Confounding in real-life cost-effectiveness studies: assessing the validity and efficiency of different correction techniques

L.M.A. Goossens[1]

C.W.M. van Gils[1]

M.J. Al[1]

C.A. Uyl-de Groot[1]

W.K. Redekop[1]

1 Institute for Medical Technology Assessment, Erasmus University Rotterdam,

Contents

1 Introduction 2
1.1 Background 2
1.2 Objective 5
1.3 Research Questions 5

2 Methods 6
2.1 Description of the empirical studies 7
2.2 Creation and description of 'complete' dataset 12
2.3 Creation and description of 'biased' datasets 16
2.4 Application of bias correction techniques 18
2.5 Comparisons of correction methods 22

3 Results 24

4 Discussion 37

5 References 43

Appendix 1 45

Appendix 2 53

1 Introduction

1.1 Background

(Problem)
Under current Dutch policy regarding expensive medicines, the cost-effectiveness of expensive new intramural drugs must be evaluated in a `real world' setting. The results of these observational studies are likely to have more external validity than randomised controlled trials. Nevertheless, internal validity must also be maintained, because problems with internal validity can greatly distort the results about the effectiveness and cost-effectiveness of a medicine.

In ideal randomised experiments (i.e. large sample size, perfectly randomized, no loss to follow up, full adherence to assigned treatment and no measurement error) internal validity is assured. Association is causation: association measures can be directly interpreted as effect measures. In contrast, observational studies can be subject to many forms of bias, which result in alternative causes of association.(1,2) In general, there are three major types of bias: confounding, selection bias, and information bias.(3)

In daily clinical practice treatments are not randomly assigned. In fact, it's the job of medical doctors to assign treatment based on patient characteristics, and preferences. Treatments are assigned based on patient prognosis. Consequently, the prognosis of patients receiving a treatment will often differ systematically from that of patients not receiving a treatment. Many epidemiologists refer to this phenomenon as confounding while econometricians label this endogeneity. Either way, if confounding is not removed or reduced, the real treatment effect will be either underestimated or overestimated.

Another problem, called "selection bias" by epidemiologists and some statisticians/ econometricians, occurs when people with a non average treatment effect are more likely to be included in the study than others. While the causes of selection bias may differ between experimental and observational studies, observational studies are not inherently more sensitive to selection bias than experimental studies.
The third form of bias is information bias, a bias that occurs from measurement errors. Observational studies may be more sensitive to measurement errors regarding either the treatment received, the outcome of interest, or confounding variables. Measurement errors regarding the treatment or the outcome can either over- or underestimate the treatment effect.

The problems described above occur as a result of systematic errors. However, random errors due to sampling variability can also occur.(4) Especially in the setting of 'real world' studies we expect the inter-individual variability to be greater than what would be seen in a carefully selected study population in an RCT. This greater variability would mean that a larger sample size would be required in real-world studies than in RCTs in order to achieve an acceptable level of statistical power.

Several fundamentally different correction methods have been proposed in the medical literature to overcome confounded results. Some are straightforward, whereas others are more sophisticated. They vary in both their ability to provide valid estimates and their ability to provide precise estimates. Correction methods with lower reliability will require larger databases and may therefore be more costly - or even impossible - to perform. To date we lack good information about the advantages and disadvantages of different correction methods to perform real-life (phase IV) cost effectiveness analyses.
We propose to conduct a study that examines both the theoretical value and the practical value of all major methods to correct for confounding. The assessment of the practical value of the methods will involve using a simulated data set, closely related to real world data, to study how well the different correction methods perform in a 'real world' setting.

(Relevance)
Policymakers face two important questions when they read reports of cost-effectiveness analyses: 1) are the results about effectiveness and cost-effectiveness valid and reliable? and, 2) are the results relevant for the decisions that have to be made?
With the current Dutch policy regarding expensive medicines, the cost-effectiveness of new expensive drugs needs to be evaluated within the first four years following addition to the list of expensive medicines. Clearly, this evaluation in 'real world' daily practice can provide policymakers with results that are much more relevant and applicable to the current situation than economic evaluations piggy-backed onto randomized controlled trials before the new medicine has even been used in daily practice. What is most worrisome here is the question about whether the study findings are valid. As described above, several forms of bias can threaten internal validity, and thereby result in an overestimation or underestimation of both the treatment effect and the cost-effectiveness of a medicine.

The major shortcoming of observational studies is the absence of a random assignment of treatment; this will lead to confounding bias. This study focused on this type of bias.

The focus of medical research regarding correction for confounding mostly lies on the accurate measurement of the clinical treatment effect.(2,5,6)

However, economic evaluation studies aim to estimate costs and incremental cost-effectiveness ratios (ICERs). These have different properties than data on clinical effectiveness. Cost data are typically skewed. When the incremental effects are small, ICERs are very sensitive to small changes in incremental costs. The ICER is negative for very attractive treatments - when a cost-saving coincides with a positive clinical effect – and very unattractive treatments – which are more expensive than the comparator and lead to inferior effects.
Only a few publications mention confounding correction techniques in relationship to the cost-effectiveness of new treatments (for instance(7)). There is no literature on techniques for the kinds of medicines listed on the expensive medications list and this study aimed to fill this gap somewhat.

To date, validity and reliability have been the only concerns in methodological research. However, practical feasibility plays an important role as well. For this reason, we not only evaluated the different methods in terms of validity and reliability, but also considered other criteria. This study therefore investigated how different confounding bias correction techniques performed according to criteria such as validity, precision, costs, data requirements, and expertise requirements, in the setting of observational cost-effectiveness research, with a focus on the list of expensive medicines.

1.2 Objective

The objective of our research was to investigate how various existing methods for dealing with confounding in observational studies perform with respect to validity and precision of effectiveness as well as health economic outcomes measures. Feasibility regarding costs, data requirements and expertise requirements of different methods were also compared. The main focus of this study was on bias resulting from 'confounding by indication'. The methods investigated included regression, propensity score matching, inverse probability weighting, and instrumental variable regression. The comparison of methods was used to formulate criteria to select a method for the analysis of observational data in a specific study. The final goal was to generalize our findings to currently listed expensive medicines.

1.3 Research Questions

1. To what extent can different statistical methods provide valid and reliable estimates of incremental treatment costs, effects and cost-effectiveness when using 'real world' observational data in cost-effectiveness studies?
2. How do the different methods compare with regard to feasibility (i.e. costs, expertise and data requirements)?
3. To what extent can optimal and inappropriate techniques be identified in different study settings?
4. How can these conclusions be applied to the categories of medicines on the list of expensive medicines?

2 Methods

To appropriately compare the different correction methods, we performed a simulation study instead of using empirical data to obtain an ideal hypothetical research situation. In this hypothetical research situation the patient would live two lives at the same time. The patient's clone 1 would receive the new treatment while clone 2 would receive the control treatment. As a consequence, a dataset based on this hypothetical setting would contain two outcome observations for each patient, one for each treatment. We referred to such a dataset as the 'complete' dataset. Having a complete dataset made it possible to assess the ability of the methods to correct for confounding. The 'real' effect could be established in the complete dataset and was used to compare the results of the different correction which were applied to biased samples from the complete set.

The drawback of creating a dataset instead of using empirical data may put some limits to the generalizability of the results. Therefore, it was of utmost importance that this created dataset reflected reality as much as possible. Our simulated study was closely linked to two empirical studies performed in The Netherlands.

Our simulated dataset consisted of hypothetical patients with metastatic colorectal cancer. The patients had two different treatment options: Sequential (treatment S) or Combination chemotherapy (treatment C). For each patient, the complete dataset contained information about baseline characteristics, prognostic factors, and two survival and two cost outcomes for each patient: one outcome when treated with treatment S, and one outcome when treated with treatment C. The cost-effectiveness of treatment S versus treatment C was the subject of investigation in this cohort.

The design of the study can be divided into the following sections: (1) Description of the empirical studies, (2) creation of an artificial or 'complete' database, (3) creation of different biased datsets from this dataset, (4) application of different methods to these datasets, (5) comparison of outcomes of the different bias-correcting methods according to diverse criteria, and (6) Formulation of a decision aid.

2.1 Description of the empirical studies

The simulated dataset was based on data from two empirical studies: (1) the CAIRO trial and (2) a real-life study. Both studies included patients diagnosed with stage IV colorectal cancer in 2003 and 2004. Between January 2003 and December 2004, in total 1957 patients were registered with stage IV colorectal cancer by the Netherlands Cancer Registry, which records all cancer patients in The Nethelands at primary diagnosis (Figure 1).

Available treatments in 2003 and 2004

For patients with stage IV colorectal cancer there were no curative treatment options and palliative systemic treatment with chemotherapy was the treatment of choice. In 2003 and 2004 three different chemotherapy agents were available: fluoropyrimidines (FL), irinotecan and oxaliplatin. Although all treatments were considered effective, no good data on the optimal strategy to use these drugs were available in 2003. For this reason additional empirical research was performed. Table 1 shows the possible treatment options at that time in the Netherlands.

Table 1 Possible (equivalent) treatment combinations

first-line	second-line	third-line
FL	FL+ oxaliplatin	(FL+) irinotecan
FL	(Fl +) irinotecan	FL+ oxaliplatin
FL+ oxaliplatin	(Fl +) irinotecan	
(FL +) irinotecan	FL+ oxaliplatin	

The first two treatment combinations can be referred to as “sequential treatment” since both treatments start with fluoropyrimidines only, followed by either oxaliplatin or irinotecan in second- and third-line. The latter two are “combination treatments” as they directly start with a combination of fluoropyrimidines and oxaliplatin or irinotecan in the first-line treatment.

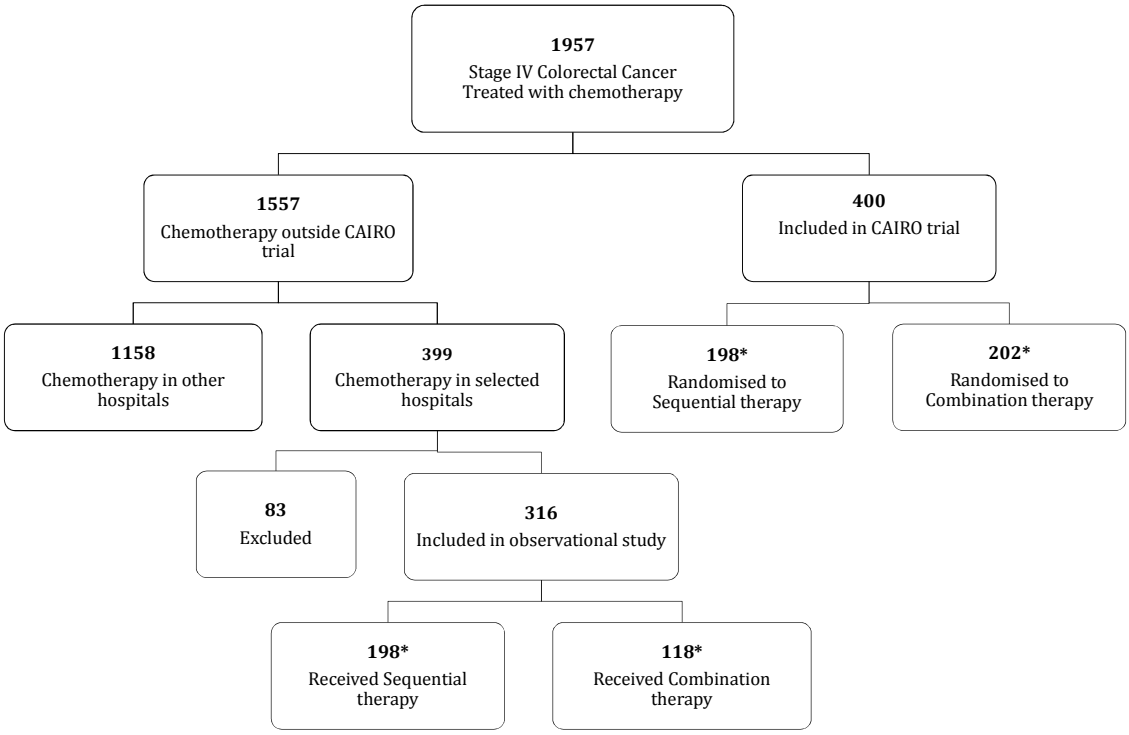
CAIRO trial

Between January 2003 and December 2004, 820 advanced colorectal cancer patients were included in the phase III randomized CAIRO trial (ClinicalTrials.gov NCT00312000) of the Dutch Colorectal Cancer Group (DCCG).(8,9) The subgroup of 400 patients diagnosed with stage IV colorectal cancer, who were identified by the Netherlands Cancer Registry, was used in the simulation study. Patients were randomized between first-line capecitabine (FL), second-line irinotecan, and third-line FL + oxaliplatin (sequential treatment arm) and first line FL + irinotecan and second-line FL + oxaliplatin (combination treatment arm). These cytotoxic drugs were administered at their recommended doses and schedules. The trial used entry criteria that also apply to the use of these drugs in general practice. A total of 79 of the approximately 100 Dutch hospitals participated to this study. Its primary outcome measure was survival. A summary of prognostic baseline characteristics and survival outcomes is provided in table 1.

Real-life study

The real-life study was conducted by iMTA as case study to obtain experience regarding the Dutch policy regulations 'expensice medicines'. In this study, patients receiving chemotherapy outside the CAIRO trial were identified in 29 selected hospitals (3 university hospitals, 14 large teaching hospitals, and 12 general hospitals, together to be considered a good representation of clinical healthcare in The Netherlands). Twenty-five of these hospitals participated to the CAIRO trial. The medical files of all stage IV colorectal cancer patients who received chemotherapy outside the CAIRO trial in these 29 hospitals were reviewed, of whom in total 316 patients were included in the real-life study. Data were collected on baseline characteristics, treatment schedule, resource use, and survival.(10) Although treatment assignment was not fixed or randomized, patients received either sequential chemotherapy treatment or combination treatment, like in the CAIRO trial. A summary of prognostic baseline characteristics, survival outcomes and costs is provided in table 2.

Figure 1 Flowchart



** Simulation (n = 20.000) based on these patients (n = 716)*

Table 2 Characteristics of stage IV CRC patients

Characteristics	Trial patients		Non-Trial patients	
	Sequential	Combination	Sequential	Combination
	(n = 198)	*(n = 202)*	*(n = 198)*	*(n = 118)*
Sex				
Male	127 (64%)	133 (66%)	115 (58%)	72 (61%)
Female	71 (36%)	69 (34%)	83 (42%)	46 (39%)
Age, years				
Median	61	61	64	59
Range	27 - 82	35 - 80	30 - 92	29 - 81
Age < 70	160 (81%)	167 (83%)	134 (68%)	103 (87%)
Age ≥ 70	38 (19%)	35 (17%)	64 (32%)	15 (13%)
Lactate dehydrogenase (LDH)				
Normal	105 (53%)	108 (53%)	77 (48%)	53 (54%)
Abnormal	93 (47%)	94 (47%)	82 (52%)	46 (46%)
Missing			39	19
Alkaline phosphatase (AF)				
Normal	83 (43%)	87 (44%)	58 (38%)	35 (36%)
Abnormal	111 (57%)	113 (56%)	94 (62%)	61 (64%)
missing	4	2	46	22
White blood count (WBC)				
Normal	134 (69%)	145 (73%)	111 (62%)	75 (71%)
Abnormal	30 (31%)	55 (27%)	67 (38%)	30 (29%)
Missing	4	2	20	13
Performance status				
0	114 (58%)	124 (61%)	50 (42%)	34 (48%)
≥ 1	84 (42%)	78 (39%)	70 (58%)	37 (52%)
Missing			78	47
Resection of primary tumour				
Yes	121 (37%)	133 (66%)	117 (61%)	72 (64%)
No	121 (63%)	68 (34%)	76 (39%)	41 (36%)
Missing	5	1	5	5
Predominant localisation of metastases				
Liver	172 (88%)	178 (88%)	166 (91%)	96 (92%)
Extrahepatic	24 (12%)	24 (12%)	17 (9%)	8 (8%)
Missing	2		15	14
Site of primary tumour				
Colon	136 (69%)	130 (64%)	133 (67%)	69 (58%)
Rectosigmoid	19 (10%)	19 (10%)	19 (10%)	15 (13%)
Rectum	42 (21%)	53 (26 %)	46 (23%)	34 (29%)
Missing	1			
Metastatic sites involved				
1	79 (41%)	90 (45%)	92 (47%)	57 (49%)
≥ 2	114 (59%)	110 (55%)	102 53%)	59 (51%)
Missing	5	2	4	2
Overall survival (months)				
Median	13,25	15,79	9,18	15,38
Range	0.39 - 69.28	0.59 - 80.03	0.07 - 60.3	0.16 - 66.33
Mean	16,97	19,51	13,31	18,54
SD	14,05	14,86	12,59	14,01
1-year survival rate	54%	68%	40%	63%
Event rate	98%	98%	91%	87%
Censored rate	2%	2%	9%	13%
Total costs (Euro) *n = 130*				
Median	*NA*	*NA*	19.237	30.008
Range	*NA*	*NA*	462 - 65,288	2200 - 109,139
Mean	*NA*	*NA*	16.208	24.458
SD	*NA*	*NA*	14.627	21.038

Weibull survival model based on empirical data

For the purpose of the simulation study, the data from the two studies were merged. Subsequently, a multivariate weibull survival analysis was conducted to analyse the effect of combination (treatment C) versus sequential therapy (treatment S) in empirical stage IV colorectal cancer patients.

Table 3 shows the results of the analysis, which included all covariates with a significant effect on survival in the Weibull survival model.

Table 3 Multivariate survival analysis

		95% CI		
Variables	Hazard Ratio	Lower Limit	Upper Limit	p value
Combination treatment	0.73	0.63	0.86	< 0.0001
Age older than 70	1.20	1.01	1.45	0.0588
Performance score ≥ 1	1.43	1.04	1.63	0.0214
Abnormal AF levels	1.56	1.32	1.84	< 0.0001
WBC elevated with 100 points	1.26	1.12	1.43	0.0003
No resection primary tumour	1.75	1.42	2.15	< 0.0001
When number of involved metastatic sites = 2+				
No resection primary tumour	1.57	1.27	1.94	< 0.0001
Abnormal LDH levels	1.64	1.35	1.99	< 0.0001

GLM cost model based on empirical data

Based on cost data obtained from the real-life study, a regression analysis was performed to analyse the effect of combination (treatment C) versus sequential therapy (treatment S) on total treatment costs. We used a Generalized linear model with power link (power=1.29) and poisson variance function. Variables included in this analysis were: (survival) days, days squared and interactions of these two variables with treatment, age, and sex.

2.2 Creation and description of 'complete' dataset

Covariates

In the first phase, the covariates for 20,048 patients were synthesized in a way that preserved the covariance structure of the original data. These covariates were the covariates with a significant effect on survival in the Weibull survival model – Age, performance score, AF levels, WBC levels, resection, number of invoved metastatic sites, and LDH levels - plus some additional ones – gender, and predominant location of metastases.

First, a gender was assigned to each synthesized patient. This was done by drawing from a binomial distribution with a probability of being female equal to the proportion of women in the original data (35%). Next, age was related to gender in the original data set in a simple ordinary least squares model:

$Age\_o_i = \beta_0 + \beta_1 * Sex\_o_i + \varepsilon_i$, with $\varepsilon_i \sim N(0,\sigma)$ (1)

In this equation, *_o* denotes variables in the original dataset and *i* denotes the individual patients. The coefficients and the root mean squared error from the OLS model were used to add the age variable to the simulation data set:

$Age\_s_i = \beta_0 + \beta_1 * Sex\_s_i + u_i$, with $u_i \sim N(0, root\ MSE)$ (2)

In this equation, *_s* denotes variables in the simulation dataset. Individual deviations u_i from the mean predicted values were drawn from a normal distribution with a mean of zero and a standard deviation which was equal to the root mean squared error from the OLS model.

Next, six covariates in the simulation dataset were synthesized on the basis of logistic models in the original dataset. The first model associated the probability of having an elevated LDH level with age and sex in the original data:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * age\_o_i + \beta_2 * sex\_o_i + \varepsilon_i \text{, with } \varepsilon_i \sim N(0,\sigma). \quad (3)$$

In this equation, p_i denotes the individual probability of having an elevated LDH level. The regression coefficients from this logistic model were combined with the synthesized aged and

sex to predict individual probabilities of having an elevated LDH level in the simulation data set:

$$p_i = \frac{\exp(\beta_0 + \beta_1 * age\_s_i + \beta_2 * sex\_s_i)}{1 + \exp(\beta_0 + \beta_1 * age\_s_i + \beta_2 * sex\_s_i)} \quad (4)$$

The element of chance was not based on the error term in the logistic regression equation. Instead, the predicted probabilities in the simulation dataset were applied in binomial distributions. Drawings from these distributions for each synthesized patients determined the allocation of elevated LDH.

This process of performing logistic regression, predicting individual probabilities and drawing from binomial distributions was repeated for the other covariates, in the following order: elevated alkalic phosphatase (yes/no), predominant location of metastasis (liver or extra hepatic), tumor resection (yes/no), number of metastases (1 or more than 1), and the WHO performance score (0, higher than 0). Each time, one covariate was added to the simulation dataset, based on the existing covariates.

The last covariate to be synthesized was the leukocyte level, for which an OLS model was used. Finally, the logarithmic and squared values for the leukocyte level were calculated and interaction terms were formed for the combination of resection and the number of metastases and for the combination of elevated LDH and the number of metastases.

The resulting covariance structure of the simulated dataset closely resembles the covariance structure of the original dataset.

Survival time

Synthesizing of survival time was based on the Weibull survival model from the previous section 2.1 in which survival was associated with treatment, performance score, elevated LDH level, leukocyte level, elevated alkalic phosphatase, number of metastases and tumor resection.

For each patient, two counterfactuals for survival time were synthesized, one for each treatment. First, individual survival functions were construed for each treatment arm, using the following equation:

$$S(t) = \exp(-\lambda t^p) \quad (5)$$

In this equation, $S(t)$ is the probability of being alive at time t (or the proportion of patients alive at time t), λ is the scale parameter of the Weibull distribution, and p is the shape parameter of the Weibull distribution. The scale parameter is a function of the shape

parameter, the covariates and the regression coefficients. In the accelerated failure-time parameterization of the Weibull model, this function is described by:

$$\lambda = \exp(-p * x\beta) \quad (6)$$

In this equation, $x\beta$ is matrix notation for the linear combination of regression coefficients (from the original data) and synthesized covariates.
In order to determine how long each patient would survive, values for $S(t)$ were drawn for each patient from a uniform distribution ranging from 0 to 1. Survival time t could then be calculated from equation (5). For each individual, one drawing for $S(t)$ was performed, which was used for both treatment arms.

Heterogeneity of treatment effect

Individual variation of the treatment effect was introducing an element of chance in the calculation of λ. Instead of fixed coefficients, random variables were used for each individual. In order to retain the mean and covariance structure of the original data, these random variables were calculated by combining random draws from a normal distribution with the Cholesky decomposition of the variance-covariance matrix from the Weibull model.
The first random coefficient to be determined was the coefficient $b_{treatment\_i}$ for the treatment effect. This was done by adding the estimated coefficient from the original data to the product of the first element of the Cholesky decomposition matrix and a drawing from a normal distribution.

$$b_{treatment\_i} = \beta_{treatment} + 5 * rand_{treatment\_i} * Cholesky\ [1,1] \quad (7a)$$

In this equation, $b_{treatment\_i}$ is the individually determined coefficient for the treatment effect, $\beta_{treatment}$ is the coefficient from the Weibull model on the original data, $rand_{treatment\_i}$ denotes variables a drawing from a normal distribution. This drawing was multiplied by 5 in order to ensure that the synthesized coefficient would be negative for a small proportion of patients. By multiplying the random draw, the synthesizing process incorporated synthesized heterogeneity across individuals as well as uncertainty about the average treatment effect in the population.
The other individual coefficients were subsequently determined by combining random drawings from a normal distribution with the relevant elements of the Cholesky decomposition matrix. The equation for the first

$$b_{performance\_i} = \beta_{performance} + 5 * rand_{treatment\_i} * Chol\ [2,1] + rand_{performance\_i} * Chol\ [2,2] \quad (7b)$$

The survival curves and survival times were determined by applying equations (5) and (6), using the random individual coefficients developed in equations 7.

Survival times under both treatments depended on two elements of change, drawings for *S(t)* and for the treatment effect, which was reflected in the coefficients for other covariates as well. For each individual, one drawing for S(t) was performed, which was used for both treatment arms. Patients who responded relatively well (compared to other patients in this treatment group) to treatment C, also responded relatively well to treatment S.

Treatment costs

Predicted mean costs were assumed to follow the GLM function that was fitted in section 2.1. Using the coefficients from this model, predicted mean costs μ_i were estimated for each synthesized patient for each treatment arm, as well as the standard deviation.

$$\mu_i = X_i\beta^{1/1.29} \quad (8)$$

In this equation, $X_i\beta$ denotes the linear combination of coefficients and covariates in the generalized linear model. Mean predicted costs depend on survival

$$SD_i = sqrt(\varphi)*sqrt(\mu_i) \quad (9)$$

In this equation, φ denotes dispersion factor, which describes the relationship between μ and variance in generalized linear model.

In order to introduce a random element in the individual costs, the predicted mean and standard deviation were used to describe gamma distributions per patient and per treatment arm, from which the individual patient's costs for each treatment were drawn.

$$Treatment\ costs_i \sim Gamma\ (k_i, \theta_i) \quad (10)$$

In this equation:

$$k_i = \mu_i^2 / SD^2 \quad (11)$$

and $$\theta_i = SD^2 / \mu_i \quad (12)$$

2.3 Creation and description of 'biased' datasets

Drawing biased samples

From this 'complete' dataset several nonrandom samples were drawn and formed new datasets in which each patient receives only one treatment, either S or C. To create these datasets, a procedure was developed such that the probability of getting a certain treatment depends on observed prognostic variables. As a consequence, the patients' baseline prognostic features differed between the two treatment groups, i.e. we created 'biased' datasets with confounding variables. The extent of the confounding was varied in four datasets by nonrandom sampling on different prognostic factors. This was done in several steps. First, different sets of prognostic variables, to be used as confounding variables in the four biased datasets, were specified. Then the probability of receiving treatment C was calculated. After this, treatment was assigned to each patient in the 'complete' dataset. Finally, a group of patients was selected for a study sample.

Confounding variables

The 'complete' dataset dataset contained information regarding nine baseline characteristics, based on the relevant baseline characteristics found in the empirical studies. In the empirical real-life study we found that age over 70 and worse performance scores were significantly associated with a decreased chance to receive combination treatment. The reason for this finding was that combination therapy was considered to be slightly more toxic compared to sequential therapy. As a consequence physicians were more reluctant to prescribe combination therapy to these elderly patients and/or patients with a worse health status.

In the biased datasets of the simulation study we decided to depend the treatment choice on all prognostic factors. First the prognostic factors were categorized into "Patient factors" and "tumor factors". We decided that negative patient factors would be associated with a decreased probability to receive treatment C and negative tumour factors would be associated with an increased probability to receive treatment C.

In order for a variable to be a confounder, it must be associated with both treatment assignment and prognosis. Table 5 gives an overview of the impact of each of the nine covariates with respect to treatment assignment and prognosis. Most of the covariates are associated with a worse prognosis regarding survival, but have no impact on total costs. Older age is associated with both improved survival as well as lower costs and sex is only associated with total costs. The variable "locmeta" is associated with neither effects nor

costs, but was included in the "biased" datasets because it can serve as an instrumental variable (see section 2.4)

Based on this categorization, we chose which variables to use to implement confounding in each of the four biased datasets.

Biased dataset Nr. 1

In this dataset, the bias was primarily related to costs; i.e. treatment assignment was associated with two variables with a prognostic impact on total costs.

Biased dataset Nr. 2

In this dataset, the bias was related to all prognostic variables which were associated with a reduced chance of receiving treatment C.

Biased dataset Nr. 3

In this dataset, the bias was related to all prognostic variables which were associated with an increased chance of receiving treatment C.

Biased dataset Nr. 4

All prognostic variables were related to treatment assignment, either in a negative or a positive way.

Table 5 Unfavorable prognostic factors and biased samples

Short name	Explanation	Impact on chance to receive treatment C	Unfavourable prognostic impact on	Biased sample nr. 1	Biased sample nr. 2	Biased sample nr. 3	Biased sample nr. 4
Patient factors							
AGE	Age over 70 years old	Negative	Effects & Costs	+	+		+
SEX	Female sex	Negative	Costs	+	+		+
PS	Performance score over 0	Negative	Effects		+		+
AF	Abnormal AF	Negative	Effects		+		+
WBC	Higher WBC	Negative	Effects				
Tumour factors							
NRM	Nr of metastasis over 1	Positive	Effects			+	+
URES	Unresected tumour	Positive	Effects			+	+
LDH	Abnormal LDH	Positive	Effects			+	+
LOCMETA	Dominant metastasis outsite liver	Negative	No impact	+	+	+	+

Treatment assignment

For each of the four specifications of confounding variables, the probability to receive treatment C was predicted by a complementary log-log model.

$$Pr(T = 1| X) = 1 - exp(-exp(XB)) \quad (13)$$

The complementary log–log link was chosen because it is not usually applied in the estimation of propensity scores, which are used in several methods to adjust for confounding. This should prevent an artificially good fit of the propensity score models.

Using the probability for each patient, treatment – with the corresponding survival time and costs - was assigned by drawings from individual binomial distributions:

$$Treatment_i \sim Binomial\ (Pr_i) \quad (14)$$

Inclusion in sample

The dataset was sorted on the assigned treatment and a random number, ensuring that patients would be randomly ordered within their treatment group. The first 400 or 2000 patients per treatment group were then selected for inclusion into the four biased study samples.

2.4 Application of bias correction techniques

Adjustment for confounding can be approached from two angles. Defined simply, confounding is the combination of two associations – the association of a variable with the outcome (making it a risk factor) and the association of this variable with treatment assignment.(3). The problem can be solved by addressing either of these associations. Regression focuses on modeling the effect of the risk factor on the outcome. Other methods eliminate the association of the confounder with treatment.

The following techniques were used to achieve unbiased estimates of the incremental effect of treatment C versus treatment S on mean survival and mean costs.

0. ***Gold standard.*** The 'real' mean survival and costs were calculated per treatment. This was done by taking the average potential outcomes for each treatment for the treated (defined as the patients who were treated with treatment C). These results were later used to compare the results of the other models.

1. ***Mean method.*** Under the first naïve method, the mean assigned survival and costs for each treatment group were calculated without any adjustment.

2. *Simple regression*. A Weibull survival model was fit on the survival data with treatment group as the only covariate. For costs, a generalized linear model with a log link and treatment group as the only covariate was estimated. The regression results were used to predict potential survival and costs for all treated patients, under both treatment regimes. Next. the mean survival and costs per treatment group were calculated as the mean for each treatment regime. Finally, the incremental cost-effectiveness of treatment C versus treatment S was expressed as net monetary benefit assuming that one additional life-year lived is worth €40,000 and €80,000.(11)

3. *Full regression.* This method was the same as the previous one, but – assuming that a researcher would use all available data - it involved all the regression covariates that were applied in synthesizing the patients, survival and costs. However, unlike the synthesizing process, a log link instead of a power link was used in the generalized linear model, while the three-way interactions were omitted.

4. *Propensity score matching (effects model), mean method*. A probit model, containing all variables that were used for synthesizing individual survival, was used to predict each patient's probability of receiving treatment C, given his or her covariates. Using these probabilities (propensity scores), each treated patient was then matched to an untreated patient with the most similar propensity score.(12-14) Treated patients could not be matched more than once, while untreated patients could be matched to more than one treated patient, giving them a higher weight in later calculations. The only restriction was that there had to be 'common support', meaning that patients were matched only if their propensity score was not higher (lower) than the highest (lowest) score among patients in the other treatment group. After matching, the mean survival, mean costs and net monetary benefit were calculated for the matched patients from each group.

Simulation studies have shown that propensity score techniques perform best when the model contains all confounders and possibly other risk factors. Variables which are associated with treatment but have no effect on the outcome should be omitted since they tend to introduce bias.(15) However, costs in our datasets were synthesized using a model that contains sex, which is not a risk factor in the survival model, and age, which is in the survival model only as a categorical variable (≥ 70 years). This suggests that a single propensity score model cannot be ideal for both costs and effects. On the other hand, it is unwise to use different models in a cost-effectiveness analysis, because this would mean that the (matched) sample from which the costs were taken would be different from the sample for effects.

5. *Propensity score matching (effects model), regression*. Patients were matched in the same manner as in the previous method. However, this time the full Weibull regression and generalized linear model were used to predict individual survival time and costs per matched patient. The mean survival and costs per treatment were then calculated as the mean predicted survival and costs for treated patients, for each treatment regime. Finally, the net monetary benefit was calculated.

6. *Propensity score matching (costs model), mean method.* Patients were matched in the same manner as in the previous two methods. This time, a propensity score probit model for costs was applied. Since costs were synthesized based on simulated survival, the variables from the survival model were confounders for costs as well and were therefore included in the cost model. The variables age and sex were added to the probit model.
It would not have been appropriate to use survival time in the propensity score model, although it was applied in the model with which the costs were synthesized. The scores from such a model could not be used for estimating survival time. Survival time would then become an explanatory variable as well as an outcome variable.

7. *Propensity score matching (costs model), regression.* This method was the same as the previously discussed method for regression after matching on propensity scores, but this time the scores from the costs model were used.

8. *Propensity score as covariate (effects model), simple regression* Under this method, all covariates from the full regression models (except for the treatment variable) were replaced by the propensity score as the sole covariate. The regression results were used to calculate the mean survival and costs per treatment regime and net monetary benefit as described for the other regression models.

9. *Propensity score as covariate (effects model), full regression.* In this method, the propensity score was added to variables in the full regression models, which could be seen as making the functional form of these variables in the regression more flexible.

10. *Propensity score as covariate (costs model), simple regression.*
11. *Propensity score as covariate (costs model), full regression.* In these methods, the propensity score from the costs model was used in regression analyses instead of the score from the survival model.

12. ***Inverse probability weighting (effects model), mean method.*** Probit regression was used to predict the probability that a patient would receive the treatment that they actually received. The probit model was based on the model that was used to synthesize survival time. For treated patients, this probability is equal to the propensity score. For untreated patients, the probability is equal to one minus the propensity score.(16,17) The predictions were used to calculate inverse probability weights by taking the inverse of the probabilities. This meant that patients with characteristics that were relatively underrepresented in a treatment group received a higher weight. These weights were applied in the calculation of sample means for survival and costs per treatment group, from which the net monetary benefit could be calculated.

13. ***Inverse probability weighting (effects model), regression method.*** The same weights were used in the Weibull regression model for survival time and the generalized linear model for costs. The regression results were used to predict survival times and costs, after which the means per treatment regime and the net monetary benefit were calculated.

14. ***Inverse probability weighting (costs), mean method***

15. ***Inverse probability weighting (costs model), regression method.*** Assuming that the rules for variable selection for propensity scores might also apply to inverse probability weighting, we also based these weights on the simulation model for costs. These weights were applied in the calculation of sample means for survival and costs per treatment group, from which the net monetary benefit could be calculated.

16. Instrumental variable regression. This method used two-stage regression models to estimate the treatment effects on survival and costs. In the first stage, the assigned treatment (C or S) was regressed on the covariates from the survival simulation model, plus an additional variable that was associated to treatment assignment but to survival or costs. The results from this ordinary least squared analysis were used to linearly predict the probability of being treated by C.(18)

In the second stage, this predicted probability was used in the full regression models instead of the actual treatment. The regression results were used to calculate the mean survival and costs per treatment regime and the net monetary benefit as described for the other regression models.

2.5 Comparisons of correction methods

The correction methods were compared on validity and reliability in samples of size 400 and 2000.

Validity

Acknowledging that the results could be affected by the stochastic process of drawing samples, we performed 1000 iterations and assessed the validity on each of them. These assessments focused on how well each method produces valid estimates of incremental effectiveness, incremental costs, and incremental cost-effectiveness (net monetary benefit). Degree of validity was based on comparisons of the results using a particular method with the gold standard, i.e. the results of the calculations based on the potential outcomes for the patients who were selected for the sample.

For each iteration, the `real' incremental effectiveness (survival gain) was subtracted from the estimate gain. The bias was then calculated as the mean of these differences over all iterations. Bias was also expressed as a percentage of the real incremental effectiveness. The same process was applied for incremental costs and net monetary benefit.

Brookhart et al. used the mean squared error (MSE) as a measure to evaluate a number of propensity score covariate regression models.(15) This was calculated by squaring the errors at the individual level (defined as the difference between the observed – or 'real' - and predicted outcomes) and taking the means of these squared errors. However, this measure cannot be calculated using the means method, because this method does not have predicted outcomes at the individual level. We therefore chose to use the mean deviation instead of MSE. Trying to make this somewhat equivalent to the MSE, or root MSE, we calculated mean deviation by squaring the bias at the sample level (which is the difference between mean `real' outcomes and the mean outcomes for the particular method) and then taking the square root of the result. This approach yields a value that could be interpreted as the average error of the estimated means.

For each iterations, we determined which of the methods produced the best results for costs, effects and net benefit. We also ranked the methods according to the proportion of iterations in which they had delivered the best results.

Reliability

Reliability was assessed by comparing the size of the 95% confidence intervals of estimates of the incremental effectiveness and the incremental costs. These intervals were derived by applying bootstrapping the bootstrap method with 1000 iterations on one randomly drawn

sample from each biased dataset.(19) Additionally, the bootstrapping results were used to construct cost-effectiveness acceptability curves for selected methods for each biased dataset.(20)

The degree of reliability of cost-effectiveness estimates was determined using bootstrapping to obtain values for the 95% confidence interval.

Summary measures

As a summary measure designed to facilitate comparisons of the methods, all methods were ranked according to each criterion. The second summary measure was based on a judgment of acceptability of each method according to each criterion (i.e., does the method yield acceptable results, yes or no?). For effects, a bias ≤ 10% (in absolute terms) of the correct value was considered acceptable. For costs, the limit was set to 15%. For the net monetary benefit with a threshold value of €40,000, the maximum acceptable bias was 30% and for the net monetary benefit with a threshold of €80,000, this was 10%. For reliability, confidence intervals were considered acceptable if they did not span more than 100 days or €5000.

3 Results

The results of all methods on all biased datasets are summarized in four tables (Tables 5-8). In table 5 and 6, the methods are ranked on performance, per dataset, for costs and effects and for validity and reliability (for n = 2000 and n = 400, respectively). In tables 7 and 8, each method is judged based on whether or not its results were considered acceptable. These tables are summary tables based on the simulation results that are presented in the appendix. In addition to summary tables we also present our results using acceptability curves (Figures 2 and 3). The results are first presented for n = 2000. Subsequently, we will highlight the differences between the results using populations of 2000 and the results using populations of 400 patients.

Biased dataset 1: bias related to costs

Validity, effects

With regards to the estimated survival gain of treatment C compared to treatment S, all methods performed reasonably well or very well. The lowest biases were achieved by the naïve methods, full regression, propensity-score covariate regression and IPW methods (range: 0.2% to 2.2%). The size of bias for PSM methods was somewhat larger, with 9% for regression after matching on propensity score based on the costs model. A bias of the same magnitude, but in the other direction, was brought about by instrumental variable regression. On the other hand, use of PSM (mean, costs) led to a very small bias: 0.2%.

Validity, costs

Estimation of incremental costs was more problematic than that of incremental effects, although the results of the most non-naïve methods were considered acceptable. All results were overestimations of the real costs. A very good result was achieved by PSM (mean, costs). The largest biases (31.7%) were found in the results of the naive methods, without adjustment for bias. With the exception of IPW (mean, costs), PSM methods performed much better than the other methods, with biases from 0.1% to 6.9%.

Validity, NMB

Due to the fact that the 'real' incremental cost-effectiveness ratio of treatment C versus S was somewhat higher than €30,000, relative biases could be expected to be higher for a net monetary benefit based on a threshold of €40,000 per life-year gained than for the net monetary benefit with a threshold of €40,000. The former net monetary benefit is relatively close to zero. In absolute terms, biases are rather similar for both thresholds. Due to the

general overestimation of incremental costs, net monetary benefits were underestimated by most methods.

When a threshold for willingness-to-pay was set to €40,000, the best results were achieved by IPW (mean, costs), PSM (mean, costs), PSM (regression, effects and costs) and PS covariate regression (simple, costs). The results of the other methods were considered unacceptable.

When the threshold was set to NMN €80,000, the results were somewhat different. PSM (mean, costs) and IPW (mean, costs) performed quite well again. The results for PSM (regression, costs) contained too much bias. On the other hand, PS covariate regression (full, effects and full, costs) as well as IPW (regression, effects) and IPW (regression, costs) were now considered sufficiently good.

Reliability, effects

The estimates with the smallest confidence intervals were produced using full regression (83.6) and PS covariate regression (87.6). PSM and instrumental regression scored the worst. Their results were considered insufficient, as were the results from PS covariate regression (simple, effects and simple, costs) and IPW (mean, costs and regression, costs).

Reliability, costs

These results were rather similar to those for reliability of effect estimates. The results for all PSM methods, instrumental variable regression, and for all IPW methods except for IPW (mean, effects) were considered insufficient.

Reliability, cost-effectiveness

Three methods would have been most helpful in reaching the correct decision that the treatment C was likely to be cost-effective at certain thresholds. They resulted in steep cost-effectiveness acceptability curves around the real ICER of €32,271. These methods were PS covariate (simple), IPW regression and PSM (mean). The results of these methods as shown in the figures are based on propensity scores for effects. However, the results using scores based on costs are similar.

Summary

With regards to validity, the most consistently good results were achieved by PSM (mean, costs), IPW (mean, costs) and PS covariate regression (simple, costs). With regards to reliability, the results of these three methods were considered insufficient for costs (except for ps covariate regression (simple, costs) and effects. In terms of reliability for costs and

effects, the other unweighted regression methods and naive methods performed better. For reliability of the cost-effectiveness estimates, the valid methods performed best.

Biased dataset 2: bias related to patient factors causally associated to costs and effects

Validity, effects

Except for the naive methods and instrumental variable, all methods performed reasonably well to well, with rather similarly sized biases (0.8% to 6.5%). Most results were underestimates of the actual survival gain. The lowest biases were achieved by full regression, propensity-score covariate regression and regression after inverse probability weighting.

Validity, costs

Very good results were achieved by PSM (mean, costs) and IPW (mean, costs): biases of 2.8% and 4.0%, respectively. The bias of all PSM methods was within acceptable bounds. The results of PS covariate regression (simple and full effects) and IPW (regression, effects and costs) were considered too strongly biased.

Validity, NMB

The degree of bias in estimating both NMB measures was acceptable for all PSM methods and also instrumental variable regression. For NMB €40,000, the performance of IPW (mean, costs) was acceptable, and for NMB €80,000 the bias of PS covariate regression (full, costs) was sufficiently small. All other methods led to estimates with too much bias.

Reliability, effects

All methods except for PSM methods and instrumental variable regression resulted in sufficiently narrow confidence intervals.

Reliability, costs

Acceptably narrow confidence intervals were produced by the naive methods as well as PS covariate regression (simple, costs) and IPW (mean, costs and effects). Intervals for PSM were consistently €1000 to €2000 broader than those for similar approaches based on PS covariate regression and IPW with similar options.

Reliability, cost-effectiveness

PSM regression, PSM mean, IPW mean, IPW regression, full regression and PS covariate regression led to steep acceptability curves that were fairly close to the vertical line at the value of the ICER. Instrumental variable regression and PS covariate regression (simple) produced less useful curves.

Summary

With regards to validity of effect estimates, the methods do not differ much in bias, except for the naive methods and instrumental regression. The most consistently valid results were achieved by PSM (mean, costs) and IPW (mean, costs). However, all PSM methods performed acceptably with regards to validity of effects, costs and NMB.

Whereas PSM (mean costs) did not produce sufficiently narrow confidence intervals for either costs or effects, the results of IPW (mean, costs) were reliable. Regarding the reliability of the cost-effectiveness estimates, the valid methods performed well, as did most others.

Biased dataset 3: bias related to tumour factors causally associated with effects

Validity, effects

The success in removing bias varied substantially between adjustment methods. When the naive methods were used, it appeared that treatment C had no survival gains compared to treatment S. These results had a bias of around 100%. PSM (mean, effects and costs) and full regression led to the best results (range in bias: 0.8-1.7%). All PS covariate methods performed well, as did regression after IPW (not more than 4%). It is notable that for PSM, mean methods achieved better results than regression methods, whereas the reverse was true for IPW.

Validity, costs

Except for the naive methods and instrumental variable regression, all methods led to good estimates of the incremental costs, with less than 7% bias. The best results were for PSM (mean, effects and costs) and IPW (regression, effects).

Validity, NMB

PSM (mean, effects and costs) had the most unbiased estimates of both NMBs. PS covariate regression (full, effects and costs), IPW (regression, effects and costs) and full regression performed well. The other methods led to unacceptably biased results.

Reliability, effects

All methods resulted in sufficiently narrow confidence intervals, except for PSM methods and instrumental variable regression.

Reliability, costs

For cost as well, all methods except for PSM methods and instrumental variable regression resulted in sufficiently narrow confidence intervals. Intervals for PSM were consistently €1500 wider.

Reliability, cost-effectiveness

All methods resulted in steep acceptability curves that were fairly close to the vertical line at the value of the ICER. The curve produced using the instrumental variable regression was furthest from the vertical line, but even this result was acceptable.

Summary

PSM (mean, effects and costs) and full regression led to the best estimates of incremental effects, but all PS covariate and IPW regression methods performed well. For incremental costs, in all methods except for the naive methods and instrumental variable regression, differences for reliability were small. Overall, the methods that estimated effects well, were most appropriate for this dataset.

Biased dataset 4: bias related to all patient and tumour factors causally associated with effects and costs

Validity, effects

The success in removing bias varied between adjustment methods. The least bias was achieved by PS covariate regression (simple, effects and costs), IPW (mean, effects and costs) and full regression: 0 to 1.5%. Other PS covariate regression and IPW methods, and instrumental variable regression were successful as well. The mean methods after PSM performed acceptably, but led to more bias. Both regression methods after PSM led to bias of the same magnitude as the naïve estimation methods.

Validity, costs

Only six of the 15 methods managed to produce acceptable estimates of the cost difference. The least biased result was achieved by PSM (mean, effects): 2.5%. The other PSM methods, IPW (mean, costs) and PS covariate regression (simple, costs) had sufficiently unbiased outcomes, as well. PS covariate regression (simple, effects) had almost the same bias as the naive methods (24.9%).

Validity, NMB

For NMB €40,000, limited bias was achieved by PSM (mean, costs), PSM (regression, effects and costs), and IPW (mean, costs). For NMB €40,000, IPW (mean, costs) and IPW (regression, effects and costs) had the best result. Full regression, PS covariate methods except for PS covariate (simple, effects) and PSM (mean, costs) also yielded acceptable results.

Reliability, effects

Only the intervals of the naive methods, full regression and two PS covariate regression methods (full, effects and costs) had confidence intervals that were narrower than the limit of 100 days. Instrumental variable regression performed worst, with an interval of 230 days.

Reliability, costs

Except for the naive methods, only full regression and PS covariates methods produced intervals within the limit of €5000. The widest intervals were seen with instrumental variable regression (€12,000) and PSM (mean, costs)(€10,094). Overall, PSM performed relatively poorly.

Reliability, cost-effectiveness

Most methods resulted in steep acceptability curves that were fairly close to the vertical line at the value of the ICER: PSM (mean), IPW (regression), IPW (mean), full regression, PSM covariate regression. The curve from the instrumental variable regression was furthest from the vertical line.

Summary

Five methods that were very successful in effects – PS covariate regression (simple and full, effects and costs), three IPW methods and full regression - led to strongly biased estimates of costs. However, the IPW (mean, costs) method performed very well in effects, and did the same in costs. PSM methods were acceptable or nearly acceptable in effects and good in

costs. They also achieved good results for NMB €40,000, and some of them for NMB €80,000, which was also estimated correctly by PS covariate methods.
With regards to reliability, only PS covariate regression (full, effects and costs) for effects and all PS covariate methods for costs were considered acceptable.

Smaller sample

When the sample size was reduced from 2000 to 400, the validity results were quite similar and there were few differences.
However, the confidence intervals for the 400-patient samples were much wider than the those for the 2000-patient samples. In fact, the intervals in the smaller sample were generally 2 to 2.5 times wider than the confidence intervals in the larger sample. None of the confidence intervals was considered sufficiently narrow.
The rankings of the interval sizes over the methods were not very sensitive to the changes in sample size. Furthermore, in acceptability curves, the same methods produced the most certainty to the decision maker. Curves for the 400 patients sample were less steep and further apart.

Table 5 Ranking of methods based on a population of 2000 patients

n = 2000	**Validity**																**Reliability**							
	Effects				**Costs**				**NMB_40000**				**NMB_80000**				**Effects**				**Costs**			
	Nr.1	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**
Mean	8	15	16	15	15	16	15	14	16	8	16	16	15	15	16	16	4	9	9	4	1	1	8	1
Simple regression	6	16	15	14	16	15	16	15	15	9	15	15	14	16	15	15	5	10	8	5	2	2	9	2
Full regression	3	1	3	5	11	13	5	11	12	15	6	12	10	12	1	8	1	1	1	1	8	7	4	5
PSM, mean (effects model)	13	7	1	12	5	3	2	1	10	5	2	8	12	2	5	11	14	14	15	14	11	12	13	14
PSM, regression (effects model)	14	10	11	13	3	4	12	4	3	3	8	3	5	1	11	10	13	12	13	12	13	15	14	11
PSM, mean (costs model)	2	9	2	10	1	1	3	2	2	1	3	1	1	4	6	6	15	15	14	15	15	13	12	15
PSM, regression (costs model)	15	12	12	16	4	6	13	5	4	2	9	4	11	3	12	13	12	13	12	13	14	14	15	13
PS covariate, simple regression (effects model)	7	8	6	4	14	10	10	13	13	16	10	13	13	13	9	12	8	8	7	7	3	3	2	3
PS covariate, full regression (effects model)	4	2	4	6	9	9	8	8	8	11	5	7	8	7	2	4	2	6	2	2	6	10	5	6
PS covariate, simple regression (costs model)	1	6	7	1	6	5	11	6	5	7	11	10	6	9	10	7	9	11	6	8	4	4	1	4
PS covariate, full regression (costs model)	9	3	5	7	7	8	9	7	6	10	4	5	3	6	3	2	3	7	3	3	7	11	6	7
IPW, mean (effects model)	5	13	13	3	13	7	7	9	11	14	12	11	9	14	13	9	7	3	11	10	5	5	11	10
IPW, regression (effects model)	11	4	9	9	8	12	1	10	7	12	7	6	4	8	8	3	6	4	4	6	10	9	7	9
IPW, mean (costs model)	12	11	14	2	2	2	4	3	1	6	13	2	2	11	14	1	11	5	10	11	12	6	10	12
IPW, regression (costs model)	10	5	8	8	12	14	6	12	9	13	1	9	7	10	7	5	10	2	5	9	9	8	3	8
Instrumental variable	16	14	10	11	10	11	14	16	14	4	14	14	16	5	4	14	16	16	16	16	16	16	16	16

Table 6 Ranking of methods based on a population of 400 patients

n=400	**Validity**																**Reliability**							
	Effects				**Costs**				**NMB_40000**				**NMB_80000**				**Effects**				**Costs**			
	Nr.1	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**
Mean	6	15	16	14	15	15	14	14	16	10	16	15	13	15	16	14	7	7	10	4	1	1	8	1
Simple regression	10	16	15	13	16	16	15	15	15	11	15	14	12	16	15	12	8	8	11	5	2	2	9	2
Full regression	1	3	3	3	11	14	8	11	11	15	3	10	9	8	1	9	1	3	1	1	9	7	1	4
PSM, mean (effects model)	13	9	1	11	4	3	1	1	7	3	6	9	10	2	5	10	15	14	15	11	12	11	14	10
PSM, regression (effects model)	15	13	13	15	5	7	12	4	8	4	14	11	15	13	13	15	12	12	13	10	14	14	12	11
PSM, mean (costs model)	5	8	2	7	1	1	2	2	1	1	5	1	1	1	6	4	14	15	14	14	13	12	15	13
PSM, regression (costs model)	16	14	14	16	6	8	13	6	12	5	13	13	16	14	14	16	13	13	12	12	15	15	13	12
PS covariate, simple regression (effects model)	2	7	4	2	14	11	3	13	13	16	8	12	11	11	9	11	11	10	6	8	4	3	4	3
PS covariate, full regression (effects model)	9	4	6	9	8	10	10	9	5	9	1	4	5	6	2	3	4	5	2	2	11	9	6	7
PS covariate, simple regression (costs model)	4	6	5	1	3	5	4	5	3	7	9	7	3	7	10	8	10	11	7	9	3	4	3	5
PS covariate, full regression (costs model)	12	5	8	10	7	9	11	7	4	8	2	3	4	4	3	1	3	6	3	3	8	10	7	8
IPW, mean (effects model)	8	11	11	5	13	6	7	8	10	12	11	6	7	12	11	7	9	4	8	15	5	5	10	15
IPW, regression (effects model)	7	1	9	8	9	12	5	10	6	13	7	5	6	10	8	5	5	1	5	6	7	13	5	9
IPW, mean (costs model)	11	10	10	4	2	2	6	3	2	6	12	2	2	5	12	2	6	9	9	13	6	6	11	14
IPW, regression (costs model)	3	2	7	6	12	13	9	12	9	14	4	8	8	9	7	6	2	2	4	7	10	8	2	6
Instrumental variable	14	12	12	12	10	4	16	16	14	2	10	16	14	3	4	13	16	16	16	16	16	16	16	16

Table 7 Acceptability of methods based on a population of 2000 patients

n = 2000	**Validity**																**Reliability**							
	Effects				**Costs**				**NMB_40000**				**NMB_80000**				**Effects**				**Costs**			
	Nr.1	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**
Mean	YES	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	YES	YES	YES	YES	YES	YES	YES	YES
Simple regression	YES	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	YES	YES	YES	YES	YES	YES	YES	YES
Full regression	YES	YES	YES	YES	-	-	YES	-	-	-	YES	-	-	-	YES	YES	YES	YES	YES	YES	YES	-	YES	YES
PSM, mean (effects model)	YES	YES	YES	YES	YES	YES	YES	YES	-	YES	YES	-	-	YES	YES	-	-	-	-	-	-	-	-	-
PSM, regression (effects model)	YES	YES	-	-	YES	YES	YES	YES	YES	YES	-	YES	YES	YES	-	-	-	-	-	-	-	-	-	-
PSM, mean (costs model)	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	-	-	-	-	-	-	-	-
PSM, regression (costs model)	YES	YES	-	-	YES	YES	YES	YES	YES	YES	-	YES	-	YES	-	-	-	-	-	-	-	-	-	-
PS covariate, simple regression (effects model)	YES	YES	YES	YES	-	-	YES	-	-	-	-	-	-	-	YES	-	-	YES	YES	-	YES	YES	YES	YES
PS covariate, full regression (effects model)	YES	YES	YES	YES	-	-	YES	-	-	-	YES	-	YES	-	YES	YES	YES	YES	YES	YES	YES	-	YES	YES
PS covariate, simple regression (costs model)	YES	YES	YES	YES	YES	YES	YES	YES	YES	-	-	-	YES	-	YES	YES	-	YES	YES	-	YES	YES	YES	YES
PS covariate, full regression (costs model)	YES	YES	YES	YES	YES	-	YES	-	-	-	YES	-	YES	YES	YES	YES	YES	YES	YES	YES	YES	-	YES	YES
IPW, mean (effects model)	YES	YES	-	YES	-	YES	YES	-	-	-	-	-	-	-	-	YES	YES	YES	YES	-	YES	YES	YES	-
IPW, regression (effects model)	YES	YES	YES	YES	YES	-	YES	-	-	-	YES	-	YES	-	YES	YES	YES	YES	YES	-	-	-	YES	-
IPW, mean (costs model)	YES	YES	-	YES	YES	YES	YES	YES	YES	YES	-	YES	YES	-	-	YES	-	YES	YES	-	-	YES	YES	-
IPW, regression (costs model)	YES	YES	YES	YES	-	-	YES	-	-	-	YES	-	YES	-	YES	YES	-	YES	YES	-	-	-	YES	-
Instrumental variable	YES	-	YES	YES	-	-	-	-	-	YES	-	-	-	YES	YES	-	-	-	-	-	-	-	-	-

Table 8 Acceptability of methods based on a population of 400 patients

n = 400	**Validity**																**Reliability**							
	Effects				**Costs**				**NMB_40000**				**NMB_80000**				**Effects**				**Costs**			
	Nr.1	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**	**Nr.1**	**Nr.2**	**Nr.3**	**Nr.4**
Mean	YES	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
									-	-	-	-	-	-	-	-								
Simple regression	YES	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Full regression	YES	YES	YES	YES	-	-	YES	-	-	-	YES	-	-	-	YES	YES	-	-	-	-	-	-	-	-
PSM, mean (effects model)	YES	YES	YES	YES	YES	YES	YES	YES	-	YES	YES	-	-	YES	YES	-	-	-	-	-	-	-	-	-
PSM, regression (effects model)	-	-	-	-	YES	-	YES	YES	-	YES	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PSM, mean (costs model)	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	-	-	-	-	-	-	-	-
PSM, regression (costs model)	-	-	-	-	YES	-	YES	YES	-	YES	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PS covariate, simple regression (effects model)	YES	YES	YES	YES	-		YES	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PS covariate, full regression (effects model)	YES	YES	YES	YES	-	-	YES	-	-	-	YES	-	YES	-	YES	YES	-	-	-	-	-	-	-	-
PS covariate, simple regression (costs model)	YES	YES	YES	YES	YES	YES	YES	YES	YES	-	-	-	YES	-	-	YES	-	-	-	-	-	-	-	-
PS covariate, full regression (costs model)	YES	YES	YES	YES	YES	-	YES	-	-	-	YES	YES	YES	YES	YES	YES	-	-	-	-	-	-	-	-
IPW, mean (effects model)	YES	YES	-	YES	-	YES	YES	-	-	-	-	-	YES	-	-	YES	-	-	-	-	-	-	-	-
IPW, regression (effects model)	YES	YES	YES	YES	-	-	YES	-	-	-	YES	-	YES	-	YES	YES	-	-	-	-	-	-	-	-
IPW, mean (costs model)	YES	YES	-	YES	YES	YES	YES	YES	YES	YES	-	YES	YES	-	-	YES	-	-	-	-	-	-	-	-
IPW, regression (costs model)	YES	YES	YES	YES	-	-	YES	-	-	-	YES	-	-	-	YES	YES	-	-	-	-	-	-	-	-
Instrumental variable	YES	YES	-	-	-	YES	-	-	-	YES	-	-	-	YES	YES	-	-	-	-	-	-	-	-	-

Figure 2 Acceptability Curves

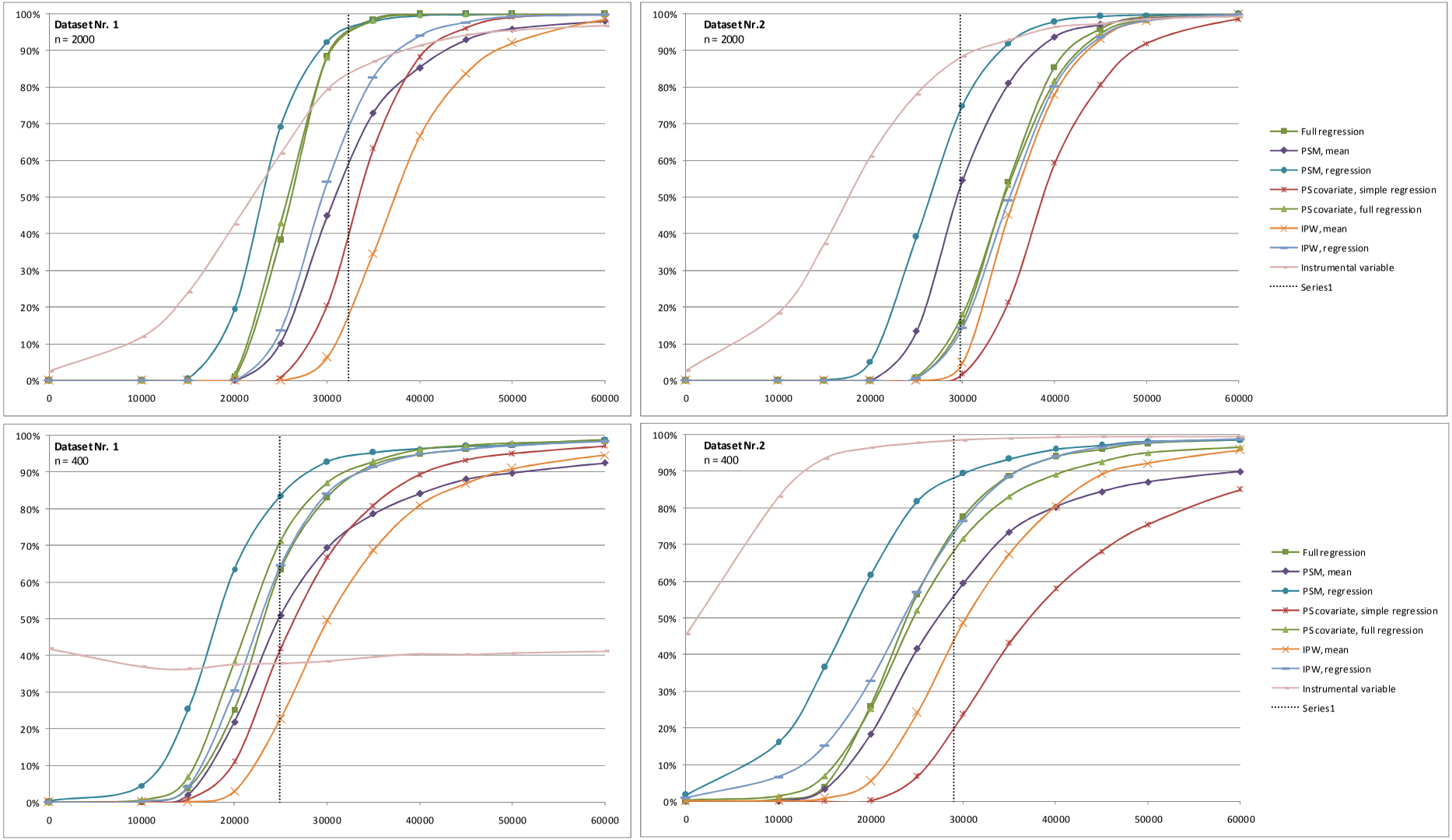
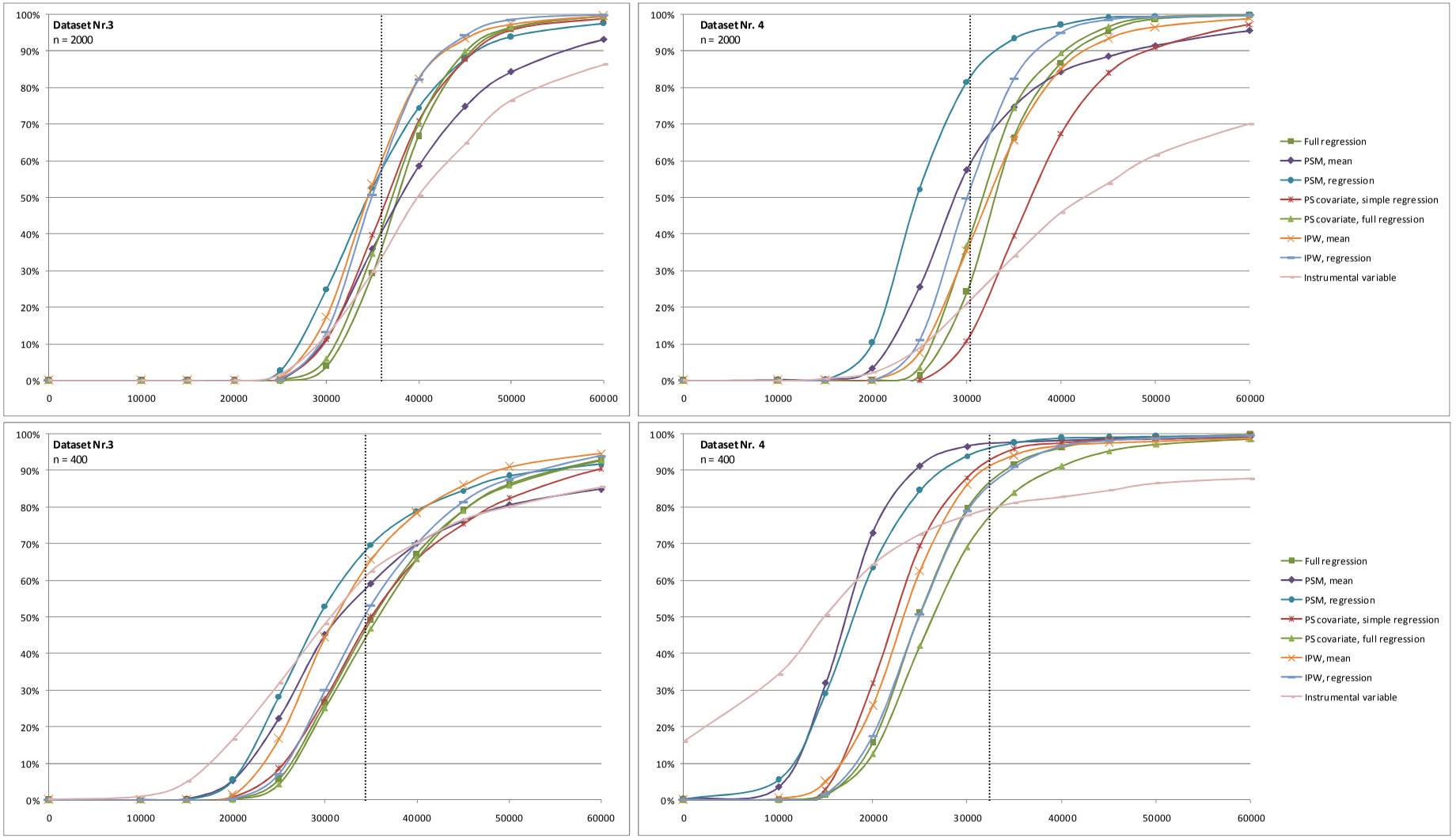


Figure 3 Acceptability Curves



4 Discussion

Evaluation of health care medication in 'real world' daily practice can provide policymakers with results that are much more relevant and applicable to the current situation than economic evaluations piggy-backed onto randomised controlled trials before the new medicine has even been used in daily practice. This requires observational studies, a major shortcoming of which is the absence of a random assignment of treatment. Lack of random assignment can lead to substantial problems with confounding bias.

Several methods have been proposed to address this problem in medical research, mostly focused on the clinical effectiveness of the treatment. However, economic evaluations estimate costs and health effects simultaneously, combine this information in cost-effectiveness ratios (ICERs) of net monetary benefit measures, and determine the likelihood that the intervention is cost-effective given the available information. Cost data have different properties than data on clinical effectiveness. For example, they are typically skewed. This study investigated how well different confounding bias correction techniques performed according to criteria such as validity, precision and practical requirements in the setting of observational cost-effectiveness research, with a focus on the list of expensive medicines. In the next paragraphs we will answer the four research questions we constructed. We simulated a cohort of patients with stage IV colorectal carcinoma. Treatment assignment was non-random, which resulted in four differently biased datasets. This bias was evident in the results of naive analysis methods, which calculate the means of costs and effects in a straightforward manner, or regression analysis with treatment as the sole covariate.

(1) To what extent can different statistical methods provide valid and reliable estimates of incremental treatment costs, effects and cost-effectiveness when using 'real world' observational data in cost-effectiveness studies? and (3) To what extent can optimal and inappropriate techniques be identified in different study settings?

With regards to validity, three methods stood out in their ability to consistently produce relatively unbiased results: 1) taking the mean of costs and effects after inverse probability weighting based on a weighting model for costs, 2) taking the mean of costs and effects after propensity score matching based on a matching model for costs, and 3) regression with two covariates (treatment and the propensity score based on a model for costs). Of these three, IPW and propensity-score-as-covariate regression were most likely to have acceptably narrow confidence intervals in large samples. However, in samples of 400 instead of 2000 patients, reliability was problematic for all methods. More models were successful at estimating effects (incremental effectiveness) than estimating costs (incremental costs). With

respect to practical requirements, all methods were feasible to conduct by researchers with regularly available statistical software.

Several authors have argued that – in order to prevent or reduce bias - models to estimate propensity scores should not include covariates that have no relationship with the outcome. On the other hand, propensity scores should include covariates that are associated with treatment assignment. We have seen that these rules may conflict when two outcomes have to be analysed using the same sample. Some variables were required to be in the propensity score model for cost estimates, whereas they should have been excluded from the model for effect estimates. In our example, sex and – to a lesser extent – age had an impact on cost outcomes, but not on health effects.

Indeed, our results suggest that the choice about which covariates to include in the propensity score can have a substantial impact on the results of the analysis. Both types of models performed equally well when estimating differences in health effects. However, on balance, the cost models achieved better results in cost estimates.

All propensity score matching methods succeeded in eliminating bias in cost estimates, even when they were based on the effects model to develop the propensity scores. However, the reliability of these methods was disappointingly low. Propensity-score-as-covariate regression methods successfully adjusted for bias in effect estimates, but they were less successful with cost estimates, especially when these estimates were based on the effects model for propensity scores. The results of these models were inside our limits of acceptability for reliability. Inverse probability weighting performed similarly to propensity score regression methods.

In most cases, simpler models for the outcome (means of regression with no additional covariates except treatment and, if applicable, propensity score) resulted in better cost estimates than regression models with more covariates. It appears that adding covariates also adds assumptions about the functional form of their relationship with the outcome and mitigates some of the balance achieved by matching or weighting. Cost estimates may be particularly sensitive to model misspecification. In our example, the analysis model was conventional – a log link, two-way interactions and squared covariates – but not exactly the same as the model which was used to synthesise the data. The simpler models do not have this drawback. However, the use of mean methods is not always feasible or desirable.

In our synthesised dataset, all patients died before the analysis was performed. This made it possible to use the mean method as well as Weibull survival regression analysis. If not all patients have died – e.g. if not all ‘events’ have taken place – only survival regression remains as a way to estimate survival differences. Thus technique can used to calculate mean survival even when some patients have not yet died by the end of the follow-up period. Costs can also be estimated using regression in this case, if the model takes into account the

fact that some patients have reached the final episode of their treatment, while others have not.

Among the methods examined in our study, only propensity-score-as-covariate regression can accommodate censoring and meet the desire for simplicity in the outcome model. It is not unlikely, however, that propensity-score matching or inverse probability weighting followed by regression with a simple model would also achieve good results. These would not require an assumption about the functional form of the relationship between the propensity score and the outcome.

Standard regression without using propensity scores performed very well in terms of reliability and effect estimates, but not on costs and cost-effectiveness. Neither did instrumental variable regression.

Apart from differences in performance, methods also differ on a conceptual level. Adjustment for confounding can be approached from two angles. Defined simply, confounding is the combination of two associations – the association of a variable with the outcome (making it a risk factor) and the association of this variable with treatment assignment.(3) The problem can be solved by addressing either of these associations.

Regression focuses on modelling the effect of the risk factor on the outcome. Other methods eliminate the association of the confounder with treatment. The treatment effect that is estimated by regression is the average effect of treatment over all observed values of covariates. The possible number of covariates is not limitless. It is restricted by the number of subjects and the number of events. A ratio of 10 to 15 subjects or events per independent variable has been mentioned.(21,22) If the sample is too small to support a large number of variables, estimates will either be biased due to the incidental parameter problem if all variables are included or residual confounding (omitted variable bias) if some are omitted.

A final disadvantage of this technique is the danger that the analysis will still give results even if the treatment groups are very dissimilar and overlap is too limited.(23) The results may therefore be invalid as a consequence.

The dissimilarities between treatment groups will be noticed when patients in the treatment group are matched with treatment candidates who did not get the treatment. Matching can be done individually or group-wise using a number of covariates (e.g., age, sex) that can influence the effect of treatment. This technique makes treatment groups more similar, but in order to match each patient a limited number of covariates and continuous variables may have to be reclassified in order to be able to find matches.

Propensity score matching is gaining popularity among epidemiologists as a more sophisticated and feasible way of matching or adjusting. The sample size problem is partly resolved if the outcome is a rare count while treatment is not rare. In the treatment model, statistical significance is not a requirement for covariates to be included. The performance of

propensity scores in clinical effects has been shown to lead to somewhat different, and possibly better, results than regression without adjustment for differences in treatment groups. However, this point is still being debated. The choice of the most appropriate propensity score matching method in specific cases was discussed by Basur.(24)

Inverse probability weighting (IPW) has similarities with propensity score matching and it therefore shares some strengths and weaknesses. A major difference with matching methods and regression is that it is a form of direct standardisation: the average treatment effect is estimated for the full population from which the study sample was drawn. The effect reflects the treatment effect if everybody in the sample were treated, compared to when nobody were treated. This may be different from the average treatment effect in the treated, which is the result after propensity score matching. This method is an example of indirect standardisation. The estimation is limited to people who were similar to people who were actually treated. In our example, the results for propensity scores techniques and inverse probability weighting were not very consistently different.

Sometimes an instrumental variable can be used to provide an unbiased estimate of the causal effect of treatment. Like a randomiser an instrumental variable must be predictive of the treatment choice but may not cause the outcome nor be associated with any of the potential observed and unobserved confounders, so it can only impact the outcome through the treatment. Finding a suitable instrument is often problematic. There is no way to test whether that this is the case, unless several instruments can be identified. In our example, instrumental variable regression did not perform well, although the instrument was predictive of treatment assignment and did not cause the outcome. There was some correlation with other confounders, which led to biased estimates of costs.

Without some form of correction, inverse probability weighting leads to artificially small standard errors, due to the fact that the samples are enlarged by the application of the inverse probability weights. Using stabilized weights has been proposed as a way to address this issue, but we used the non-parametric bootstrap technique.

Bootstrapping has a two advantages in cost-effectiveness studies. It is likely to provide correct estimates of standard errors for costs, the residuals of which may be distributed in a non-normal way. Furthermore, because of the large number of iterations with separate estimates for incremental costs and effects, they also relate to other tools used in economic evaluation: CE planes and acceptability curves.

(2) How do the different methods compare with regard to feasibility?

All applied methods were comparably feasible. Regarding data requirements we can conclude that to adequately enable correction for confounding, data on all confounding variables, as well as incremental effects and costs outcomes, is essential in every method.

The results of all methods can be used for the construction of CE planes and acceptability curves. Also regarding expertise requirements the correction methods were comparable. All analyses were performed in Stata (Statacorp 2009), for which contains a pre-programmed command for propensity-score estimation and matching is available. The elements into which all methods can be divided – such as calculating and applying weights, estimating mean predicted survival an costs - are comparable across methods.

(4) How can these conclusions be applied to the categories of medicines on the list of expensive medicines?

The list of expensive medicines contains many therapies against cancer and other life-threatening diseases, for which survival is the primary outcome measure. For this reason, our study was based on the simulation of cancer patients.

We applied our adjustment methods on four quite differently biased datasets. Still, the best methods performed consistently across these datasets. Because of the diversity of the bias in the synthesized datasets, we expect that these results are generalizable to medicines on the list of expensive medicines, even though each listed medicine may have its own real-world setting and its own characteristics. From a conceptual point of view as well, there seems to be no reason to expect otherwise. A choice could be based on the desire to produce an average treatment effect in the treated, in which case propensity-score matching or propensity-score-as-a-covariate regression would be appropriate – or a population-average treatment effect, in which case inverse probability weighting is applicable.

The only exception is when there is much unmeasured confounding. In that case, it is recommended that an analysis with instrumental variable regression is attempted, since only this method can deal with unmeasured confounding by mimicking the randomisation through the instrumental variable.

Limitations

This study has a number of limitations. A study based on simulation is not the same as a study on real data. However, without simulation the 'real' effect would not be known and bias would not have been quantified. The synthesised data was based on real data, from a randomised controlled trial and an observational study of chemotherapy in patients with colorectal carcinoma.

We assumed that no censoring took place. This accommodated the straightforward method of calculating means. Censoring would have led to a higher degree of uncertainty.

We also assumed that there was no unmeasured confounding. When treatments are assigned to patients, something like the intuition of the treating physician may play a role. This cannot be explicitly expressed in a variable for which adjustment can take place. On the

other hand, if balance is achieved on the major predictors of the outcome, this may not always be a problem. However, it cannot be ruled out that the physician somehow has more information than the data show.

In cost-effectiveness studies, survival is often adjusted for health-related quality of life, in quality-adjusted life years (QALYs). This was not addressed in the current study, because it would make calculations more complicated without shedding more light on the problems.

Conclusions

Adjusting for confounding bias is most likely to be successful when the association of treatment and confounders is addressed, instead of the association of outcome and confounders. After estimating propensity scores in order to achieve covariate balance across treatment groups, the model estimating the outcome should be kept as simple as possible. This reduces the risk of misspecification of the functional form or the link function.

Inverse probability weighting and propensity-score-as-covariate regression are preferred because they provide estimates with narrower confidence intervals than one-to-one propensity score matching.

5 References

(1) Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999 Jan;10(1):37-48.

(2) Klungel OH, Martens EP, Psaty BM, Grobbee DE, Sullivan SD, Stricker BH, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol 2004 Dec;57(12):1223-31.

(3) Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed.: Lippincott Williams & Wilkins; 2008.

(4) Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health 2004 Apr;58(4):265-271.

(5) Hak E, Verheij TJ, Grobbee DE, Nichol KL, Hoes AW. Confounding by indication in non-experimental evaluation of vaccine effectiveness: the example of prevention of influenza complications. J Epidemiol Community Health 2002 Dec;56(12):951-955.

(6) Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007 Jan 17;297(3):278-85.

(7) Crown WH. Antidepressant selection and economic outcome: a review of methods and studies from clinical practice. Br J Psychiatry Suppl 2001 Sep;42:S18-22.

(8) Koopman M, Antonini NF, Douma J, Wals J, Honkoop AH, Erdkamp FL, et al. Randomised study of sequential versus combination chemotherapy with capecitabine, irinotecan and oxaliplatin in advanced colorectal cancer, an interim safety analysis. A Dutch Colorectal Cancer Group (DCCG) phase III study. Ann Oncol 2006 Oct;17(10):1523-1528.

(9) Koopman M, Antonini NF, Douma J, Wals J, Honkoop AH, Erdkamp FL, et al. Sequential versus combination chemotherapy with capecitabine, irinotecan, and oxaliplatin in advanced colorectal cancer (CAIRO): a phase III randomised controlled trial. Lancet 2007 Jul 14;370(9582):135-142.

(10) Mol L, Punt CJA, van Gils CWM, Ottevanger PB, Koopman M. Trial participation in a multicentre phase III trial (CAIRO) in advanced colorectal cancer patients in the Netherlands, and a comparison of outcome between trial and non-trial patients. (abstract 581PD). presented at the 35th ESMO Congress Milan, Italy 8-12 October 2010

.

(11) Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Med Decis Making 1998 Apr-Jun;18(2 Suppl):S68-80.

(12) Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 1984;79(387):516-24.

(13) Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55.

(14) B. DR,Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998 Oct 15;17(19):2265-81.

(15) Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol 2006 Jun 15;163(12):1149-1156.

(16) Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004 Sep;15(5):615-625.

(17) Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health 2006 Jul;60(7):578-86.

(18) Newhouse J, McClellan M. Econometrics in outcomes research: the use of instrumental variables. Annu Rev Public Health 1998;19:17–34.

(19) DiCiccio TJ EB. Bootstrap Confidence Intervals. Statistical Science 1996;11:189-212.

(20) van Hout BA, Al MJ, Gordon GS, Rutten FF. Costs, effects and C/E-ratios alongside a clinical trial. Health Econ 1994 Sep-Oct;3(5):309-319.

(21) Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996 Dec;49(12):1373-9.

(22) Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995 Dec;48(12):1503-10.

(23) Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 2005 Jun;58(6):550-559.

(24) Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. Value Health 2006 Nov-Dec;9(6):377-385.

Appendix 1 Results Validity

Biased dataset Nr.1 *n = 2000*	**Incremental effects (days)**		**Incremental costs (euro)**		**Net Monetary Benefit (euro)** *Threshold € 40,000*		*Threshold €8 0,000*	
	Mean		**Mean**		**Mean**		**Mean**	
Gold standard	133		11065		3448		15966	
	Bias (%)	**Deviation**	**Bias** (%)	**Deviation**	**Bias (%)**	**Deviation**	**Bias (%)**	**Deviation**
Mean	-2.5 (-1.9%)	17.9	3510 (31.7%)	3510	-3787 (-109.8%)	3832	-4063 (-25.4%)	4947
Simple regresssion	-2.1 (-1.6%)	18.0	3510 (31.7%)	3510	-3744 (-108.6%)	3797	-3979 (-24.9%)	4924
Full regression	-0.4 (-0.3%)	15.2	1897 (17.1%)	1970	-1946 (-56.4%)	2258	-1995 (-12.5%)	3492
PSM, mean (effects model)	-6.3 (-4.7%)	27.4	979 (8.9%)	1387	-1674 (-48.6%)	2700	-2350 (-14.7%)	5540
PSM, regression (effects model)	7.0 (5.3%)	23.5	523 (4.7%)	1204	232 (6.7%)	2167	1006 (6.3%)	4643
PSM, mean (costs model)	0.3 (0.2%)	29.3	-87 (-0.8%)	1366	136 (3.9%)	2357	186 (1.2%)	5461
PSM, regression (costs model)	11.9 (9%)	26.2	626 (5.7%)	1297	694 (20.1%)	2334	2015 (12.6%)	5064
PS covariate, simple regression (effects model)	-2.3 (-1.7%)	18.5	2632 (23.8%)	2636	-2879 (-83.5%)	3028	-3126 (-19.6%)	4446
PS covariate, full regression (effects model)	1.7 (1.3%)	16.2	1732 (15.7%)	1838	-1549 (-44.9%)	2081	-1367 (-8.6%)	3474
PS covariate, simple regression (costs model)	0.1 (0.1%)	18.9	1045 (9.4%)	1215	-1032 (-29.9%)	1875	-1019 (-6.4%)	3788
PS covariate, full regression (costs model)	3.0 (2.3%)	17.0	1620 (14.6%)	1743	-1293 (-37.5%)	1996	-967 (-6.1%)	3506
IPW, mean (effects model)	2.1 (1.6%)	20.5	2105 (19%)	2138	-1875 (-54.4%)	2425	-1646 (-10.3%)	4268
IPW, regression (effects model)	3.0 (2.3%)	17.9	1655 (15%)	1784	-1322 (-38.3%)	2005	-989 (-6.2%)	3604
IPW, mean (costs model)	3.7 (2.8%)	25.7	303 (2.7%)	1113	99 (2.9%)	2300	501 (3.1%)	5024
IPW, regression (costs model)	3.0 (2.3%)	20.9	1973 (17.8%)	2056	-1645 (-47.7%)	2342	-1317 (-8.2%)	4208
Instrumental variable	-12.0 (-9.1%)	11065	1810 (16.4%)	4215	-3125 (-90.6%)	5641	-4440 (-27.8%)	9621

Biased dataset Nr.2 *n = 2000*	**Incremental effects (days)**		**Incremental costs (euro)**		**Net Monetary Benefit (euro)** *Threshold €40,000*		*Threshold €80,000*	
	Mean		**Mean**		**Mean**		**Mean**	
Gold standard	143		10913		4707		17708	
	Bias (%)	**Deviation**	**Bias** (%)	**Deviation**	**Bias (%)**	**Deviation**	**Bias (%)**	**Deviation**
Mean	60.7 (42.5%)	60.7	4778 (43.8%)	4778	1865 (39.6%)	2203	8508 (48%)	8579
Simple regresssion	60.7 (42.5%)	60.7	4778 (43.8%)	4778	1868 (39.7%)	2226	8513 (48.1%)	8588
Full regression	-0.7 (-0.5%)	16.5	2466 (22.6%)	2488	-2543 (-54%)	2731	-2620 (-14.8%)	3890
PSM, mean (effects model)	3.7 (2.6%)	29.6	1067 (9.8%)	1508	-750 (-15.9%)	2611	-394 (-2.2%)	5673
PSM, regression (effects model)	7.6 (5.3%)	25.4	1254 (11.5%)	1572	-507 (-10.8%)	2338	280 (1.6%)	4892
PSM, mean (costs model)	5.8 (4.1%)	31.3	309 (2.8%)	1385	284 (6%)	2519	892 (5%)	5846
PSM, regression (costs model)	8.7 (6.1%)	26.4	1312 (12%)	1668	-402 (-8.5%)	2349	521 (2.9%)	5006
PS covariate, simple regression (effects model)	-4.5 (-3.1%)	19.8	2255 (20.7%)	2269	-2743 (-58.3%)	2976	-3231 (-18.2%)	4664
PS covariate, full regression (effects model)	1.1 (0.8%)	17.3	2186 (20%)	2227	-2068 (-43.9%)	2403	-1951 (-11%)	3749
PS covariate, simple regression (costs model)	-3.3 (-2.3%)	19.7	1273 (11.7%)	1389	-1630 (-34.6%)	2218	-1988 (-11.2%)	4156
PS covariate, full regression (costs model)	1.7 (1.2%)	17.3	2111 (19.3%)	2157	-1923 (-40.8%)	2309	-1734 (-9.8%)	3702
IPW, mean (effects model)	-9.3 (-6.5%)	21.5	1494 (13.7%)	1602	-2514 (-53.4%)	2813	-3534 (-20%)	4908
IPW, regression (effects model)	2.0 (1.4%)	18.8	2421 (22.2%)	2462	-2201 (-46.8%)	2554	-1982 (-11.2%)	3978
IPW, mean (costs model)	-8.5 (-6%)	22.1	438 (4%)	1053	-1373 (-29.2%)	2140	-2307 (-13%)	4437
IPW, regression (costs model)	2.4 (1.7%)	19.6	2552 (23.4%)	2580	-2291 (-48.7%)	2645	-2030 (-11.5%)	4135
Instrumental variable	-16.2 (-11.3%)	10913	-2369 (-21.7%)	4322	598 (12.7%)	5404	-1173 (-6.6%)	10344

Biased dataset Nr.3 *n = 2000*	**Incremental effects (days)**		**Incremental costs (euro)**		**Net Monetary Benefit (euro)** *Threshold € 40,000*		*Threshold €8 0,000*	
	Mean		**Mean**		**Mean**		**Mean**	
Gold standard	126		11466		2370		16263	
	Bias (%)	**Deviation**	**Bias** (%)	**Deviation**	**Bias (%)**	**Deviation**	**Bias (%)**	**Deviation**
Mean	-130.4 (-103.2%)	130.4	-3597 (-31.4%)	3597	-10682 (-450.8%)	10682	-24962 (-153.5%)	24962
Simple regresssion	-125.7 (-99.5%)	125.7	-3597 (-31.4%)	3597	-10174 (-429.3%)	10174	-23944 (-147.2%)	23944
Full regression	2.2 (1.7%)	14.9	418 (3.6%)	884	-180 (-7.6%)	1441	57 (0.4%)	2942
PSM, mean (effects model)	1.0 (0.8%)	21.5	92 (0.8%)	935	70 (2.9%)	1941	279 (1.7%)	4194
PSM, regression (effects model)	13.5 (10.7%)	26.6	729 (6.4%)	1215	802 (33.8%)	2237	2381 (14.6%)	5073
PSM, mean (costs model)	1.5 (1.2%)	21.7	99 (0.9%)	988	126 (5.3%)	1912	399 (2.5%)	4231
PSM, regression (costs model)	13.9 (11%)	26.4	744 (6.5%)	1227	842 (35.5%)	2197	2475 (15.2%)	5013
PS covariate, simple regression (effects model)	3.1 (2.5%)	16.4	-607 (-5.3%)	846	947 (39.9%)	1700	1286 (7.9%)	3394
PS covariate, full regression (effects model)	3.0 (2.4%)	15.1	480 (4.2%)	922	-154 (-6.5%)	1462	171 (1%)	2979
PS covariate, simple regression (costs model)	3.2 (2.6%)	16.4	-645 (-5.6%)	861	1000 (42.2%)	1705	1354 (8.3%)	3391
PS covariate, full regression (costs model)	3.1 (2.4%)	15.1	481 (4.2%)	919	-146 (-6.2%)	1457	188 (1.2%)	2978
IPW, mean (effects model)	16.3 (12.9%)	22.8	462 (4%)	784	1320 (55.7%)	1995	3101 (19.1%)	4454
IPW, regression (effects model)	4.0 (3.1%)	15.4	54 (0.5%)	829	380 (16%)	1504	814 (5%)	3087
IPW, mean (costs model)	16.3 (12.9%)	22.9	390 (3.4%)	753	1400 (59.1%)	2040	3190 (19.6%)	4511
IPW, regression (costs model)	4.0 (3.1%)	15.4	441 (3.9%)	891	-8 (-0.4%)	1466	425 (2.6%)	3035
Instrumental variable	12.5 (9.9%)	11466	2989 (26.1%)	3500	-1622 (-68.4%)	4150	-255 (-1.6%)	8587

Biased dataset Nr.4 *n = 2000*	**Incremental effects (days)**		**Incremental costs (euro)**		**Net Monetary Benefit (euro)** *Threshold € 40,000*		*Threshold €8 0,000*	
	Mean		**Mean**		**Mean**		**Mean**	
Gold standard	131		10766		3578		16343	
	Bias (%)	**Deviation**	**Bias** (%)	**Deviation**	**Bias (%)**	**Deviation**	**Bias (%)**	**Deviation**
Mean	-17.0 (-12.9%)	21.6	2677 (24.9%)	2677	-4533 (-126.7%)	4538	-6389 (-39.1%)	6586
Simple regresssion	-14.9 (-11.4%)	20.7	2677 (24.9%)	2677	-4313 (-120.5%)	4321	-5949 (-36.4%)	6220
Full regression	1.9 (1.5%)	15.2	1996 (18.5%)	2041	-1787 (-49.9%)	2207	-1579 (-9.7%)	3376
PSM, mean (effects model)	-11.3 (-8.6%)	31.2	269 (2.5%)	1321	-1468 (-41%)	2829	-2630 (-16.1%)	6120
PSM, regression (effects model)	14.2 (10.8%)	27.7	828 (7.7%)	1361	764 (21.3%)	2447	2391 (14.6%)	5365
PSM, mean (costs model)	-7.3 (-5.6%)	31.9	-392 (-3.6%)	1489	-317 (-8.8%)	2485	-1028 (-6.3%)	5873
PSM, regression (costs model)	17.2 (13.1%)	29.9	975 (9.1%)	1471	1001 (28%)	2577	2975 (18.2%)	5750
PS covariate, simple regression (effects model)	-1.5 (-1.2%)	18.0	2613 (24.3%)	2615	-2779 (-77.7%)	2961	-2945 (-18%)	4273
PS covariate, full regression (effects model)	4.2 (3.2%)	16.7	1763 (16.4%)	1855	-1305 (-36.5%)	2021	-847 (-5.2%)	3421
PS covariate, simple regression (costs model)	-0.1 (0%)	18.1	1534 (14.3%)	1587	-1541 (-43.1%)	2100	-1548 (-9.5%)	3753
PS covariate, full regression (costs model)	5.0 (3.8%)	16.9	1686 (15.7%)	1793	-1139 (-31.8%)	1957	-592 (-3.6%)	3417
IPW, mean (effects model)	1.0 (0.7%)	24.2	1839 (17.1%)	1949	-1733 (-48.4%)	2613	-1627 (-10%)	4916
IPW, regression (effects model)	5.7 (4.3%)	21.8	1894 (17.6%)	1999	-1273 (-35.6%)	2252	-653 (-4%)	4222
IPW, mean (costs model)	0.1 (0.1%)	25.2	611 (5.7%)	1247	-599 (-16.7%)	2306	-586 (-3.6%)	4929
IPW, regression (costs model)	5.3 (4%)	22.4	2049 (19%)	2117	-1470 (-41.1%)	2364	-890 (-5.4%)	4387
Instrumental variable	-7.6 (-5.8%)	10766	3231 (30%)	3921	-4063 (-113.5%)	5441	-4894 (-29.9%)	9344

Biased dataset Nr.1 *n = 400*	**Incremental effects (days)**		**Incremental costs (euro)**		**Net Monetary Benefit (euro)** *Threshold €40,000*		*Threshold €80,000*	
	Mean		**Mean**		**Mean**		**Mean**	
Gold standard	133		11036		3517		16150	
	Bias (%)	**Deviation**	**Bias** (%)	**Deviation**	**Bias (%)**	**Deviation**	**Bias (%)**	**Deviation**
Mean	2.0 (1.5%)	40.5	3605 (32.7%)	3665	-3386 (-96.3%)	4531	-3166 (-19.6%)	8224
Simple regresssion	2.6 (2%)	41.0	3605 (32.7%)	3665	-3318 (-94.3%)	4543	-3032 (-18.8%)	8291
Full regression	0.2 (0.2%)	36.3	1973 (17.9%)	2692	-1946 (-55.3%)	3731	-1920 (-11.9%)	7253
PSM, mean (effects model)	-4.1 (-3.1%)	62.6	1056 (9.6%)	2736	-1586 (-45.1%)	5535	-2010 (-12.4%)	12160
PSM, regression (effects model)	26.6 (20%)	56.3	1100 (10%)	2750	1736 (49.4%)	4979	4677 (29%)	10874
PSM, mean (costs model)	-1.6 (-1.2%)	66.8	-120 (-1.1%)	3012	-5 (-0.1%)	5369	-114 (-0.7%)	12474
PSM, regression (costs model)	32.6 (24.5%)	60.1	1293 (11.7%)	2932	2323 (66%)	5200	5953 (36.9%)	11491
PS covariate, simple regression (effects model)	-0.6 (-0.5%)	41.9	2688 (24.4%)	2960	-2754 (-78.3%)	4366	-2820 (-17.5%)	8500
PS covariate, full regression (effects model)	2.6 (1.9%)	37.9	1689 (15.3%)	2566	-1409 (-40.1%)	3766	-1129 (-7%)	7485
PS covariate, simple regression (costs model)	1.1 (0.9%)	43.9	1001 (9.1%)	2070	-875 (-24.9%)	3953	-749 (-4.6%)	8558
PS covariate, full regression (costs model)	3.4 (2.6%)	39.4	1516 (13.7%)	2481	-1141 (-32.4%)	3887	-766 (-4.7%)	7846
IPW, mean (effects model)	2.5 (1.9%)	48.2	2167 (19.6%)	2751	-1890 (-53.7%)	4476	-1614 (-10%)	9441
IPW, regression (effects model)	2.4 (1.8%)	42.1	1833 (16.6%)	2708	-1565 (-44.5%)	3929	-1297 (-8%)	8086
IPW, mean (costs model)	3.3 (2.5%)	54.9	473 (4.3%)	2401	-110 (-3.1%)	4810	254 (1.6%)	10628
IPW, regression (costs model)	1.1 (0.9%)	46.7	1976 (17.9%)	2844	-1851 (-52.6%)	4352	-1726 (-10.7%)	9068
Instrumental variable	-8.6 (-6.5%)	11036	1875 (17%)	8090	-2816 (-80.1%)	11983	-3757 (-23.3%)	22605

Biased dataset Nr.2 *n = 400*	**Incremental effects (days)**		**Incremental costs (euro)**		**Net Monetary Benefit (euro)** *Threshold €40,000*		*Threshold €80,000*	
	Mean		**Mean**		**Mean**		**Mean**	
Gold standard	143		10860		4790		17989	
	Bias (%)	**Deviation**	**Bias** (%)	**Deviation**	**Bias (%)**	**Deviation**	**Bias (%)**	**Deviation**
Mean	64.4 (45.1%)	69.2	4866 (44.8%)	4869	2190 (45.7%)	4015	9246 (51.4%)	11161
Simple regresssion	64.6 (45.2%)	69.6	4866 (44.8%)	4869	2206 (46%)	4075	9278 (51.6%)	11260
Full regression	1.2 (0.9%)	37.5	2721 (25.1%)	3193	-2586 (-54%)	4138	-2452 (-13.6%)	7650
PSM, mean (effects model)	5.1 (3.6%)	67.3	1107 (10.2%)	2831	-887 (-18.5%)	5830	-480 (-2.7%)	13013
PSM, regression (effects model)	28.7 (20.1%)	58.7	1701 (15.7%)	2985	1103 (23%)	5082	4096 (22.8%)	11164
PSM, mean (costs model)	5.1 (3.6%)	67.5	261 (2.4%)	2799	46 (1%)	5551	472 (2.6%)	12769
PSM, regression (costs model)	29.0 (20.3%)	60.2	1756 (16.2%)	3096	1172 (24.5%)	5141	4220 (23.5%)	11378
PS covariate, simple regression (effects model)	-4.3 (-3%)	44.1	2290 (21.1%)	2685	-2759 (-57.6%)	4535	-3228 (-17.9%)	8970
PS covariate, full regression (effects model)	1.9 (1.3%)	39.3	2250 (20.7%)	2884	-2046 (-42.7%)	4118	-1841 (-10.2%)	7891
PS covariate, simple regression (costs model)	-3.0 (-2.1%)	44.4	1274 (11.7%)	2172	-1600 (-33.4%)	4197	-1926 (-10.7%)	8811
PS covariate, full regression (costs model)	2.4 (1.7%)	39.8	2096 (19.3%)	2802	-1837 (-38.3%)	4099	-1578 (-8.8%)	7969
IPW, mean (effects model)	-7.4 (-5.2%)	48.7	1600 (14.7%)	2522	-2414 (-50.4%)	4621	-3229 (-17.9%)	9647
IPW, regression (effects model)	0.3 (0.2%)	45.8	2543 (23.4%)	3234	-2506 (-52.3%)	4526	-2469 (-13.7%)	8980
IPW, mean (costs model)	-5.6 (-3.9%)	49.7	593 (5.5%)	2299	-1207 (-25.2%)	4385	-1821 (-10.1%)	9585
IPW, regression (costs model)	1.1 (0.8%)	47.2	2700 (24.9%)	3238	-2577 (-53.8%)	4702	-2454 (-13.6%)	9308
Instrumental variable	-11.1 (-7.8%)	10860	-1123 (-10.3%)	7311	-90 (-1.9%)	10784	-1303 (-7.2%)	21988

Biased dataset Nr.3 *n = 400*	**Incremental effects (days)**		**Incremental costs (euro)**		**Net Monetary Benefit (euro)** *Threshold €40,000*		*Threshold €80,000*	
	Mean		**Mean**		**Mean**		**Mean**	
Gold standard	127		11456		2451		16849	
	Bias (%)	**Deviation**	**Bias** (%)	**Deviation**	**Bias (%)**	**Deviation**	**Bias (%)**	**Deviation**
Mean	-126.2 (-99.4%)	126.4	-3460 (-30.2%)	3547	-10360 (-422.7%)	10381	-24180 (-143.5%)	24216
Simple regresssion	-121.5 (-95.7%)	121.8	-3460 (-30.2%)	3547	-9848 (-401.8%)	9885	-23155 (-137.4%)	23214
Full regression	4.9 (3.9%)	33.0	584 (5.1%)	1795	-47 (-1.9%)	3233	491 (2.9%)	6649
PSM, mean (effects model)	2.9 (2.3%)	49.6	46 (0.4%)	2155	530 (21.6%)	4425	1323 (7.9%)	9710
PSM, regression (effects model)	31.8 (25%)	56.9	1262 (11%)	2573	2471 (100.8%)	5081	6420 (38.1%)	11255
PSM, mean (costs model)	3.1 (2.4%)	51.0	129 (1.1%)	2146	488 (19.9%)	4591	1327 (7.9%)	10121
PSM, regression (costs model)	32.1 (25.3%)	56.7	1324 (11.6%)	2629	2469 (100.8%)	5057	6485 (38.5%)	11245
PS covariate, simple regression (effects model)	8.6 (6.7%)	37.1	-200 (-1.7%)	1570	1139 (46.5%)	3559	2077 (12.3%)	7482
PS covariate, full regression (effects model)	9.2 (7.2%)	35.2	1001 (8.7%)	2023	3 (0.1%)	3368	1007 (6%)	6969
PS covariate, simple regression (costs model)	8.8 (6.9%)	37.5	-230 (-2%)	1572	1193 (48.7%)	3599	2156 (12.8%)	7580
PS covariate, full regression (costs model)	9.5 (7.5%)	35.4	1019 (8.9%)	2029	22 (0.9%)	3395	1064 (6.3%)	7022
IPW, mean (effects model)	20.6 (16.3%)	44.1	568 (5%)	1664	1693 (69.1%)	3988	3954 (23.5%)	8684
IPW, regression (effects model)	9.6 (7.6%)	35.1	479 (4.2%)	1824	575 (23.5%)	3427	1629 (9.7%)	7058
IPW, mean (costs model)	20.5 (16.1%)	44.1	485 (4.2%)	1653	1760 (71.8%)	4035	4005 (23.8%)	8738
IPW, regression (costs model)	9.5 (7.5%)	35.0	688 (6%)	1836	351 (14.3%)	3365	1390 (8.2%)	6997
Instrumental variable	22.6 (17.8%)	11456	3823 (33.4%)	5829	-1348 (-55%)	9705	1126 (6.7%)	21564

Biased dataset Nr.4 *n = 400*	**Incremental effects (days)**		**Incremental costs (euro)**		**Net Monetary Benefit (euro)** *Threshold €40,000*		*Threshold €80,000*	
	Mean		**Mean**		**Mean**		**Mean**	
Gold standard	131		10770		3577		16367	
	Bias (%)	**Deviation**	**Bias** (%)	**Deviation**	**Bias (%)**	**Deviation**	**Bias (%)**	**Deviation**
Mean	-16.0 (-12.2%)	42.2	2688 (25%)	2810	-4436 (-124%)	5179	-6184 (-37.8%)	9325
Simple regresssion	-14.1 (-10.7%)	42.2	2688 (25%)	2810	-4229 (-118.2%)	5082	-5769 (-35.2%)	9233
Full regression	2.2 (1.7%)	37.4	2043 (19%)	2708	-1799 (-50.3%)	3796	-1556 (-9.5%)	7401
PSM, mean (effects model)	-12.2 (-9.3%)	69.1	278 (2.6%)	2888	-1602 (-44.8%)	6080	-2726 (-16.7%)	13424
PSM, regression (effects model)	36.3 (27.7%)	63.4	1362 (12.6%)	2939	2629 (73.5%)	5535	6820 (41.7%)	12356
PSM, mean (costs model)	-5.4 (-4.1%)	68.7	-375 (-3.5%)	3060	-64 (-1.8%)	5688	-397 (-2.4%)	13029
PSM, regression (costs model)	40.6 (31%)	66.0	1587 (14.7%)	2948	3014 (84.3%)	5714	7721 (47.2%)	12818
PS covariate, simple regression (effects model)	-1.2 (-0.9%)	43.9	2613 (24.3%)	2923	-2747 (-76.8%)	4540	-2882 (-17.6%)	8879
PS covariate, full regression (effects model)	6.8 (5.2%)	40.7	1823 (16.9%)	2690	-1075 (-30.1%)	3873	-326 (-2%)	7899
PS covariate, simple regression (costs model)	0.2 (0.1%)	44.6	1506 (14%)	2304	-1487 (-41.6%)	4161	-1467 (-9%)	8790
PS covariate, full regression (costs model)	8.0 (6.1%)	41.6	1716 (15.9%)	2638	-843 (-23.6%)	3931	30 (0.2%)	8070
IPW, mean (effects model)	3.5 (2.7%)	56.4	1808 (16.8%)	2950	-1428 (-39.9%)	5186	-1048 (-6.4%)	11017
IPW, regression (effects model)	6.2 (4.8%)	47.3	1988 (18.5%)	2944	-1305 (-36.5%)	4294	-621 (-3.8%)	9051
IPW, mean (costs model)	3.3 (2.5%)	61.1	641 (6%)	2686	-280 (-7.8%)	5438	81 (0.5%)	11910
IPW, regression (costs model)	5.3 (4.1%)	50.3	2119 (19.7%)	2946	-1538 (-43%)	4577	-956 (-5.8%)	9714
Instrumental variable	-14.0 (-10.7%)	10770	3042 (28.2%)	6455	-4577 (-128%)	10203	-6113 (-37.3%)	20467

Appendix 2 Results reliability

n = 2000	**Biased dataset Nr. 1**		**Biased dataset Nr. 2**		**Biased dataset Nr.3**		**Biased dataset Nr. 4**	
	Width		**Width**		**Width**		**Width**	
	Effects (days)	**Costs(euro)**	**Effects (days)**	**Costs(euro)**	**Effects (days)**	**Costs(euro)**	**Effects (days)**	**Costs(euro)**
Mean	95.3	3594	95.5	3764	95.4	3815	98.2	3702
Simple regression	95.7	3594	95.9	3764	95.2	3815	98.6	3702
Full regression	83.6	4878	86.0	5006	76.1	3726	91.0	4801
PSM, mean (effects model)	151.1	6250	140.2	5950	117.0	5180	173.4	8142
PSM, regression (effects model)	124.3	6499	119.8	6905	116.1	5192	135.1	6268
PSM, mean (costs model)	151.5	7425	148.2	6198	116.9	5153	187.8	10094
PSM, regression (costs model)	123.8	6626	127.8	6905	115.6	5273	140.9	6343
PS covariate, simple regression (effects model)	100.0	4044	95.1	4060	82.8	3633	103.1	4425
PS covariate, full regression (effects model)	87.6	4710	89.6	5296	76.5	3795	93.4	4923
PS covariate, simple regression (costs model)	104.0	4331	96.8	4073	82.8	3579	104.4	4539
PS covariate, full regression (costs model)	91.1	4801	90.5	5302	76.5	3800	96.3	4983
IPW, mean (effects model)	99.4	4406	87.3	4195	95.7	3961	114.9	5865
IPW, regression (effects model)	98.1	5726	88.2	5191	76.7	3804	101.4	5687
IPW, mean (costs model)	117.4	6480	89.6	4972	95.7	3918	118.4	6308
IPW, regression (costs model)	114.3	5657	87.3	5009	76.7	3710	104.4	5051
Instrumental variable	245.1	1.8E+04	232.9	1.7E+04	222.4	1.1E+04	230.4	1.2E+04

n = 400	Biased dataset Nr. 1		Biased dataset Nr. 2		Biased dataset Nr. 3		Biased dataset Nr.4	
	Width		Width		Width		Width	
	Effects (days)	Costs(euro)	Effects (days)	Costs(euro)	Effects (days)	Costs(euro)	Effects (days)	Costs(euro)
Mean	207.1	7636	199.7	7779	211.4	8522	226.7	8733
Simple regresssion	209.9	7636	203.1	7779	211.6	8522	227.1	8733
Full regression	189.0	10088	188.6	11163	168.9	7908	200.2	11497
PSM, mean (effects model)	321.3	11910	313.5	12590	269.5	11605	320.9	14899
PSM, regression (effects model)	281.2	13841	260.6	16193	256.8	11039	294.0	16950
PSM, mean (costs model)	320.4	13156	326.2	13443	268.7	12012	384.0	17629
PSM, regression (costs model)	282.1	14884	283.5	17231	255.8	11332	321.0	17385
PS covariate, simple regression (effects model)	226.8	7985	218.1	9599	179.3	8206	245.6	11128
PS covariate, full regression (effects model)	202.5	10190	194.2	11947	175.8	8410	221.1	12777
PS covariate, simple regression (costs model)	220.2	7954	218.9	9685	180.0	8195	249.0	11807
PS covariate, full regression (costs model)	201.6	9991	195.1	12000	176.5	8450	223.9	12821
IPW, mean (effects model)	218.3	7996	190.5	10334	204.6	8811	402.6	27036
IPW, regression (effects model)	203.2	9672	180.7	14328	177.6	8380	228.9	13457
IPW, mean (costs model)	203.8	8010	210.2	10962	206.1	8826	346.4	24513
IPW, regression (costs model)	199.5	10185	186.9	11309	177.5	8049	230.8	12248
Instrumental variable	691.1	4.3E+04	475.3	3.5E+04	436.3	2.4E+04	487.0	2.8E+04