain a considerable amount of the dependent variables should be included in the imputation model. All four TTO-questions were however far from normally distributed, as can be seen from the (approximate) probability density functions of the TTO questions in figure 6.1. It can be seen that there is a considerable amount of observations in the extremes. Naturally, the largest amount of 0 values is found for SEVERE, and the largest amount of *10* values for MCI. But still there are quite some 0 values for MODERATE, and *10* values for MILD.

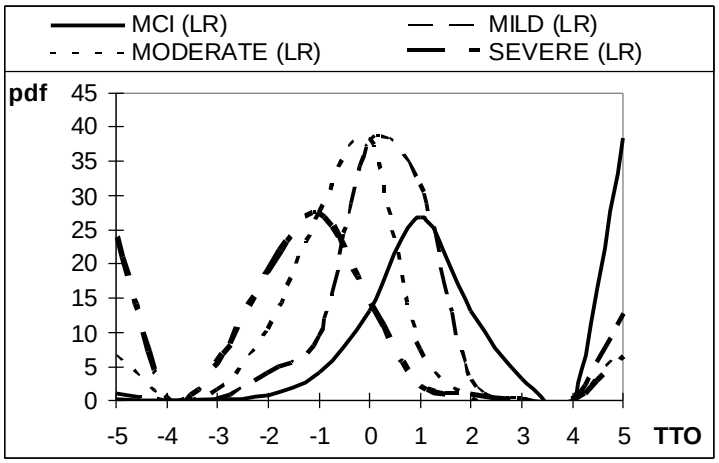
Figure 6.1 Approximate probability density functions of TTO-variables



[37] Nagelkerke R^2, a goodness-of-fit measure used in logistic regression, was 0.24; 0.24; 0.28; 0.27.

The TTO-questions were also bound to the interval *0-10* (equal to health utilities in the interval *0-1*). To enhance the normality assumptions of the TTO-questions the variables were transformed through logged odds-ratio (LR) transformations[38], see figure 6.2.

Figure 6.2 Approximate probability density functions of LR-transformed TTO-variables



The LR-transformation seemed to improve the shape of the distributions for MCI and for SEVERE, except for the large amounts on the right for MCI and on the left for SEVERE. These variables were therefore also modelled semicontinuously, with a continuous variable plus dummies for the large amount of *0* and *10* observations. Looking at the Bowman-Shelton goodness-of-fit static, see table 6.4, supported by QQ-plots, none of the continuous variables were believed to be normally distributed. With the LR-transformation, assumptions of normality were though improved for MCI and SEVERE, but including dummies did not improve the normality assumption. Only MCI and SEVERE were therefore LR-transformed.

Table 6.4 Bowman-Shelton statistic of TTO-variables

Transformation	MCI	MILD	MODERATE	SEVERE
Untransformed				
Without dummy	236	12	38	221
Logged odds-ratio transformation				
Without dummy	25	55	99	89
With *0* (left) as dummy	-	-	-	222
With *10* (right) as dummy	100	-	-	-

The hypothesis that the variable is normally distributed can be rejected at the 5% significance level if the Bowman-Shelton static is larger than 4.74 (4.83) in a sample of 400 (500).

All four TTO-variables were also found to be highly correlated, especially for neighbouring scenarios, see table 6.5. Collinearity among the variables were though quite modest when all variables were regressed on each other (condition index=10), and improved only somewhat when either MCI (8) or SEVERE (7) was discarded.[39]

[38] In order to do this, 0.1 had to be added to all 0 values and withdrawn from all 10 values, or otherwise logarithmic odds-ratios would not exist.

[39] Condition index is a diagnose measure of multicollinearity, where <10 indicates mild; 10-30 moderate; and >30 severe multicollinearity.

Table 6.5 Correlation coefficients of TTO-variables

	MCI (LR)	MILD	MODERATE	SEVERE (LR)
MCI (LR)	1	0.66***	0.33***	0.16***
MILD	0.64***	1	0.73***	0.55***
MODERATE	0.34***	0.73***	1	0.82***
SEVERE (LR)	0.21***	0.58***	0.86***	1

***=significant at 0.001 level. Pearson's correlation coefficient (parametric) in lower left corner, and Spearman's correlation coefficient (non-parametric) in upper right corner.

To find variables that could explain a considerable amount of variance, testing for differences in the TTO-values were done both parametrically and non-parametrically, see table 6.6. Except for already included or excluded variables in the model, only MARITAL was significant for more than one TTO-variable, but was superfluous when controlling for already included variables. No variable were therefore included in the third step.

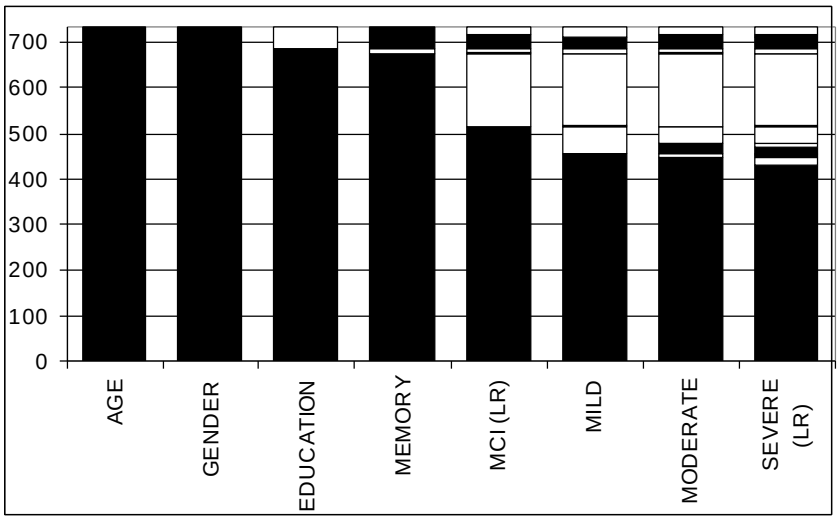
Table 6.6 Significance levels from tests of differences among groups

	MCI(LR)		Mild		Moderate		Severe (LR)	
	Param.	Non-param.	Param.	Non-param.	Param.	Non-param.	Param.	Non-param.
AGE	*	**						*
GENDER			*	*	*	**		
MARITAL			*	*			**	
EDUCATION					*	*		*
EMPLOY			*	*				*
OWNEXP								
RELEXP								
EXPDEM								
HEALTH								
MEMORY								

Parametric tests: 2-groups, t-test; >2-groups, ANOVA. Non-parametric tests: 2-groups, Mann-Whitney U; >2-groups Kruskal-Wallis H. *=significant at 0.05 level, **=significant at 0.01 level.

In the fourth step, variables selected in the second and the third step should be removed if they had many missing values. From figure 6.3 it can be seen that this is not the case for any of the included variables here, and therefore none is excluded. In total, 15 % of the values from the observed cases with variables included in the model were missing. Thus, with this amount of missing values, an approximate relative efficiency of 97% compared to infinite many imputations would be reached with five imputations. Still, this is only valid among the respondents.

Figure 6.3 Missing data patterns of variables included in the imputation model



6.5 Results from applying missing data methods

The final point estimates of health utilities and their associated estimates of variances assuming ignorable models, are presented in table G.1 in Appendix G.[40] Looking at the estimated totals point estimates, there are only tiny differences between the methods, MCI (in range 0.83-0.84); MILD (0.62-0.62); MODERATE (0.40-0.40) and SEVERE (0.25-0.27).

Point estimates of the means seem more stable between methods within groups, than between groups classified by AGE and GENDER. The largest difference between two methods are the estimates for SEVERE among females in agegroup 65-74 and males in agegroup 75-84. In the first case, MI (AMELIA) is 25% higher than ACA, and in the second MI (NORM) is 25% higher then both CCA. Within no other subgroup the point estimates differ between two methods with more than 20%. The point estimates retrieved from the two nonignorable MI methods are very similar, see table 6.7. The estimated means are about 5-10% higher (NMAR+30%) or lower (NMAR-30%) compared to MAR, and is primarily a function of the amount of missing data.

Table 6.7 Point estimates of health utilities using imputations models assuming MAR and NMAR.

		TTO-variables			
Method	Assumption	MCI	Mild	Moderate	Severe
MI (A)	MAR	0.83	0.62	0.40	0.26
	NMAR+30%	0.88	0.66	0.43	0.29
	NMAR-30%	0.78	0.57	0.37	0.24
MI (N)	MAR	0.83	0.62	0.40	0.27
	NMAR+30%	0.88	0.66	0.43	0.29
	NMAR-30%	0.79	0.57	0.38	0.24

The estimates of variances also only differ slightly for totals, as well as within most subgroups. One exception is the estimated variances for MCI in agegroup 55-64, were all estimates are between 14%-34% higher for ACA or any MI, compared to both CCA (except for ACA male which is slightly lower then for both CCA). The estimate for male 75-84 on MODERATE is also 23% higher using MI (NORM) than using ACA. No other variance estimates differ between two methods with more than 20% within a subgroup.

[40] The corresponding number of cases in each model is found in table G.2 in appendix G.

7 Summary of results

In this section an overview is given on the main findings in the literature on the issue of missing data in studies used in health economic evaluations. Focus of these studies is typically to collect health effects and resource use data, in order to estimate costs and outcomes. Individual patient data may be collected from several sources such as RCTs, cohort studies, case-control studies and administrative data. Specific data is usually collected directly from patients using interviews, questionnaires, record forms and diary cards, or from registers or proxy respondents, Questionnaires to a general public are also common in order to estimate WTP.

Missing data is a common problem within health economic evaluations, especially due to the use of surveys that may be accompanied with quite high rates of nonresponse. The two main problems associated with missing data is bias and incorrect variance estimation, both affecting the ability of drawing valid conclusions. An example of how bias appears is cost data that are skewed due to rare high cost events, where the sickest and most costly patients are more likely to have missing data. But even though bias is usually more critical, variance estimation can also be problematic. The use of piggybacked studies does for example pose problems, where the economic evaluation competes with the clinical evaluation on study design issues. Since clinical outcomes in general are less variable than cost outcomes, the requested sample size is typically lower than what is demanded for achieving significant results in the economic evaluation.

There can be several reasons for missing data. Unit nonresponse, when no information is retrieved from a questionnaire, is usually due to refusal or unavailability. Withdrawal in longitudinal studies is often related to the study, such as recovery, lack of improvement, side-effects, or even death. Item nonresponse, where one or more questions are skipped, may for example be due to difficult or sensitive questions, skipped instructions, extent of survey, technical errors by interviewers, or questions missed by accident.

A dataset with missing data can be seen as a complete dataset which is only partly observed due to an unknown stochastic mechanism. This missing data mechanism is said to be MAR when the missing data can be seen as related to the observed data. A special case of MAR is MCAR, where missing data neither is related to observed or unobserved data. A mechanism that is MAR is usually also ignorable, meaning that the parameters that govern the missing data mechanism are unrelated the parameters that are being estimated. This implies that the mechanism needs not to be modelled since any nonresponse bias can be corrected for. But when the missing data mechanism is NMAR and non-ignorable, so that missing data is related to unobserved data and the parameters of the data and the mechanism are related, data has to be modelled in order to correct for nonresponse bias.

The best solution to missing data is not to have it at all. Since this is seldom possible, one should try to measure the reason for nonresponse within the data, in order to make the ignorable assumption more plausible, and then use adjustment methods to correct for the nonresponse bias. This can for example be through collecting auxiliary variables, baseline values and other information about the cause for missing data. Follow-up methods can also be useful to make the mechanism ignorable. Since the size of the nonresponse bias is a result of both the nonresponse rate and the difference between respondents and nonrespondents, increasing response rate can also be used to reduce bias. There are several parts of a study that will affect response rates, such as pilot studies, incentives, the choice of data sources and data collection modes, and questionnaire design.

Adjustments methods include procedures based on complete-observed-units, weighting, imputation, and maximum likelihood. Simple methods such as ACA and CCA might be acceptable with very small amounts of missing data that are MCAR. It is though even better to use imputation

methods based on regression or hot deck, which may also be acceptable when missing data is MAR. Weighting is usually applied to unit nonresponse or monotone patterns, and do not generalize to arbitrary patterns of missing data that are MAR. It is effective in reducing nonresponse bias, but may under some circumstances cause excessive variance. All these simple methods have the advantage of ease of use, while their drawback is lack of efficiency and the restriction to ignorable mechanisms. With higher rates of missing data, the choice should be between ML and MI, which are more efficient than the simple methods. Still, they are primarily large sample tools, unless they are applied within a Bayesian framework.

In a theoretical sense, even though the statistical properties of MI approaches that of ML, it is possible to formulate a ML model for each missing data problem that is more efficient than the corresponding MI model. Still, there are often only a few imputed datasets needed to reach acceptable levels of efficiency using MI. In the estimation of health utilities, five imputed datasets gave an approximate relative efficiency of 97% among respondents. Also, in practice MI is easier to undertake and is less model-dependent than ML and thus is more robust to deviations from distributional assumptions of the data. In the estimation of health utilities the two MI methods differed most from CCA, both for point and variance estimates. Most differences were quite marginal though, and did not lead to any radically different conclusions.

If the missing mechanism is NMAR, ML and MI can still be used in selection or pattern-mixture models. Since the assumptions of a non-ignorable model are in practice untestable, such models should be used cautiously if they are assumed for the final model, and always be accompanied with sensitivity analysis. The robustness of ignorable assumptions can also be tested by assuming non-ignorable models, as was done within the estimation of health utilities. Even if data is non-ignorable it is seldom solely ignorable, so assuming an ignorable model can still adjust outcomes in the correct direction and thus reduce bias and improve variance estimation.

8 Discussion and concluding remarks

This thesis has presented methods aimed at handling missing data in studies used in health economic evaluations. Even if the problem is accentuated in these studies, the results are also to a large extent generalizable to other studies within social science, especially those based on survey methodology. Through limiting uncertainty that is due to missing data in these studies, decision making related to the economic evaluations will naturally improve.

The first conclusion is that there is no non-treatment solution to missing data with a realized incomplete dataset. Of course there are limitations to how much attention should be paid to the missing data, but just overlooking it will almost certainly lead to costly collected information not being utilized, and in the worst case that invalid conclusions are drawn. Awareness of the appropriate adjustment methods are therefore important to allow the use of as much information as possible in minimizing bias and estimating correct variances. It is also important to let the choice of adjustment method be based on the missing data mechanism, and to treat different methods as complimentary rather than as rivals. Since there is often uncertainty about the actual mechanism, this issue should be addressed in the sensitivity analysis. Even if the largest coverage has been of adjustment methods, the most important conclusion is that if possible missing data should be avoided, and otherwise measures should be taken to make the data ignorable. Still, when budgets for economic evaluation studies are constrained, and without awareness of what could be done about missing data, it comes as no surprise if the matter is not dealt with in a satisfactory way.

Since the sample size of the dataset used for estimation of health utilities was quite large, at least for the overall estimates, bias was of a larger concern than that of variance estimation. The possibilities of discovering potential biases were though quite small due to the limited information from the data. Some improvements could perhaps have been taken in order to prepare for missing data, even though a tight budget would probably not have allowed all of them. The first suggestion would have been to collect auxiliary variables for the whole sample, which probably would have been an admissible cost. More reminders would have been useful to try to raise response rate as well. A more extensive follow-up of a subgroup of the nonrespondents, for example with telephone interviews, would have been accurate, in order to be able to compare nonrespondents with respondents. This would also have made an assumption of ignorability more appealing. Even if the pilot samples dealt with missing data in some part, since they were aimed at the question of how to reduce the rate of missing data, it could have been further developed as to encompass qualitative aspects of missing data. This would mainly involve how to measure the reason for nonresponse as part of the final sample. Two other suggestions on tracking the missing mechanism would have been to categorize the reason for nonresponse, and to register the arrival-time of questionnaires.

Since quality aspects can never be overlooked, and the knowledge of how to deal with missing data in the best manner is probably not always satisfactory in the health economic field, my hope is that this thesis will contribute on strengthening the awareness on this topic. And even though it was difficult to discover any large biases in the examined study, there ought to be many other health economic evaluation studies, probably both with datasets that are more extensive and analysis methods that are more complex, but where little concern has been spent to the missing data. It would therefore be interesting to at least apply the sort of adjustment methods that are covered in this thesis on those studies. As long as there is an avoidable uncertainty in a study, there is always an excessive risk of making the wrong decision.

Appendix A. Distributional assumptions and transformations

Two assumptions are usually necessary when applying missing data adjustment methods made. The first is that the values on all variables of each case, that is the row of a dataset, are independent identically distributed draws[41]; and second that the draws come from some probability distribution that seem to fit the data. The most common assumption is that data is distributed according to a multivariate normal model. This would imply that all variables have normal distributions and each variable can be represented as a linear function of all the other variables together with a normal homoscedastic error term [28]. If data is solely categorical, either a multinominal or a loglinear model may also be assumed, and with a mix of both continuous and categorical data, a general location model could be a reasonable assumption. A more detailed description of these distributional models is given in [29].

Since a model is only an approximation to reality, some departures are almost always inevitable. In order to improve the fit, variables that differ from the assumed distribution may therefore be transformed. A drawback of transformation is that it often makes the model less intuitive and interpretable [1]. If an imputation model is built under the assumption of multivariate normality, skewed variables could be transformed through taking the first logarithm, and then be transformed back to the original scale after the imputation. Categorical variables could be turned into a set of dummy-variables. Ordinal variables might also be turned into dummy-variables, but are preferably imputed as continuous and afterwards rounded of to the closest category. Count data can also be imputed as continuous, but might not be needed to be rounded off unless they will be treated as count data in the analysis. Proportions, which are bound to the range between *0* and *1*, can be transformed with a logged odds-ratio transformation to make them unbounded and symmetric.

Variables with a high proportion of values equal to a single value may be modelled in two parts: a continuous variable, plus a dummy variable for the single value. Such data are common when analyzing costs, since costs can not take on negative values and often a large amount of zero costs and only a few high costs. If there are missing values in this variable, this will probably not work well because the probability mass of zero is tied to the location and scale of the continuously distributed values. Instead it should be modelled through a seemingly unrelated regression.[29] In practice this means that the semicontinuous variable is coded as a dummy variable and a continuous variable, but for baseline value of the dummy (usually *0*), the values on the continuous variable are missing. In this way the effect from the dummy on the continuous variable is omitted, but the effect from the dummy on other continuous variables is retained.[45]

Distributional assumptions may play a crucial role, and if violated the models for estimation and missing data adjustment may perform poorly. Distributional assumptions should therefore be investigated if possible, for example through plotting probability density histograms and performing test on distributions of the variables. Another solution is to use bootstrap to test robustness of the data to assumptions.[1] Considering the two advanced methodologies for missing data adjustment, MI is less sensitive then ML to the choice of model, since the model is only used to impute the data and not to estimate the parameters and are thus innocuous to variables without missing data [28]. Experience has also repeatedly shown that MI tends to be quite forgiving to departures from the imputation model. Therefore it is often acceptable to impute binary and ordered categorical variables under a multivariate normal model [31].

[41] This assumption is needed in order for the central limit theorem to apply.

Appendix B. Maximum likelihood estimation

This appendix is mainly based on [25]. Maximum likelihood (ML) is a general approach to statistical estimation. ML estimators have many desirable properties under a wide range of conditions. First they are known to be both consistent and efficient, and thus minimum-variance-unbiased-estimator. ML estimators are also asymptotically normal, which means that they in repeated sampling are normally distributed. Estimation with ML is built on the theorem that estimates should be selected so that, if they were true, the probability of observing what has been observed is maximized.

For example, assume that Y is a dataset with independent identically distributed variables, and that interest is in estimating the parameter θ. Then $f(y_i|\theta)$, where $i=1,2...n$, will be the probability density functions[42] of observing each value of Y given θ. Now, given all observations of y_i, so that Y is fixed, the likelihood for θ will be equal to the joint probability of all $f(y_i|\theta)$:

$$L(\theta \mid Y) = \prod_{i=1}^{n} f(y_i \mid \theta).$$

Estimation with ML therefore means that θ should be selected so that the probability of observing all y_i values is maximized. If Y is assumed to be a dichotomous variable with outcomes *0* and *1*, wherefrom n independent observations y_i are made, and that interest is in estimating the probability that Y is equal to 1. The likelihood $L(\theta|Y)$ is then equal to the product of the likelihoods of observing all y_i=1 and y_i=0, and is maximized when θ is selected as to maximize the function:

$$L(\theta \mid Y) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}.$$

There are a variety of techniques available to maximize the likelihood function. Usually it is achieved through differentiating the likelihood with respect to θ, set the result equal to zero, and then solve for θ. Sometimes there can be more than one ML estimate that maximizes the function, or there may not be one at all, but in the above case and in most other cases a unique ML estimate can be found.

A simple example with missing data is the univiariate missing data pattern similar to a) in figure 3.1 in section 3.5, where all n values of Y_1, $Y2$ and Y_3 are fully observed, but only $m<n$ values of Y_4. If assumed that the missing data mechanism is MAR and that the missing data mechanism is ignored, the likelihood to be maximized is $L(\theta|Y_{obs})$. In this case the likelihood can be shown to separate into the conditional distribution of the unobserved variable Y_4 given the fully observed variables and θ, and the marginal distribution of the fully observed variables given θ, which can be maximized separately.[29]

$$L(\theta \mid Y_{obs}) = \prod_{i=1}^{m} f(y_{i4} \mid y_{i1}, y_{i2}, y_{i3}, \theta) \prod_{i=1}^{n} f(y_{i1}, y_{i2}, y_{i3} \mid \theta).$$

[42] When y_i is discrete instead of continuous the density function is a mass function.

Appendix C. Bayesian estimation

This appendix is mainly based on [29]. Bayesian inference from a random sample Y about a population parameter θ assumes that probabilities for the outcome of the sample is interpreted as a degree of belief (or uncertainty) about θ, compared to classical inference, where probabilities for θ is assumed to be the outcome from repeatedly drawing a random sample.[43] This implies that in Bayesian inference, θ is allowed to follow a probability distribution for fixed Y, while in classical inference, Y is variable with θ taken to be fixed. Bayesian inference is connected to the well known Bayes theorem, which relates the conditional and marginal probabilities for the stochastic events A and B to each other:

$$P(A \mid B)P(B) = P(B \mid A)P(A) \Leftrightarrow P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Bayesian estimation can be seen as a modification (or extension) of ML estimation[44]. This can be seen in that ML estimation draws inference about θ based on the likelihood $L(\theta|Y)$ of the data, while Bayesian estimation uses the same likelihood, but extends it by the prior belief $\pi(\theta)$ about θ. In accordance with Bayes theorem this may be written as:

$$P(\theta \mid Y) = \frac{L(\theta \mid Y)\pi(\theta)}{P(Y)}.$$

The posterior density distribution $P(\theta|Y)$ will consequently be the summary of the prior knowledge $\pi(\theta)$ updated with the likelihood $L(\theta|Y)$ from the observed sample data. The term $P(Y)$ in the denominator is referred to as a normalizing constant or the marginal likelihood of the data, and can be found by integrating out the likelihood over the prior densities:

$$P(Y) = \int L(\theta \mid Y)\pi(\theta)d\theta .$$

Since Y is taken to be fixed, $P(Y)$ will be constant and can therefore be regarded as a proportionality factor. The formula for the posterior density function can thus be simplified into:

$$P(\theta \mid Y) \propto L(\theta \mid Y)\pi(\theta).$$

The updated beliefs about the prior knowledge $\pi(\theta)$, with the sample data evidence $L(\theta|Y)$ can thus be expressed as the posterior density function $P(\theta|Y)$. If interest is in estimating θ with Bayesian estimation, the posterior can be written as:

$$P(\theta,\xi \mid Y_{obs}, R) \propto L(\theta,\xi \mid Y_{obs}, R)\pi(\theta,\xi)$$ [45]

Here, ξ is the parameters that govern the response indicator matrix R, see appendix E for an extended explanation.

[43] As a consequence, Bayesian inference allows that probabilities are assigned to all propositions, while probabilities in classical inference always need to be relative to something.
[44] See appendix B.
[45] Y_{mis} have been integrated out of the joint density $Y=(Y_{obs},Y_{mis})$.

When θ and ξ are distinct so that R is ignorable, $\pi(\theta,\xi)$ can be factorized into $\pi_\theta(\theta)\pi_\xi(\xi)$. When ξ is integrated of the posterior $P(\theta,\xi|Y_{obs},R)$, R will also disappear on the right hand side, and the marginal posterior thus becomes:

$$P(\theta \mid Y_{obs}, R) = P(\theta \mid Y_{obs}) \propto L(\theta \mid Y_{obs})\pi_\theta(\theta).$$

This implies that with an ignorable missing data mechanism, all information about θ can actually be summarized in the posterior $P(\theta|Y_{obs})$. Also, when Bayesian estimation is used with an ignorable missing data mechanism, for example when producing MI:s with MCMC[46], the estimation is far more sensitive to the data model rather than the choice of the prior, and almost any reasonable prior would lead to essentially the same result. The prior is therefore usually noninformative or holds very little information.[47] Though, with small samples, sparse data, or high rates of missing values, it may be necessary to apply an informative prior (partly determined by data) [31].

On the other hand, when the missing data mechanism is nonignorable, inference has to be based on the full data posterior:

$$P(\theta,\xi \mid Y_{obs}, R) \propto L(\theta,\xi \mid Y_{obs}, R)\pi(\theta,\xi).$$

This implies that both parameters θ and ξ have to be estimated jointly, which is typically not possible solely from Y_{obs} and R. This might though be solved through imposing an informative prior distribution on (θ,ξ) or to impose a prior restriction on the joint parameter space of θ and ξ.

[46] See appendix D.
[47] A noninformative prior contains very little information about the parameters θ that are to be estimated.

Appendix D. Simulation techniques

Simulation techniques are popular in statistics and are used widely in order to examine properties of statistics and estimators that are otherwise not easily computed, for example with complex integrations or when no parametric model fits the data well. Two important techniques are resampling and Monte Carlo simulation. Both techniques are based on repetitive sampling and direct examination of the results, but while resampling methods only use already sampled observations from a population, Monte Carlo simulation sets up a data generating process[48] for a hypothetical population.

In general Monte Carlo is a technique that can be used to generate random numbers, to solve intractable mathematical systems. A typical use is the simulation of sampling distributions of estimators. The basic Monte Carlo procedure goes as follows. First a pseudo-population is specified through an algorithm that is believed to be the true data generating process. The pseudo-population thus consists of mathematical procedures that can generate sets of numbers, and these sets resemble the samples of data that are believed to be draws from the true population. A number of t samples are then created from the pseudo-population, and for each trial t, the estimator of interest is calculated and stored in a vector. The probability distribution of that vector will then be the Monte Carlo estimate of the sampling distribution for the estimator.[52]

Resampling methods are a general term for methods that compute summary estimates using redraws with replacement from an already drawn random sample. The most thorough of these methods is the bootstrap, since it usually uses many more sub-samples than other methods. The purpose is often to derive robust estimates of the variance of a population parameter when the parameter does not seem to follow any known probability distribution. Principally a bootstrap can be viewed as first treating the initial sample as the actual population, and then perform a Monte Carlo simulation where new samples are drawn from and of the same size as this new population, but where each value is replaced after it have been drawn.[53]

Focus is here on methods that can be used in missing data adjustment methods to represent imputation uncertainty. First a Markov Chain Monte Carlo (MCMC) method called data augmentation is described. More specifically this method can be used to simulate posterior distributions of the parameters θ within MI. In the description of the bootstrap, the approximate bayesian bootstrap is given as an example. This is a method that can be used to represent hot deck imputation uncertainty.

D.1 MCMC - Data augmentation

Section D.1 is mainly based on [25]. The MCMC method, which was introduced in section 2.5, can be used for sampling from probability distributions. The Markov Chain is then assumed to be a stationary[49] process of the probability distribution of interest. A brief description is given here of how the data augmentation algorithm could be used to estimate θ, which is typically the means and covariances of the variables in a dataset.

Data augmentation is aimed at making draws from $\theta^{(t)}$ from the approximation of the posterior distribution of θ with an algorithm called Imputation Posterior, where t is the number of the iteration. The process can be thought of as a Bayesian equivalent to EM, where features of the EM

[48] A data generating process is a description of a model that is believed to generate the actual data.
[49] A stochastic process is said to be (weakly) stationary if its mean and variance is constant over time, and the covariance between two time periods only depends on the distance between the two periods and not on at which time point it is calculated.

algorithm are combined with MI. Though, when EM converges to a single set of values, data augmentation will converge to a probability distribution. The imputation step corresponds to the E-step of EM, and consists of imputing the missing values with predicted values. The posterior step then corresponds to the M-step of EM, where a MCMC model estimates the posterior distribution of θ given an augmented complete dataset consisting of both observed and predicted imputed values. The word augmentation stems from the fact that the posterior is drawn from the posterior based on the artificially complete data posterior $P(\theta|Y_{obs,}Y_{mis})$[50], which is the posterior $P(\theta|Y_{obs})$ augmented by the imputed values Y_{mis}. The reason for this is that the observed data posterior $P(\theta|Y_{obs})$ generally is intractable.[51]

The logarithm is as follows. First starting values are chosen for the $\theta^{(0)}$. These values correspond to the prior of the likelihood function, which is usually noninformative with only little or no information about the parameters is included.[52] For example, in a multivariate normal model these are typically the means and covariances achieved from listwise deletion, pairwise deletion or EM. Then for each pattern of missing data, the means are covariances are used to obtain estimates of regression imputation coefficients, in order to generate predicted values for Y_{mis}. A randomly drawn residual from the variables residual normal distribution, is then added to each predicted value. This is similar to drawing $Y_{mis}^{(t+1)}$ from the density $P(Y_{mis}|Y_{obs},\theta^{(t)})$. Then the means and covariances $\theta^{(t+1)}$ are reestimated based on the dataset including both observed and imputed values, and a random draw is made from the posterior distribution of the means and covariances $P(\theta|Y_{obs},Y_{mis}^{(t+1)})$. The algorithm then returns to obtaining estimates of the regression coefficients, and the whole procedure iterates until it converges to a stationary distribution of the parameters.

After the process has converged, imputations can be drawn. The draws are either made in intervals, for example every thousand cycle starting at the thousandth cycle, or only once at the end of the process. In the latter case, multiple simulation runs are needed to produce multiple datasets. In general, the larger the amount of missing values, the more iterations are needed for convergence. In his book, Schafer [29] uses between 50 and 1000 iterations. One suggestion is to use at least as many iterations as is needed for the EM algorithm to converge. Another is to study the autocorrelations of the worst parameter, in order to detect where serial dependence disappears, see for example [31] for a further explanation.

D.2 Bootstrap – Approximate Bayesian Bootstrap

The bootstrap technique can be used to draw inference about a parameter when the underlying distribution is awkward, and it is often used to derive robust estimates of standard errors and confidence intervals of population parameters. First assume that interest is in drawing inference about the population parameter θ, based on a sample of size n. A bootstrap then consists of redrawing a sample of the same size n from the initial sample, but now with replacement so that each element is allowed to be drawn multiple times. Then the estimator of interest can be calculated for the new sample that is redrawn. This procedure should then be undertaken several times, and the probability distribution resulting from all calculated estimators may be approximated to the true sampling distribution of the estimator. Typically 50-200 bootstraps are needed to estimate standard error of an estimator, and at least 1.000 to estimate a confidence interval of an estimator.

The logic behind the bootstrap technique is that it ought to be better to draw inference from the sample at hand, which is assumed to be a random sample of the population of interest, rather than to use classical inference based on unrealistic assumptions about the data. This is critical for several

[50] The EMis procedure that is used in AMELIA is based on the same posterior distribution.[48]
[51] Here it is assumed that ξ is integrated out, see appendix C.
[52] See appendix C.

reasons, since if the sample is non-representative for the population, the bootstrap will be non-representative as well. Thus, in general a smaller sample will be less able to fully represent the full dispersion of the true distribution. Also, a non-random sample may lead to biased bootstraps. On the other hand, it has also been shown that when estimating the mean from a sample of size 20 that is drawn from a normally distributed variable, the sampling distribution of the mean will be a good approximation to the normal distribution.[53]

An application of bootstrap to missing data adjustment methods is the approximate Bayesian bootstrap, which can be used when producing hot deck MI:s. The problem here is how to select donors as to maintain the natural variability. For example, assume a univariate pattern similar to a) in figure 3.1 in section 3.5, which can be divided into adjustment cells, for example based on *gender*=Y_1, *age categorized*=Y_2, and *educational level*=Y_3. Then assume that within a specific cell, for example *men*, *<25 years*, and *primary school*, there are n_1 cases that has missing values on the fourth variable Y_4=*quartile income*, and n_2 cases that are fully observed. From the n_2 donors of complete cases, a random sample of n_2 cases with replacement should then be drawn. From the new sample, another random sample of size n_1 cases with replacement should be drawn, and then imputed for the missing values. The trick that is used here to better preserve the natural variability, is thus to make the random draw twice, compared to if n_1 cases had been drawn directly from the fully observed n_2 cases. To produce several multiply imputed datasets, the whole procedure is simply repeated.[33]

Appendix E. The missing data mechanism

Appendix E is mainly based on [29]. Assumptions about the occurrence of missing data are usually formulated as a missing data mechanism, where the incomplete dataset is only partly observed due to a possibly unknown process. It is then assumed that this process can be described as a stochastic mechanism for the relation between the missing and the observed data, possibly related to the values of the hypothetically complete dataset.

First it is assumed that interest is in estimating the parameters θ from the complete dataset *Y*, as presented in section 3.5. Then let y_{ij} be the value of the *j*:th variable on the *i*:th case, and assume that all cases are independent, identically distributed random draws from a multivariate probability distribution. The probability (or density) of the complete dataset, which is the product of the density functions for the *n* cases, conditioned on the unknown parameters θ that are to be estimated is then:

$$P(Y|\theta)=\prod_{i=1}^{n} f(y_i|\theta).$$

Then if *Y* is only hypothetically complete it can be further decomposed into observed values Y_{obs} and unobserved values Y_{mis}.[53] Also let a response indicator matrix *R* be defined as; $R_{ij}=0$ if y_{ij} is not observed, and $R_{ij}=1$ if y_{ij} is observed. A missing data mechanism can then be defined as the conditional distribution of *R* given the complete data *Y*, $P(R|Y,\xi)$, where ξ is some unknown vector of parameters that govern the missing data process. The probability density of a complete dataset can therefore be seen as the joint probability density of *Y* and *R*, or as factorized into the density of *Y* and the conditional distribution of *R* given *Y*:

$$P(Y,R|\theta,\xi)=P(Y|\theta)P(R|Y,\xi).$$

When the occurrence of missing data depends on values of the data that are observed but not on unobserved data, the missing mechanism is said to be missing at random (MAR), and has the density function $P(R|Y_{obs},\xi)$. This implies that the mechanism leading to missing data can actually be observed within the available data, since *R* depends on Y_{obs} and not on Y_{mis}.

If the missing data mechanism is independent of both unobserved and observed data, then the missing data mechanism is said to be missing completely at random (MCAR). This means that the probability for a single value to be missing is independent of the values of Y and the density function of the missing data mechanism is $P(R|\xi)$. If all missing values in an incomplete dataset are MCAR then the dataset can be thought of as a random subset of the complete data. MCAR can also be seen as a special case of MAR without relation to the observed data.

One important aspect is that of the relation between the unknown vector parameters θ and ξ. If the assumption of MAR is true, then it can be shown as with the complete data that the probability density of the observed data can be factorized into the density of Y_{obs} given θ and the conditional density of *R* given Y_{obs} and ξ:

[53] Y_{obs} could be divided further into fully observed response and design variables, and the analysis then conditioned on the design variables, in order not to put any distributional assumptions on these variables, since these are usually not random cases and thus not independent identically distributed.

$$P(Y_{obs}, R \mid \theta, \xi) = P(Y_{obs} \mid \theta) P(R \mid Y_{obs}, \xi).$$[54]

Further, if assumed that the two vector parameters θ and ξ of the two factors are distinct[55], then the ξ parameters that govern the missing data process are unrelated to the θ parameters that are to be estimated.[36] Therefore, any likelihood based inference about θ will be unaffected by ξ or by the second factor $P(R \mid Y_{obs}, \xi)$ of the factorization above. In this case MAR (and MCAR) are said to be ignorable, since the missing data mechanism is ignored by the observed data. Also, the density function of Y_{obs}, or any function proportional to it, will be the likelihood that can be maximized in order to estimate θ.

$$L(\theta \mid Y_{obs}) \propto P(Y_{obs} \mid \theta).$$

As a result, an ignorable missing data mechanism needs not to be modelled when estimating θ, in the sense that no assumptions has to be made about Y_{mis}, since all information about the mechanism is already included in the observed data. The condition of ignorability is also rarely violated. If data where MAR but not ignorable, estimation would still be valid, though not fully efficient.[27]

Finally, if the missing data mechanism depends on values of Y_{mis}, the missing data mechanism is said to be not missing at random (NMAR)[56], which means that the unobserved values is related to *R*. A distinction can also be made between models were ξ is known or not. The first may for example appear when values that are larger than a known certain value is missing, so the distribution of *R* would be determined given the hypothetically complete dataset *Y*, and thus ξ would become superfluous.

On the other hand, when some unmeasured variables are correlated to the missing values, ξ is at least partially unknown, and even with knowledge of both Y_{obs} and Y_{mis}, the distribution of *R* could not be fully determined. Therefore, the likelihood ignoring the missing data mechanism can not be used, so instead the full likelihood proportional to the complete data function (of the observed values) is required in order to estimate ξ and θ jointly.

$$L(\theta, \xi \mid Y_{obs}, R) \propto P(Y_{obs}, R \mid \theta, \xi)$$

For these reasons, with a non-ignorable missing data mechanism, both parts of the factorized complete dataset (of observed values) are typically modelled jointly. In a selection model this factorization looks like:

$$P(Y_{obs}, R \mid \theta, \xi) = P(Y_{obs} \mid \theta) P(R \mid Y_{obs}, \xi),$$

while in a pattern-mixture model it looks like:

$$P(Y_{obs}, R \mid \theta, \xi) = P(Y_{obs} \mid R, \theta) P(R \mid, \xi).$$

[54] This is through integrating Y_{mis} out of the out of the joint density $Y=(Y_{obs}, Y_{mis})$.
[55] This implies that the parameter space of (θ, ξ), that is the values that the parameters are allowed to take on, is the product of the parameter spaces of θ and ξ.
[56] MAR and NMAR are disjunct and thus constitute all possible missing data mechanisms.

Appendix F. Selection of predictor variables

When building an imputation model it is desirable to preserve the structure of the data as well as the uncertainty about this structure, even though the model needs not to be precisely correct. Using many predictors will thus minimize bias and lower the risk of excluding an important relationship in the data, which tends to makes an ignorable assumption more plausible and reduces the need to adjust for non-ignorability [2]. But it still is not feasible to include all variables because of multicollinearity, computational problems, and empty cell problems [25]. The possible lost precision when including unimportant predictors is probably small, especially if the imputed datasets will be used for several different analyses, so it is probably worse to exclude rather than to include too many predictors. Also, even if mildly important predictors are left out, MI is usually quite robust. This is because the lack of model fit will enter into the residual variance, which in a Bayesian model inflates the between-imputation variance of draws, thereby compensating for an omitted coefficient [39]. Even in the worst case with a very poorly specified model, MI can only distort part of the analysis, since it is only used to handle the missing data.

In practice the increase in explained variance is typically negligible after the best 15 variables (at most) in a linear regression, and it is therefore expedient to include no more then 15-25 variables in an MI model, including nonlinear relationships and interaction terms. To decide between first and second order (or even higher) terms, and interaction terms, R^2 can be examined by F-tests, based on only observed values. Though, it is often sensible to opt for first order models, since higher order terms tend to increase multicollinearity and consequently the variance of the parameter estimators.[36]

The imputation modelling is generally performed only once (multivariately) for the whole dataset, though it is possible to do this modelling for every incomplete variable with missing values, see for example Brand [36]. This could though lead to a cascade of auxiliary imputation problems, and in the end lead to that all variables had to be included anyway. One example which is typical in health economic is cost data that act as substitutes. Because of the dependence between them, they all should preferably be included in the same imputation model [1]. In practice there is often a small set of key variables for which imputation is needed, and focus should therefore be on modelling these key variables. Fully observed variables are always potential predictors [51].

Van Buuren et al [51] propose the following four step strategy on how to select predictor variables for a variable with missing data. First include variables that appear in the analysis model. If not, this might bias the analysis, certainly if there are strong relations in the data. Second, include all variables that are known to influence the occurrence of missing data such as stratification and reason for nonresponse on substantive grounds. Variables that differ in distribution between the missing and observed cases are also of interest. In the third step, also include variables that explain a considerable amount of the variance of the dependent variable. Such variables will help to reduce the uncertainty of the imputations. Then in the fourth and final step, variables that were added in the second and third step that have many missing values should be removed. Van Buuren et al choose to remove variables with more than half of the values missing within the subgroups of the incomplete cases.

Appendix G. Results from adjustments with missing data methods

Table G.1 Estimated means and variances for health utility estimates

			Estimated means				Estimated variances			
Age	Gender	Method	MCI	Mild	Moderate	Severe	MCI	Mild	Moderate	Severe
45-54	male	CCA (sample)	0.88	0.70	0.46	0.29	0.035	0.060	0.076	0.084
		CCA	0.88	0.70	0.46	0.29	0.035	0.060	0.076	0.084
		ACA	0.88	0.69	0.45	0.28	0.034	0.061	0.077	0.083
		MI (AMELIA)	0.88	0.68	0.45	0.28	0.036	0.063	0.077	0.083
		MI (NORM)	0.88	0.69	0.45	0.28	0.033	0.060	0.076	0.082
	female	CCA (sample)	0.86	0.64	0.41	0.26	0.030	0.055	0.070	0.078
		CCA	0.86	0.64	0.41	0.26	0.030	0.055	0.070	0.078
		ACA	0.86	0.64	0.41	0.27	0.029	0.055	0.067	0.076
		MI (AMELIA)	0.86	0.64	0.41	0.27	0.028	0.056	0.066	0.078
		MI (NORM)	0.86	0.64	0.41	0.27	0.028	0.056	0.068	0.077
	Total	CCA (sample)	0.87	0.67	0.43	0.27	0.032	0.058	0.072	0.080
		CCA	0.87	0.67	0.43	0.27	0.032	0.058	0.072	0.080
		ACA	0.87	0.67	0.43	0.27	0.031	0.058	0.071	0.079
		MI (AMELIA)	0.87	0.66	0.43	0.27	0.031	0.059	0.071	0.080
		MI (NORM)	0.87	0.66	0.43	0.27	0.030	0.058	0.072	0.079
55-64	male	CCA (sample)	0.81	0.61	0.41	0.25	0.026	0.044	0.053	0.068
		CCA	0.81	0.61	0.41	0.25	0.026	0.044	0.053	0.068
		ACA	0.81	0.61	0.41	0.25	0.025	0.044	0.053	0.067
		MI (AMELIA)	0.83	0.62	0.43	0.27	0.029	0.047	0.058	0.079
		MI (NORM)	0.82	0.61	0.41	0.27	0.032	0.048	0.054	0.071
	female	CCA (sample)	0.84	0.63	0.39	0.24	0.039	0.064	0.071	0.069
		CCA	0.84	0.63	0.39	0.24	0.039	0.064	0.071	0.070
		ACA	0.82	0.61	0.37	0.24	0.051	0.065	0.071	0.068
		MI (AMELIA)	0.82	0.60	0.37	0.24	0.052	0.064	0.073	0.070
		MI (NORM)	0.82	0.61	0.37	0.25	0.049	0.064	0.075	0.074
	Total	CCA (sample)	0.83	0.62	0.40	0.24	0.034	0.056	0.064	0.068
		CCA	0.83	0.62	0.40	0.24	0.034	0.056	0.064	0.068
		ACA	0.82	0.61	0.39	0.24	0.040	0.056	0.064	0.067
		MI (AMELIA)	0.82	0.61	0.39	0.25	0.042	0.056	0.067	0.074
		MI (NORM)	0.82	0.61	0.39	0.26	0.042	0.057	0.066	0.072
65-74	male	CCA (sample)	0.84	0.65	0.46	0.26	0.042	0.069	0.072	0.087
		CCA	0.84	0.65	0.46	0.26	0.042	0.069	0.072	0.087
		ACA	0.86	0.65	0.45	0.26	0.039	0.066	0.073	0.087
		MI (AMELIA)	0.86	0.65	0.45	0.28	0.039	0.064	0.071	0.089
		MI (NORM)	0.86	0.65	0.46	0.27	0.036	0.062	0.069	0.085
	female	CCA (sample)	0.80	0.52	0.30	0.18	0.063	0.064	0.071	0.068
		CCA	0.80	0.52	0.30	0.18	0.063	0.064	0.072	0.068
		ACA	0.82	0.53	0.32	0.17	0.055	0.064	0.072	0.064
		MI (AMELIA)	0.81	0.56	0.34	0.23	0.054	0.060	0.070	0.079
		MI (NORM)	0.82	0.55	0.34	0.23	0.054	0.066	0.076	0.079
	Total	CCA (sample)	0.82	0.58	0.38	0.22	0.053	0.070	0.077	0.078
		CCA	0.82	0.58	0.38	0.22	0.053	0.070	0.078	0.078
		ACA	0.84	0.59	0.38	0.21	0.047	0.068	0.076	0.076
		MI (AMELIA)	0.83	0.60	0.39	0.25	0.048	0.064	0.073	0.084
		MI (NORM)	0.84	0.60	0.39	0.25	0.046	0.067	0.076	0.082

Table G.1 (continued)

			Estimated means				Estimated variances			
Age	Gender	Method	MCI	Mild	Moderate	Severe	MCI	Mild	Moderate	Severe
75-84	male	CCA (sample)	0.78	0.60	0.38	0.24	0.048	0.052	0.054	0.083
		CCA	0.78	0.60	0.38	0.24	0.047	0.051	0.053	0.082
		ACA	0.77	0.61	0.39	0.25	0.050	0.051	0.057	0.080
		MI (AMELIA)	0.78	0.59	0.40	0.27	0.052	0.058	0.061	0.086
		MI (NORM)	0.79	0.62	0.42	0.30	0.055	0.060	0.066	0.097
	female	CCA (sample)	0.77	0.59	0.39	0.26	0.067	0.063	0.070	0.088
		CCA	0.77	0.59	0.39	0.26	0.066	0.061	0.068	0.086
		ACA	0.74	0.57	0.37	0.28	0.071	0.070	0.073	0.092
		MI (AMELIA)	0.76	0.56	0.37	0.27	0.067	0.068	0.072	0.088
		MI (NORM)	0.74	0.56	0.38	0.30	0.070	0.072	0.076	0.102
	Total	CCA (sample)	0.77	0.60	0.39	0.25	0.055	0.055	0.059	0.083
		CCA	0.77	0.60	0.39	0.25	0.054	0.055	0.059	0.082
		ACA	0.76	0.59	0.38	0.26	0.059	0.059	0.064	0.085
		MI (AMELIA)	0.77	0.58	0.38	0.27	0.059	0.063	0.066	0.086
		MI (NORM)	0.76	0.59	0.40	0.30	0.062	0.066	0.071	0.099
Total	male	CCA (sample)	0.84	0.65	0.44	0.26	0.036	0.058	0.065	0.079
		CCA	0.83	0.64	0.43	0.26	0.037	0.057	0.064	0.079
		ACA	0.84	0.64	0.43	0.26	0.037	0.056	0.065	0.078
		MI (AMELIA)	0.84	0.64	0.43	0.28	0.038	0.058	0.067	0.082
		MI (NORM)	0.84	0.64	0.43	0.28	0.038	0.057	0.066	0.081
	female	CCA (sample)	0.83	0.61	0.38	0.24	0.044	0.063	0.072	0.073
		CCA	0.83	0.61	0.38	0.24	0.045	0.062	0.071	0.074
		ACA	0.82	0.60	0.37	0.24	0.048	0.063	0.071	0.073
		MI (AMELIA)	0.82	0.60	0.38	0.25	0.048	0.062	0.070	0.077
		MI (NORM)	0.82	0.60	0.38	0.26	0.047	0.064	0.073	0.080
	Total	CCA (sample)	0.84	0.62	0.40	0.25	0.040	0.061	0.069	0.076
		CCA	0.83	0.62	0.40	0.25	0.041	0.060	0.069	0.076
		ACA	0.83	0.62	0.40	0.25	0.043	0.060	0.069	0.075
		MI (AMELIA)	0.83	0.62	0.40	0.26	0.043	0.060	0.069	0.079
		MI (NORM)	0.83	0.62	0.40	0.27	0.043	0.061	0.071	0.080

Table G.2 Number of cases (n) in each method

		CCA	ACA				MI
Age	Gender	(all)	MCI	Mild	Moderate	Severe	(all)
45-54	male	55	57	56	56	56	58
	female	64	70	67	69	67	75
	total	119	127	123	125	123	133
55-64	male	47	50	47	49	48	68
	female	70	78	75	78	74	89
	total	117	128	122	127	122	157
65-74	male	62	76	68	68	62	91
	female	66	83	70	75	71	115
	total	128	159	138	143	133	206
75-84	male	56	75	61	61	59	120
	female	43	62	49	50	46	118
	total	99	137	110	111	105	238
Total	male	220	258	232	234	225	337
	female	243	293	261	272	258	397
	Total	463	551	493	506	483	734

References

[1] Johnston, K. J., Buxton, M. J., Jones, D. R., Fitzpatrick, R. (1999), "Assessing the costs of healthcare technologies in clinical trials." Health Technology assessment **3**(6), 1-76.

[2] Briggs, A., Clark, T., Wolstenholme, J., Clarke, P. (2003), "Missing… presumed at random: cost-analysis of incomplete data." Health Economics **12**, 377-932.

[3] Rutten-Van Mölken, M. P. M. H., Van Doorslaer, E. K. A., Van Vliet, R. C. J. A. (1994), "Statistical analysis of cost outcomes in a randomized controlled clinical trial." Health economics **3**, 333-345.

[4] O'Sullivan, A. K., Thompson, D., Drummond, M. F. (2005), "Collection of health-economic data alongside clinical trials: is there a future for piggyback evaluations?" Value in health **8**(1), 67-79.

[5] Kleinbaum, D. G. (1996), Survival analysis: A self-learning text. New York: Springer.

[6] Drummond, M. F., Sculpher, M. J., Torrance G. W., O'Brien B. (2005), Methods for the economic evaluation of health care programmes. New York: Oxford university press

[7] Mandelblatt, J. S., Fryback, D. G., Weinstein, M. C., Russell, L. B., Gold, M. R. (1997), "Assessing the Effectiveness of Health Interventions for Cost-Effectiveness Analysis." Journal of General Internal Medicine **12**(9), 551-558.

[8] Johannesson, M. (1996), Theory and Methods of Economic Evaluation of Health Care. Dordrecht: Kluwer academic publishers.

[9] Boardman, A. E., Greenberg, D. H., Vinning, A. R., Weimer, D. L. (2001), Cost-benefit analysis - concepts and practice. New Jersey: Prentice Hall, Inc.

[10] Oostenbrink, J. B., Al, M. J., Rutten-van Mölken, M. P. M. H. (2003), "Methods to analyze cost data of patients who withdraw in a clinical trial setting." Pharmacoeconomics **21**(15), 1103-1112.

[11] Briggs, A., Gray, A. (1998), "The distribution of health care costs and their statistical analysis for economic evaluation." Journal of Health Service Research Policy **3**(4), 233-245.

[12] Biemer, P. P., Lyberg, L. E. (2003), Introduction to survey sampling. Hoboken: John Wiley & Sons, Inc.

[13] Lundström, S., Särndal, CE. (1999), "Calibration as a method for deriving nonresponse adjusted weights." Bulletin of the International statistical institute **58**(2), 313-316.

[14] Zethraeus, N., Löthgren, M. (2000)," On the equivalence of the net benefit and the Fieller's methods for statistical inference in cost-effectiveness analysis." SSE/EFI Working paper series in econometrics and finance No. 379.

[15] Bland, M. (2000), An introduction to medical statistics. New York: Oxford university press.

[16] Ekman, M. (2002), Studies in health economics – Modelling and data analysis of costs and survival. Doctoral dissertation. The economic research institute, Stockholm school of economics

[17] Collett, D. (2003), Modelling survival data in medical research. Boca Raton: Chapman & Hall/CRC

[18] Drummond, M. F., Sculpher, M. (2005), "Common methodological flaws in economic evaluations." Medical care 43(7 suppl), 5-14.

[19] Buxton, M. J., Drummond, M. F., Van Hout, B. A., Prince, R. L., Sheldon, T. A., Szucs, T., Vray, M. (1997), "Modelling in economic evaluation: an unavoidable fact of life." Health economics **6**, 217-227.

[20] Barber, J. A., Thompson, S. G. (1998), "Analysis and interpretation of cost data in randomised controlled trials: review of published studies." British Medical Journal **317**, 1195-1200.

[21] Eklöf, J. A., Karlsson S. (1999), "Testing and correcting for sample selection bias in discrete choice contingent valuation studies." SSE/EFI Working paper series in economics and finance, No 171.

[22] Diggle, P. Kenward, M. G. (1994), "Informative drop-out in longitudinal data analysis." Applied statistics 43(1), 49-93.

[23] Gujarati, D. N. (1995), Basic econometrics. Singapore: McGraw-Hill, Inc.

[24] Research methods knowledge database (2006). Available [online]: http://www.socialresearchmethods.net/kb [2006-01-01].

[25] Little, R. J. A., Rubin, D. B. (2002), Statistical analysis with missing data. Hoboken: John Wiley & Sons, Inc.

[26] Graham, J. W., Hofer, S. M., Piccinin, A. M. (1994), "Analysis with missing data in drug prevention research." National Institute on Drug Abuse Research Monograph **142**, 13-63.

[27] Allison, P. D., (2003), "Missing data techniques for structural equation modeling." Journal of abnormal psychology **112**(4), 545-557.

[28] Allison, P. D. (2001). Missing data. Thousand Oaks: Sage publications, Inc.

[29] Schafer, J. L. (1997), Analysis of incomplete multivariate data. London: Chapman & Hall.

[30] Curran, D., Bacchi, M., Schmitz S. F. H., Molenberghs, G., Sylvester, R. J. (1998), "Identifying the types of missingness in quality of life data from clinical trials." Statistics in medicine **17**, 739-756.

[31] Schafer, J. L., Olsen, M. K. (1998), " Multiple Imputation for Multivariate missing data problems: a data analyst's perspective." Multivariate behavioural research, 33, 545-571.

[32] Fitzmaurice, G. M., Molenberghs, G., Lipsitz, S. R. (1995), "Regression models for longitudinal binary responses with informative drop-outs." Journal of the royal statistical society. Series B (Methodological) **57**(4), 691-704.

[33] Rubin, D. B. (1987), Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons, Inc.

[34] Cialdini, R.B., (1993), Influence: the new psychology of modern persuasion, New York: Morrow.

[35] Demirtas, H. Schafer, J. L. (2003), "On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out." Statistics in medicine **22**, 2553-2575.

[36] Brand, J. P. L. (1999), Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. Doctoral Dissertation. Department of Medical Informatics, Erasmus University

[37] Little, R. J. A., Su, HL. (1989), "Item nonresponse in panel surveys." In Kasprzyk, D., Duncan, G., Kalton, G. (eds.), Panel surveys. New York: John Wiley & Sons, Inc.

[38] Little, R. J. A., (1988), "Missing data adjustments in large surveys." Journal of business and economic statistics **6**(3), 287-296.

[39] Rubin, D. B. (1996), "Multiple imputation after 18+ years." Journal of the american statistical association. **91**(434), 473-489.

[40] Raghunathan, T. E. (2004), "What do we do with missing data? Some options for analysis of incomplete data." Annual reviews in public health 25, 99-117.

[41] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[42] Buck, S.F. (1960), "A method of estimation of missing values in multivariate data suitable for use with an electronic computer", Journal of the royal statistical society, B, **22**, 302-306.

[43] Dempster, A.P. (1969), Elements of continuous multivariate analysis. Reading: Addison-Wesley.

[44] Statistical services FAQ Texas University (2005). Available [online]: http://www.utexas.edu/its/rc/answers/general/gen25.html [2005-01-01].

[45] Schafer, J. L., Olsen, M. K. (1999), Modeling and imputation of semicontinuous survey variables. In Proceedings of Federal Committee on Statistical Methodology (FCSM) Reseach Conference, Nov, 1999. 15 http://www.fcsm.gov/events/papers1999.html

[46] Allison, P. D. (2000), "Multiple imputation for missing data: A cautionary tale." Sociological methods and research **28**, 301-309.

[47] Yang, X., Belin, T. R., Boscardin, W. J. (2005), "Imputation and variable selection in linear regression models with missing covariates." Biometrics **61**, 498-506.

[48] King, G., Honaker, J., Joseph, A., Scheve, K. (2001), "Analyzing incomplete political science data: an alternative algorithm for multiple imputation." American political science review. **95**(1), 49-69.

[49] Ekman, M., Berg, J., Wimo, A., Jönsson, L., McBurney, C. (2006) "Health utilities in mild cognitive impairment and dementia: A population study in Sweden". Unpublished paper.

[50] Statistics in Sweden (SCB). Available [online]: http://www.scb.se [2006-01-01].

[51] Van Buuren, S., Boshuizen, H. C., Knook, D. L. (1999), "Multiple imputation of missing blood pressure covariates in survival analysis" Statistics in medicine **18**, 681-694.

[52] Mooney, C.Z., (1997), Monte Carlo simulation. Thousand Oaks: Sage publications, Inc.

[53] Mooney, C.Z., Duval, R.D. (1993), Bootstrapping –a nonparametric approach to statistical inference. Newbury Park: Sage publications, Inc.