An instrumental variable approach to statistical matching

Merbel, Anka van de

An instrumental variable approach to statistical matching

Anka Esmée van de Merbel

Thesis advisors CBS:

Prof. Ton de Waal, Dr. Arnout van Delden, Dr. Sander Scholtus

Thesis advisor Universiteit Leiden:

Prof. Mark de Rooij

Defended on March 20th, 2024

MASTER THESIS

STATISTICS AND DATA SCIENCE

UNIVERSITEIT LEIDEN

Dear reader

Welcome to this journey that is my Master's thesis. Before you read any further I must take a moment to express my gratitude to those who have helped me throughout this project.

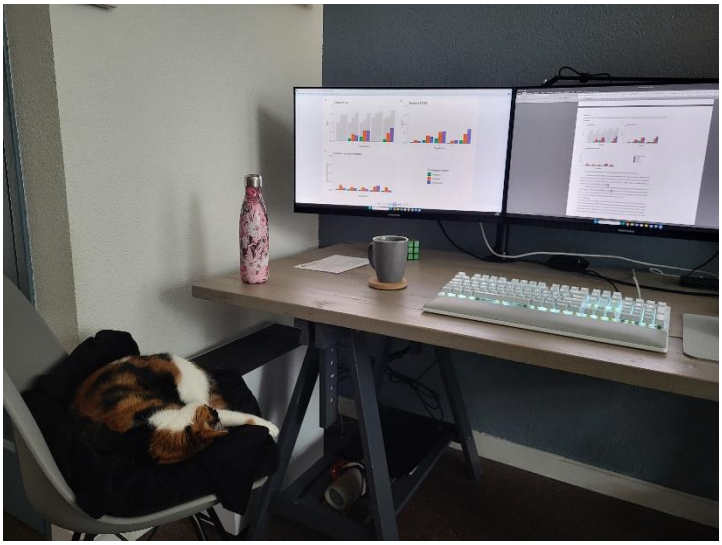
First and foremost, I would like to thank my thesis supervisors at CBS, Dr. Sander Scholtus, Dr. Arnout van Delden and Prof. Ton de Waal. Your guidance and extensive feedback have been invaluable to this process. Your expertise on various topics and the insightful discussions have deepened my understanding of statistical matching but also of probability and statistics in general. I would also like to extend my appreciation to my thesis supervisor at Leiden University, Prof. Mark de Rooij for his feedback and support.

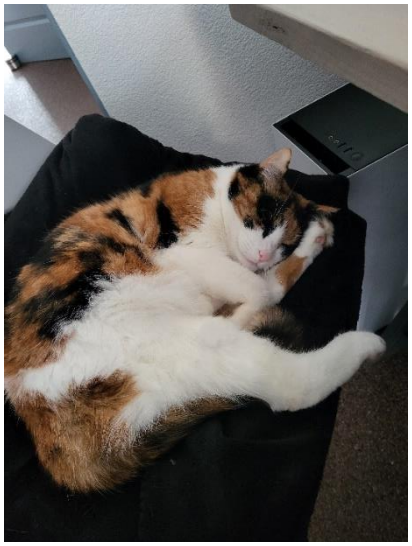
I am eternally grateful to my friends and family. Enduring my tales of obscure statistical analyses and the woes of interning at CBS took a great deal of patience, I am sure. At times I doubted myself, but your unwavering belief in my abilities have kept me (somewhat) sane. Special thanks to Foxie the cat, who was a crucial player in the writing process of this thesis, providing both spiritual and emotional support as well as a calm working environment.

To the reader; I would like to thank you for your interest in this project and I truly hope you will learn something on the way.

Sincerely,

Anka





Abstract

Statistical matching is a technique which can be applied when one wants to investigate the joint relationship between two variables that are observed in different datasets, using one or more variables that overlap in both datasets. This joint relationship cannot be estimated without relying on assumptions or additional data. Classically, statistical matching is based on the Conditional Independence Assumption (CIA) which asserts the non-overlapping variables to be independent given the overlapping variable. This assumption is inflexible, untestable and often does not hold. The current project proposes to use an approach based on the Instrumental Variable Assumption (IVA).

An instrumental variable is a variable that, given the value of some mediating variable, has no effect on some outcome variable. In the context of statistical matching this gives rise to three scenarios: the mediating variable overlaps, the outcome variable overlaps, or the instrumental variable overlaps. The IVA approach is more flexible than the CIA approach. This is because the IVA approach does not make any assumptions on which variable is the overlapping variable, whereas the CIA always conditions on the overlapping variable. The aims of the current study were twofold: 1) how does the IVA approach perform when the assumption is violated to various degrees and 2) how does the IVA approach compare to the CIA approach. To answer these questions, a simulation study was performed. For each scenario, joint probabilities of the non-overlapping variables were estimated under both the IVA and the CIA in populations which violate the IVA to various degrees. Measures for the bias, accuracy and precision were estimated and compared.

The results indicate that the IVA approach is moderately robust against slight violations of the assumption. When the IVA is not violated, estimations are unbiased and for all matching scenarios the method outperforms the CIA. When the IVA is violated it is advisable to rely on the CIA, since results of the current simulation study suggest the CIA to be more robust against violations in general.

Table of contents

Chapter 1

Introduction

National Statistical Institutes (NSIs) such as Statistics Netherlands (CBS) have the responsibility of providing society with correct and extensive statistics. For this reason NSIs have access to a wide variety of data sources, from survey data to administrative data. CBS alone has over 4600 datasets published for public use (CBS, 2023). In general, CBS tries to minimize the burden on citizens which means they do not invite citizens to answer questions too often. Social surveys are designed such that the overlap between the units in different samples is minimized. A consequence of this policy is that two variables of interest might not be observed in the same dataset. Combining survey and selective administrative datasets may be another reason why two variables are not observed in the same dataset. In this situation it is (or seems) impossible to infer the relationship between these variables, as they have not been measured for the same units, which is where the technique of statistical matching comes in.

Statistical matching is a method where two (or more) disjoint datasets with overlapping and non-overlapping variables are integrated (D'Orazio et al., 2006). The main goal is to estimate the relationship between the variables which have been observed independent of each other. For example, consider one dataset where health outcomes and tax rate are observed together and a second dataset where tax rate and smoking are observed together. Through statistical matching one can use the information on tax rate to investigate the relationship between health outcomes and smoking even though these variables were not observed for the same persons.

An integral assumption for many statistical matching methods is the Conditional Independence Assumption (CIA). The CIA contends that, given the overlapping variables, the non-overlapping variables are independent (D'Orazio et al., 2006). In practice it is not possible to test whether this assumption holds using the data available and the assumption is often violated (Leulescu & Agafitei, 2013). The CIA is usually reasonable when the overlapping variable is a proxy for one of the target variables (D'Orazio et al., 2024). Violating the CIA results in an incorrectly specified model, causing biased outcomes. When the CIA is violated there are two commonly proposed options; 1) one can opt to use auxiliary information, such as a third dataset where all variables are jointly observed or 2) one can use some plausible value for the unknown parameters e.g. based on literature or on past data (D'Orazio et al., 2006). This project proposes a third option: making use of an instrumental variable (IV).

An IV is often used in linear regression to account for confounding variables (Newhouse & McClellan, 1998). This variable has no direct effect on the outcome variable but does have an effect on

one of the regressors. In essence, the regressor serves as the mediating variable between the instrumental and the outcome variable. An illuminating example can be found on Wikipedia: suppose a researcher wishes to estimate the effect of smoking on health and knows that the correlation between smoking and health does not imply smoking to cause bad health. The researcher can opt to use tax on tobacco as an instrument, assuming tax on tobacco affects health only through smoking. Any correlation found between tax and health can then be attributed to smoking ("Instrumental variables estimation", 2024, Example).

The use of IVs is well-documented in several research fields, however it has been scarcely utilized in statistical matching. In fact, only one publication was identified where the researchers made use of an IV in a statistical matching setting (Kim et al., 2016). In this study fractional imputation was used to statistically match continuous variables. To make their model identifiable, instead of relying on the CIA, the researchers included an IV in their model. This IV was also the overlapping variable between the to be matched datasets.

The study by Kim et al. (2016) uses continuous variables and does not directly assess the performance of using an IV compared to using the CIA. Additionally, they only assess the situation where the IV is the overlapping variable. The use of IVs in statistical matching might be a viable alternative to relying on the CIA, since the latter is often violated. The aim of this project is then to estimate the bias, accuracy and precision of statistical matching using the Instrumental Variable Assumption (IVA) and compare it to the bias, accuracy and precision when using the CIA. This is done for three possible matching situations; where the instrumental, mediator or outcome variable overlaps in the data. Since data at CBS is often in categorical form and analyses are frequently done on contingency tables and the IVA approach has only been used with continuous data before, this project focusses on categorical data.

Chapter 2

Background

2.1 Statistical Matching

As mentioned, statistical matching is concerned with estimating the joint distribution of variables that have not been observed in the same dataset using information from variables that were observed in both datasets. In this section the statistical matching problem is explained in detail, and a working example will be provided and used throughout the thesis.

2.1.1 The statistical matching problem

The statistical matching technique can be seen as a missing data problem where each dataset has all values from one variable missing (Kim et al., 2016). Consider two samples *A* and *B* with n_A and n_B i.i.d. observations. Also consider the categorical random variables X, Y and Z with categories $x = 1, \dots, C_x$, $y = 1, \dots, C_y$ and $z = 1, \dots, C_z$, with joint probability distribution *P*(X, Y, Z). Sample *A* includes variables X and Y and has variable Z missing. Sample *B* includes variables X and Z and has variable Y missing. The goal of statistical matching is to use the information from both samples *A* and *B* to estimate the joint distribution of the two non-overlapping variables; *P*(Z,Y). p_{zy} will be used to denote the probability of variable Z taking value z and variable Y taking value y. Figure 2.1 depicts a schematic overview of these two datasets.

2.1.1a A working example

The following is an example that will be used throughout this thesis: suppose an NSI has data on tobacco tax (T $with\ t = high, low$), smoking (S $with\ s = yes, no$) and health outcomes (H $with\ h = adverse, good$) and a researcher is interested in the relationship between smoking and health outcomes. They have access to two samples; one where tobacco tax and smoking are observed together, and one where tobacco tax and health outcomes are observed together. The common variable is tobacco tax and the researcher is then interested in the distribution $P(\mathrm{S}, \mathrm{H})$.

2.1.2 Approaches to statistical matching

Broadly speaking, there are two approaches to statistical matching; the macro and the micro approach. The macro approach is concerned with estimating the joint distribution of variables that do not overlap in the data. The micro approach goes a step further, creating a synthetic dataset that includes all

Figure 2.1

Schematic overview of two datasets suitable for statistical matching

	Variables		
Dataset *A*	$y_1^{(A)}$	$x_1^{(A)}$	
	$y_2^{(A)}$	$x_2^{(A)}$	
	$y_{n_A}^{(A)}$	$x_{n_A}^{(A)}$	
Dataset *B*		$x_1^{(B)}$	$z_1^{(B)}$
		$x_2^{(B)}$	$z_2^{(B)}$
		$x_{n_B}^{(B)}$	$z_{n_B}^{(B)}$

variables. The micro approach uses the estimated joint distribution and is therefore an extension to the macro approach (D'Orazio et al., 2006).

Approaches to statistical matching can further be divided into parametric and non-parametric approaches. Parametric methods are based on the assumption that the unknown joint distribution comes from a specific model (e.g. conditional mean matching). Non-parametric methods do not make such model assumptions (e.g. hot deck methods). For an extensive overview of statistical matching methods, the reader is referred to D'Orazio et al. (2006).

2.1.3 The Conditional Independence Assumption

The main problem in statistical matching is that there is no sufficient information in $A \cup B$ to construct an identifiable model for $(\mathrm{X}, \mathrm{Y}, \mathrm{Z})$ and consequently to estimate $P(\mathrm{Z}, \mathrm{Y})$ (Conti et al., 2017). The assumption that is generally made in statistical matching for the model to be identifiable is the CIA. This assumption asserts that the two non-overlapping variables *Z* and *Y* are independent given the value of the overlapping variable *X*. Formally this means that under the CIA $P(\mathrm{Z}, \mathrm{Y})$ can be written as

$$P(\mathrm{Z}, \mathrm{Y}) = \sum_x P(\mathrm{Z} \mid \mathrm{X} = x) P(\mathrm{Y} \mid \mathrm{X} = x) P(\mathrm{X} = x) \tag{1}$$

Applying this idea to the working example: to statistically match the samples, the researcher at the NSI assumes that smoking is independent of health outcomes, given the tobacco tax (the overlapping variable), meaning: $P(\mathrm{S}, \mathrm{H}) = \sum_t P(\mathrm{S} \mid \mathrm{T} = t) P(\mathrm{H} \mid \mathrm{T} = t) P(\mathrm{T} = t)$.

In practice, this assumption often does not strictly hold and cannot be tested from $A \cup B$ because there is no information on the joint distribution of the non-overlapping variables in the samples (D'Orazio et al., 2006; Conti et al., 2017). The consequence of using a wrong assumption is that the

model used for matching is incorrectly specified and the result does not reflect the true underlying model, but instead the model assumption which causes biased results (D'Orazio et al., 2006).

2.2 The Instrumental Variable approach

Another approach to statistical matching rather than relying on the CIA is proposed: making use of an IV. The following section outlines what IVs are and how they can be used in statistical matching.

2.2.1 Instrumental variables

IVs are used to indirectly measure the effect of a regressor on an outcome variable, where the regressor serves as a mediator (Newhouse & McClellan, 1998). Figure 2.2a depicts a schematic overview of the relationship between instrumental (I with $i = 1 .., C_i$), mediating (M with $m = 1, \dots, C_m$) and outcome (O with $o = 1, \dots, C_o$) variables.

By the law of total probability, the conditional probability $P(\mathrm{O} \mid \mathrm{I} = i)$ can be written as

$$P(\mathrm{O} \mid \mathrm{I} = i) = \sum_m P(\mathrm{O}, \mathrm{M} = m \mid \mathrm{I} = i) = \sum_m P(\mathrm{O} \mid \mathrm{M} = m, \mathrm{I} = i) P(\mathrm{M} = m \mid \mathrm{I} = i)$$

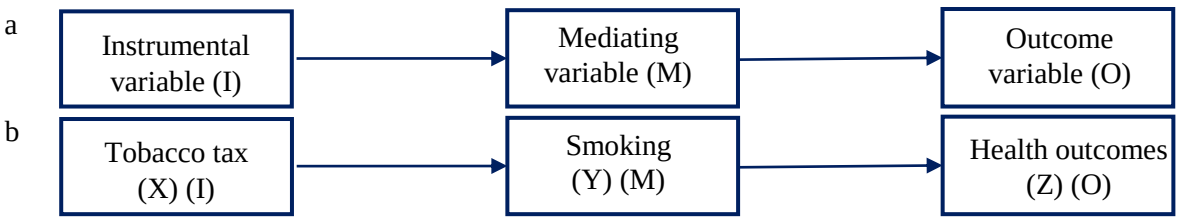
An IV is said to be independent of the outcome variable, given the value of the mediating variable, so under the IVA the term $P(\mathrm{O} \mid \mathrm{M} = m, \mathrm{I} = i)$ reduces to $P(\mathrm{O} \mid \mathrm{M} = m)$, giving

$$P(\mathrm{O} \mid \mathrm{I} = i) = \sum_m P(\mathrm{O} \mid \mathrm{M} = m) P(\mathrm{M} = m \mid \mathrm{I} = i) \qquad (2)$$

Extending this to the working example; suppose the researcher knows from the literature that tobacco tax is an instrumental variable for health outcomes, implying health outcomes to be independent of tobacco tax, given smoking. In this case it holds that $P(\mathrm{H} \mid \mathrm{T} = t) = \sum_s P(\mathrm{H} \mid \mathrm{S} = s) P(\mathrm{S} = s \mid \mathrm{T} = t)$. Figure 2.2b shows a schematic overview of this situation.

Figure 2.2

Schematic overview of the Instrumental Variable mechanism in general (a) and for a specific example (b)



Note: I, M, and O denote the instrumental, mediator and outcome variables. X, Y and Z denote the overlapping (X) and non-overlapping variables (Y, Z)

2.2.2 An instrumental variable approach to statistical matching

In the context of statistical matching the IVA might be more appropriate than the CIA. The advantage of the IVA, as opposed to the CIA, is that no assumptions are made regarding which variables overlap in samples *A* and *B*. Therefore, the overlapping variable can be either the mediating, outcome or instrumental variable. This leads to three separate matching scenario's, outlined in Table 2.1. In scenario's two and three, a set of equations must be solved to estimate the target distribution, which is specified in Appendix A.

The idea of estimating the target distribution using a set of equations can be illustrated using the working example. The instrumental variable, tobacco tax, is overlapping so the researcher is concerned with scenario three. Table 2.2 describes example conditional probabilities the researcher estimates from their data. Using Equation (2), the following four equations can be used to estimate $P(\mathrm{H} = h \mid \mathrm{S} = s)$, $(h = adverse, good\,; s = yes, no)$:

$$
\begin{aligned}
P(\mathrm{H} = adverse \mid \mathrm{T} = high) = \; & P(\mathrm{S} = yes \mid \mathrm{T} = high)P(\mathrm{H} = adverse \mid \mathrm{S} = yes) + \\
& P(\mathrm{S} = no \mid \mathrm{T} = high)P(\mathrm{H} = adverse \mid \mathrm{S} = no) \\
P(\mathrm{H} = good \mid \mathrm{T} = high) = \; & P(\mathrm{S} = yes \mid \mathrm{T} = high)P(\mathrm{H} = good \mid \mathrm{S} = yes) + \\
& P(\mathrm{S} = no \mid \mathrm{T} = high)P(\mathrm{H} = good \mid \mathrm{S} = no) \\
P(\mathrm{H} = adverse \mid \mathrm{T} = low) = \; & P(\mathrm{S} = yes \mid \mathrm{T} = low)P(\mathrm{H} = adverse \mid \mathrm{S} = yes) + \\
& P(\mathrm{S} = no \mid \mathrm{T} = low)P(\mathrm{H} = adverse \mid \mathrm{S} = no) \\
P(\mathrm{H} = good \mid \mathrm{T} = low) = \; & P(\mathrm{S} = yes \mid \mathrm{T} = low)P(\mathrm{H} = good \mid \mathrm{S} = yes) + \\
& P(\mathrm{S} = no \mid \mathrm{T} = low)P(\mathrm{H} = good \mid \mathrm{S} = no)
\end{aligned}
$$

Table 2.1

Statistical matching scenario's that are possible when using the IVA

Scenario	Overlapping variable	Probabilities estimated from data	Target distribution	Direct estimation using Equation (2)
1	Mediator	$P(\mathrm{M} \mid \mathrm{I} = i)$ and $P(\mathrm{O} \mid \mathrm{M} = m)$	$P(\mathrm{O} \mid \mathrm{I} = i)$	Yes
2	Outcome	$P(\mathrm{O} \mid \mathrm{I} = i)$ and $P(\mathrm{O} \mid \mathrm{M} = m)$	$P(\mathrm{M} \mid \mathrm{I} = i)$	No, equations need to be solved (appendix A)
3	Instrumental	$P(\mathrm{O} \mid \mathrm{I} = i)$ and $P(\mathrm{M} \mid \mathrm{I} = i)$	$P(\mathrm{O} \mid \mathrm{M} = m)$	No, equations need to be solved (appendix A)

Note: M, I, and O represent the mediator, instrumental and outcome variables.

Table 2.2

Example estimated conditional probabilities for health outcomes given tobacco tax and smoking given tobacco tax

Sample A	
$P(\mathrm{H} = adverse \mid \mathrm{T} = high)$	.64
$P(\mathrm{H} = good \mid \mathrm{T} = high)$	.36
$P(\mathrm{H} = adverse \mid \mathrm{T} = low)$	.25
$P(\mathrm{H} = good \mid \mathrm{T} = low)$	.75
Sample B	
$P(\mathrm{S} = yes \mid \mathrm{T} = high)$	.80
$P(\mathrm{S} = no \mid \mathrm{T} = high)$	.20
$P(\mathrm{S} = yes \mid \mathrm{T} = low)$	.30
$P(\mathrm{S} = no \mid \mathrm{T} = low)$	.70

Note: health outcomes given tobacco tax probabilities are estimated from sample A, smoking given tobacco tax probabilities are estimated from sample B.

Filling in the values from Table 2.2:

$$.64 = .8 \times P(\mathrm{H} = adverse \mid \mathrm{S} = yes) + .2 \times P(\mathrm{H} = adverse \mid \mathrm{S} = no)$$
$$.36 = .8 \times P(\mathrm{H} = good \mid \mathrm{S} = yes) + .2 \times P(\mathrm{H} = good \mid \mathrm{S} = no)$$
$$.25 = .3 \times P(\mathrm{H} = adverse \mid \mathrm{S} = yes) + .7 \times P(\mathrm{H} = adverse \mid \mathrm{S} = no)$$
$$.75 = .3 \times P(\mathrm{H} = good \mid \mathrm{S} = yes) + .7 \times P(\mathrm{H} = good \mid \mathrm{S} = no)$$

Solving this set of equations gives the conditional distribution of health outcomes and smoking:

$$P(\mathrm{H} = adverse \mid \mathrm{S} = yes) = 0.796$$
$$P(\mathrm{H} = adverse \mid \mathrm{S} = no) = .016$$
$$P(\mathrm{H} = good \mid \mathrm{S} = yes) = .204$$
$$P(\mathrm{H} = good \mid \mathrm{S} = no) = .984.$$

2.2.2a CIA versus IVA

If the overlapping variable (X) is the mediator and the non-overlapping variables (Y and Z) are the instrumental and outcome variables, the expression under the IVA for the conditional distribution

$P(\mathrm{O} \mid \mathrm{I} = i)$ (Equation 2) is equivalent to the expression for the joint distribution $P(\mathrm{Z}, \mathrm{Y})$ (Equation 1) under the CIA with X being the common variable. This can be shown by assigning variable X the role of mediator, variable Y the role of instrumental and variable Z the role of outcome variable (scenario 1). It is then possible to rewrite $P(\mathrm{Z}, \mathrm{Y})$ as follows

$$P(\mathrm{Z}, \mathrm{Y}) = P(\mathrm{Z} = z \mid \mathrm{Y} = y)P(\mathrm{Y} = y) =$$
$$\sum_x P(\mathrm{Z} = z \mid \mathrm{Y} = y, \mathrm{X} = x)P(\mathrm{X} = x \mid \mathrm{Y} = y)P(\mathrm{Y} = y) =$$
$$\sum_x P(\mathrm{Z} = z \mid \mathrm{X} = x)P(\mathrm{X} = x \mid \mathrm{Y} = y)P(\mathrm{Y} = y) =$$
$$\sum_x P(\mathrm{Z} = z \mid \mathrm{X} = x)P(\mathrm{Y} = y \mid \mathrm{X} = x)P(\mathrm{X} = x)$$

In the third line the IVA is applied and in the final line the expression under the CIA is given. The last step emerges from $P(\mathrm{X} = x \mid \mathrm{Y} = y)P(\mathrm{Y} = y) = P(\mathrm{Y} = y, \mathrm{X} = x) = P(\mathrm{Y} = y \mid \mathrm{X} = x)P(\mathrm{X} = x)$.

In general, the procedures for estimating the unknown distribution $P(\mathrm{Z}, \mathrm{Y})$ under the IVA and the CIA are the same. The crucial difference is that the CIA always conditions on the overlapping variable X, whereas the IVA always conditions on the mediating variable, irrespective of which variable is commonly observed in the data. This difference between the two assumptions can be illustrated using the working example of tobacco tax, smoking and health outcomes. Under the IVA, the relationship between smoking and health outcomes would be estimated using

$$P(\mathrm{H} \mid \mathrm{T} = t) = \sum_s P(\mathrm{H} \mid \mathrm{S} = s)P(\mathrm{S} = \mathrm{s} \mid \mathrm{T} = t)$$

While under the CIA the same relationship would be estimated using

$$P(\mathrm{S}, \mathrm{H}) = \sum_t P(\mathrm{S} \mid \mathrm{T} = t)P(\mathrm{H} \mid \mathrm{T} = t)P(\mathrm{T} = t)$$

This last expression is evidently incorrect since it is known that tobacco tax is independent of health outcomes given smoking, rather than smoking being independent of health outcomes given tobacco tax.

2.3 Aim of the project

The IVA approach to statistical matching could be a viable alternative to relying on the CIA, therefore the current project's aims are two-fold:

1) Evaluate the IVA approach to statistical matching by assessing its bias, root mean squared error (RMSE) and standard deviation when the IVA is violated to various degrees
2) Compare the performances of the IVA approach and CIA approach to statistical matching by comparing the biases, RMSEs and standard deviations of both methods

Chapter 3

Methods

3.1 Estimators

The goal is to estimate the joint distribution $P(\mathrm{Z},\mathrm{Y})$ using the information available in the two disjoint samples. To be able to contrast the CIA and the IVA, this was done using both methods. Under the CIA, $P(\mathrm{Z},\mathrm{Y})$ can be estimated using Equation (1) where the probabilities on the right hand side are estimated from the data, formally

$$\hat{P}(\mathrm{Z},\mathrm{Y}) = \sum_{x} \hat{P}(\mathrm{Z} \mid \mathrm{X} = x)^{(B)}\, \hat{P}(\mathrm{Y} \mid \mathrm{X} = x)^{(A)} \hat{P}(\mathrm{X} = x)^{(AB)} \tag{3}$$

where X denotes the overlapping variable, $\hat{P}^{(A)}$ denotes probabilities estimated from sample A, $\hat{P}^{(B)}$ denotes probabilities estimated from sample B and $\hat{P}^{(AB)}$ denotes probabilities estimated from both samples. $P(\mathrm{X} = x)$ was estimated using both samples for a more precise estimate (see section 2.1 of D'Orazio et al. (2006)).

Under the IVA, the conditional probability of the outcome and instrumental variables can be estimated using Equation (2) where the probabilities are estimated from the data, formally

$$\hat{P}(\mathrm{O} \mid \mathrm{I} = i) = \sum_{m} \hat{P}(\mathrm{O} \mid \mathrm{M} = m)\, \hat{P}(\mathrm{M} = m \mid \mathrm{I} = i) \tag{4}$$

where M denotes the mediating variable, I denotes the instrumental variable and O denotes the outcome variable. Which term in Equation (4) can be denoted $\hat{P}(\mathrm{Z} \mid \mathrm{Y} = y)$ and which probabilities can be estimated from what dataset (A or B) depends on which of the three variables overlaps in the data (see Table 2.1).

One of the research aims was to contrast the CIA and IVA approaches, so it was useful to convert the estimated conditional probability under the IVA, $\hat{P}(\mathrm{Z} \mid \mathrm{Y} = y)$, to an estimated joint probability $\hat{P}(\mathrm{Z},\mathrm{Y})$. This is possible by multiplying the estimated conditional probability $\hat{P}(\mathrm{Z} \mid \mathrm{Y} = y)$ by the marginal probability $\hat{P}(\mathrm{Y})$. However, since $\hat{P}(\mathrm{Y})$ is a non-overlapping variable it can only be estimated from one of the datasets. A more accurate estimate might be $\hat{P}_{comb}(\mathrm{Y})$, which can be estimated using data from both datasets

$$\hat{P}_{comb}(\mathrm{Y}) = \sum_{x} \hat{P}(\mathrm{Y} \mid \mathrm{X} = x)^{(A)} \hat{P}(\mathrm{X} = x)^{(AB)}$$

This measure $\hat{P}_{comb}(\mathrm{Y})$ was then used to get an estimate for the joint probability $\hat{P}(\mathrm{Z}, \mathrm{Y})$

$$\hat{P}(\mathrm{Z}, \mathrm{Y}) = \hat{P}(\mathrm{Z} \mid \mathrm{Y} = y)\hat{P}_{comb}(\mathrm{Y})$$

3.2 Simulation study

To assess how the IVA performs and how it compares to using the CIA in statistical matching, a simulation study was performed. This section discusses the data generation procedure and the design of the study. All analyses were performed using R Statistical Software (v4.2.3; R Core Team 2023).

3.2.1 Data generation

Populations of the three variables – mediator, outcome and instrumental – were generated. Each variable in the simulation study had two categories.

3.2.1a Populations from which to sample

The first research aim was to evaluate the bias and variance of the statistical matching procedure when the IVA is violated. The IVA is based on two premises: 1) the instrumental and outcome variables are independent given the mediating variable, and 2) the association between the instrumental and mediating variables is substantial (Newhouse & McClellan, 2010). When the former is violated this is expected to result in a biased estimate of $P(\mathrm{Z}, \mathrm{Y})$. A weak association between the instrumental and mediating variables is not strictly a violation and will probably not result in large biases, however the use of the IVA is expected to work less well under such a condition. In that sense, a population where the association is strong would be more favorable for the use of the IVA approach to statistical matching. In the rest of this thesis a violation with respect to the association between instrumental and outcome variables will be called a type A violation. An unfavorable population where the association between instrumental and mediating variables is weak, will be called a type B violation.

To quantify the association between the different variables, odds ratios (OR) were used. Chen et al. (2010) linked different OR levels to Cohen's *d* measure of effect size, which were used as a reference. In the current project the ORs used were 1 (independence), 2 (weak association), 4 (moderate association) and 7 (strong association). Table 3.1 shows the population distributions that were used to sample from in the simulation study.

The first population shown in Table 3.1 represents the "ideal" situation in which the IVA holds perfectly. The results are expected to be unbiased. In the second population there is only a type B violation of the IVA. The matching procedure is expected to be robust against this type of violation

because the main assumption on which the calculations are based, conditional independence of the outcome and the instrumental variables, still holds. In the third population, the IVA is slightly violated (type A) and in the fourth population the IVA is severely violated (type A). The statistical matching procedure is expected to be biased in these situations. The last population includes both types of violation, where the procedure is expected to be biased.

Violation of the CIA

In order to fully be able to assess the difference between the IVA and CIA approaches it is useful to quantify the degree of violation of the CIA in the used populations as well. Under the CIA the odds ratio between the non-overlapping variables Z and Y, given the value of overlapping variable X is assumed to be 1. Table 3.1 includes these odds ratios in the population for the various matching scenarios. These odds ratios were obtained using the Fisher's Exact Test function from the *stats* package in R (v4.2.3; R Core Team 2023).

3.2.1b Data generation procedure

Populations were generated using the settings outlined in Table 3.1. In the first step of generating the data, the marginal distributions, $P(\mathrm{M}=1)$, $P(\mathrm{O}=1)$ and $P(\mathrm{I}=1)$, were specified. These probabilities were randomly chosen from a uniform distribution between .25 and .75. Additionally, the association between the instrumental and mediator, $OR(\mathrm{I},\mathrm{M})$, and the mediator and outcome variables, $OR(\mathrm{O},\mathrm{M})$, were defined. In the next step, the joint distributions $P(\mathrm{I},\mathrm{M})$ and $P(\mathrm{O},\mathrm{M})$ were calculated by solving for the defined odds ratios (see intermezzo).

Intermezzo: Solving for an odds ratio

In order to set the association between the variables in the population, odds ratios were used. The formula for the odds ratio is $OR = \frac{ad}{bc}$ for the contingency table:

		Y		
		1	2	
X	1	a	b	x_1
	2	c	d	x_2
		y_1	y_2	

where x_1 and x_2 are the row totals for variable X and y_1 and y_2 are the column totals for variable Y. When the row and column totals are known, it is possible to rewrite the odds ratio equation in terms of one of the cells in the contingency table, for instance in terms of a:

$$OR = \frac{a * (x_2 - (y_1 - a))}{(x_1 - a) * (y_1 - a)}$$

where b is substituted by $x_1 - a$, c is substituted by $y_1 - a$ and d is substituted by $x_2 - (y_1 - a)$. The desired odds ratio and the row and column totals are known, so it is possible to solve for a. Then a can be used to derive b, c and d. The result is the full contingency table of variables X and Y for a given association as measured by the odds ratio.

3.2.1b Data generation procedure (continued)

The row and column totals of the variables were obtained by multiplying the population size (1,000,000 in each population) by the marginal distributions of the variables. The resulting contingency table contained counts, which were transformed to proportions again by dividing by the population size. After obtaining joint distributions $P(\mathrm{I}, \mathrm{M})$ and $P(\mathrm{O}, \mathrm{M})$, the conditional distribution $P(\mathrm{O} \mid \mathrm{M} = m)$ was calculated using the property $P(\mathrm{O} \mid \mathrm{M} = m) = \frac{P(\mathrm{O},\mathrm{M})}{P(\mathrm{M})}$.

Table 3.1

Population settings in the simulation study evaluating the IVA approach to statistical matching and to what degree the IVA and CIA are violated.

General settings	N	$P(\mathrm{M} = 1)$	$P(\mathrm{O} = 1)$	$P(\mathrm{I} = 1)$	$OddsRatio(\mathrm{O}, \mathrm{M})$
	1 000 000	.615	.652	.429	4
Specific settings	Population – IVA violation				
	1 – no violation	2 – type B violation	3 – slight type A violation	4 – severe type A violation	5 – type B and slight type A violation
$\mathrm{OddsRatio}(\mathrm{I}, \mathrm{M})$	7	2	7	7	2
$\mathrm{OddsRatio}(\mathrm{O}, \mathrm{I} \mid \mathrm{M} = m)$	1	1	2	4	2
Overlapping variable (scenario)	$\mathrm{OddsRatio}(\mathrm{Z}, \mathrm{Y} \mid \mathrm{X} = x)$				
X = M (scenario 1)	1.0	1.0	2.0*	4.0**	2.0*
X = O (scenario 2)	7.0**	2.0*	5.9**	5.4**	1.6
X = I (scenario 3)	4.0**	4.0**	3.1*	2.5*	3.7*

Note: IVA denotes the Instrumental Variable Assumption, CIA denotes the Conditional Independence Assumption. M ,O, and I denote the mediator, outcome and instrumental variables. X denotes the overlapping variable and Z,Y denote the non-overlapping variables. * indicates a slight violation of the CIA. ** indicates a severe violation of the CIA. A type A violation of the IVA happens when the outcome and instrumental are not independent, a type B violation denotes the situation where the association between the instrumental and the mediator is weak.

The final step was calculating $P(\mathrm{I}, \mathrm{M}, \mathrm{O})$. In general it holds that

$$P(\mathrm{I}, \mathrm{M}, \mathrm{O}) = P(\mathrm{O} \mid \mathrm{M} = m\,, \mathrm{I} = i)P(\mathrm{I}, \mathrm{M})$$

Since under the IVA the instrumental and the outcome variables are independent the above equation can be reduced to

$$P_{IVA}(\mathrm{I}\,, \mathrm{M}, \mathrm{O}) = P(O \mid \mathrm{M} = m)\; P(\mathrm{I}, \mathrm{M})$$

For populations where the odds ratio between the instrumental and outcome variables was equal to 1, $P_{IVA}(\mathrm{I}, \mathrm{M}, \mathrm{O})$ could be used for sampling.

To generate a population with a type A violation, some additional steps were required. Under the IVA, independence exists conditional on the mediator. In order to use the odds ratio to capture this independence in a population, the joint probability needs to be split according to the levels of the mediator. Then it is possible to define an odds ratio (unequal to 1) and calculate a joint distribution for each of the levels of the mediator. Specifically, $P_{IVA}(\mathrm{I}, \mathrm{M}, \mathrm{O})$ was adjusted to get a joint distribution with a type A violation: $P_{notIVA}(\mathrm{I}, \mathrm{M}, \mathrm{O})$.

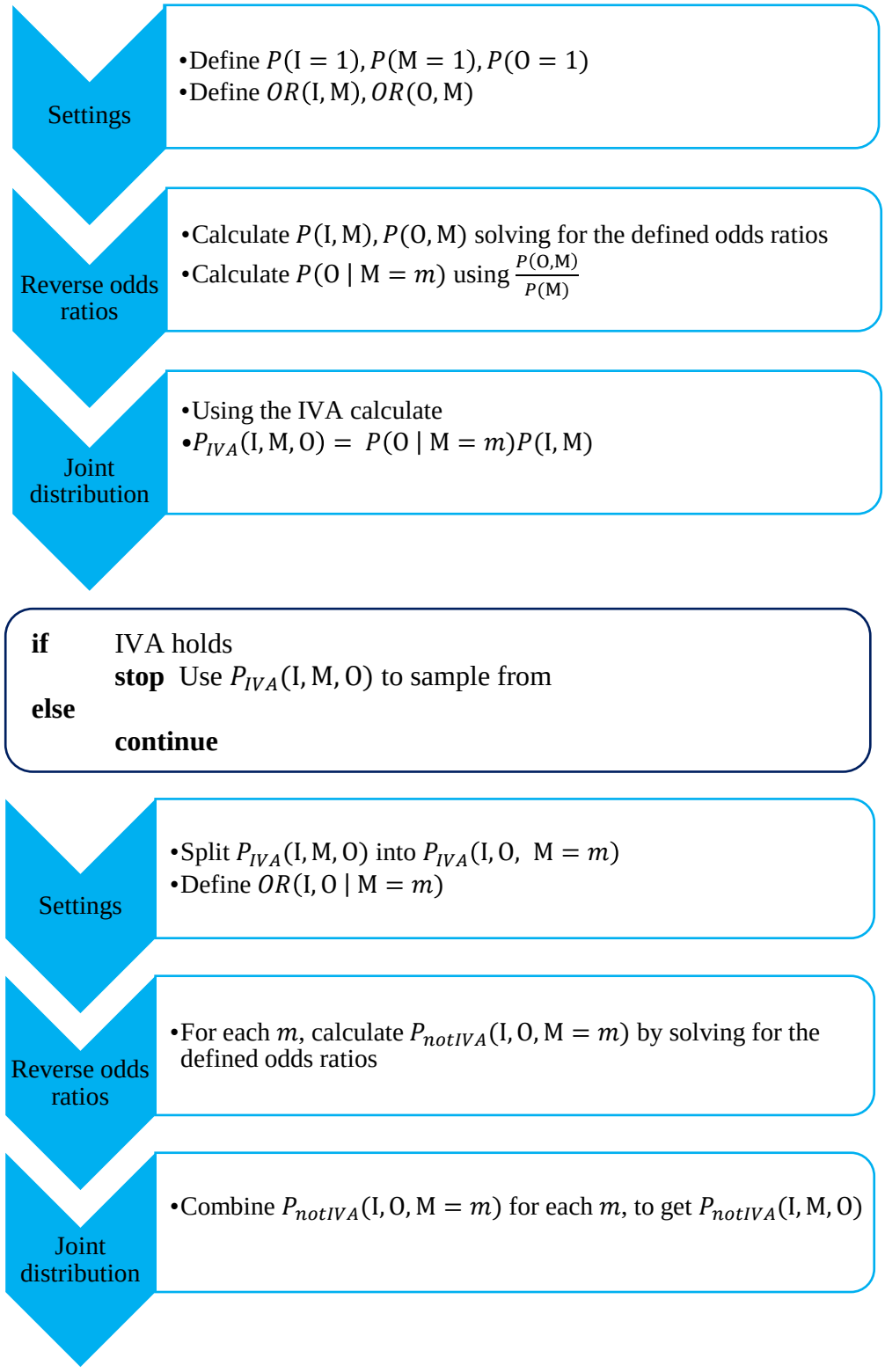
First, $P_{IVA}(\mathrm{I}, \mathrm{M}, \mathrm{O})$ was split according to the two levels of M; $P_{IVA}(\mathrm{I}, \mathrm{O}, \mathrm{M} = m)$ $(m = 1{,}2)$. The odds ratio $OR(\mathrm{I}, \mathrm{O} \mid \mathrm{M} = m)$ was specified, which was the same for each level of M. The next step was adjusting $P_{IVA}(\mathrm{I}, \mathrm{O},\ \mathrm{M} = m)$ to get $P_{notIVA}(\mathrm{I}, \mathrm{O}, \mathrm{M} = m)$ by solving for the defined odds ratio using $OR(\mathrm{I}, \mathrm{O} \mid \mathrm{M} = m)$. Finally, $P_{notIVA}(\mathrm{I}, \mathrm{M}, \mathrm{O})$ was obtained by combining $P_{notIVA}(\mathrm{I}, \mathrm{O},\ \mathrm{M} = m)$ for each level of M. A schematic overview of the procedure is outlined in Figure 3.1.

3.2.2 Design of the study

A Monte Carlo simulation study was performed with $R = 500$ replications for each population (see Appendix B for a justification of the number of simulations). Within each replication r $(r = 1, \ldots, R)$, two samples of $n = 2000$ were drawn for each of the three matching scenarios, resulting in 500×3 sample pairs. All variables were drawn from the population for each sample pair, but estimation was performed as if one variable was missing in each sample. For instance, in the first scenario, where the mediating variable is overlapping, both samples ($n = 2000$) were drawn with all variables present but estimation proceeded as if one sample had a missing outcome and the other a missing instrumental variable. The same was done for the other scenarios in the same replication. For each of these sample pairs, the joint distribution of the non-overlapping variables was estimated under the IVA and the CIA. In situations where the assumptions are violated it can happen that the estimated probabilities are negative or larger than one. This is evidently impossible so if a negative probability was estimated it was coerced to zero, when a probability was estimated to be larger than one it was coerced to one. This was done for both the estimated conditional probabilities and the estimated joint probabilities.

Figure 3.1

Schematic overview of the data generation procedure used to generate populations under the IVA and for various violations of the IVA



Note: M denotes the mediator variable, O the outcome variable and I the instrumental variable. In the current project, $OR(\mathrm{I},\mathrm{O} \mid \mathrm{M}=m)$ was the same for each of the two levels m ($m = 1,2$) of M

3.2.3 Simulated parameters

For all R sample pairs, the probability $\hat{p}_{zy}$ for all four category combinations of the non-overlapping variables is estimated. The population mean is estimated by $\hat{\bar{p}}_{zy} = \frac{1}{R}\sum_r \hat{p}_{zyr}$ and the standard deviation by $\sqrt{Var(\hat{p}_{zy})} = \sqrt{\frac{1}{R-1}\sum_r(\hat{p}_{zyr} - \hat{\bar{p}}_{zy})^2}$ for each category combination. Table 3.2 shows a simplified example of how the statistics are calculated. Confidence intervals around the estimates were derived by ordering all R $\hat{p}_{zy}$'s and taking the 2.5th and 97.5th percentiles as lower and upper bounds respectively.

3.2.4 Quality of the simulated parameter estimators

The accuracy of an estimator can be assessed using the bias and the Root Mean Squared Error (RMSE). The bias of an estimator $\hat{\theta}$ is quantified as the expected difference between the estimated and the true value of a parameter

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

where θ is the population parameter. The RMSE of an estimator $\hat{\theta}$ is calculated as

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

If we take $\theta = p_{zy}$ and $\hat{\theta} = \hat{p}_{zy}$, the bias and RMSE can be approximated for R sample pairs from the population according to the following recipe:

1. Generate two disjoint samples with one overlapping variable from a given distribution.
2. Estimate the four p_{zy}'s, for each sample pair r.
3. Calculate $\hat{p}_{zyr} - p_{zy}$ and $(\hat{p}_{zyr} - p_{zy})^2$ for each of the four estimates, for each sample pair r.
4. Estimate the bias by $Bias(\hat{p}_{zy}) = \frac{1}{R}\sum_r(\hat{p}_{zyr} - p_{zy})$ and the RMSE by $RMSE(\hat{p}_{zy}) = \sqrt{\frac{1}{R}\sum_r(\hat{p}_{zyr} - p_{zy})^2}$

To get one measure of the bias and RMSE (instead of four for each category combination), the RMSE and absolute value of the bias were averaged over the category combinations. The bias and RMSE were used to assess the accuracy of the statistical matching procedures using the IVA or the CIA.

In principle it is sufficient to evaluate the quality of the statistical matching by using these bias, RMSE and variance estimates using a large set of simulated, drawn samples. Still, it might be interesting

Table 3.2

Simplified example of how the true parameters would be estimated from three simulations

Y	Z	$r = 1$	$r = 2$	$r = 3$	Parameters	
		$\hat{p}_{zy}$	$\hat{p}_{zy}$	$\hat{p}_{zy}$	$\hat{\bar{p}}_{zy}$	$\sqrt{V(\hat{p}_{zy})}$
1	1	.30	.25	.31	.287	.032
1	2	.03	.04	.04	.037	.006
2	1	.09	.11	.07	.090	.020
2	2	.58	.60	.58	.587	.012

Note: r denotes the simulation replication, $\hat{p}_{zy}$ denotes the estimated probability for a given category combination, $\hat{\bar{p}}_{zy}$ the estimated population mean for a given category combination and $\sqrt{V(\hat{p}_{zy})}$ the standard deviation for a given category combination.

to study how close bootstrap estimates of the bias, RMSE and variance based on a single sample r come to the estimates based on a large set of samples.

3.2.5 Bootstrap

Bootstrap analyses were performed in order to study how close bootstrap estimates of the bias, RMSE and variance of a single sample come to estimates based on a large set of samples. Samples from $S = 10$ simulated sample pairs were saved and used for bootstrapping. For each of these sample pairs, $B = 100$ bootstrap samples were drawn and the mean, standard deviation, confidence interval, bias and RMSE were derived according to the following recipe:

For each sample s, with $s = 1, \dots, S$:

1. Let b denote a bootstrap sample, with $b = 1, \dots, B$. Further let $\hat{\bar{p}}_{zys}^{(B)} = \frac{1}{B}\sum_b \hat{p}_{zysb}$ denote the mean over B bootstraps of sample s. Let p_{zys} be the true proportion in cell z, y for sample s.
2. Estimate the standard deviation by

$$\sqrt{\widehat{Var}(\hat{p}_{zys})} = \sqrt{\frac{1}{B-1}\sum_b \left(\hat{p}_{zysb} - \hat{\bar{p}}_{zys}^{(B)}\right)^2}$$

   and average over category combinations.
3. the bias by

$$\widehat{Bias}(\hat{p}_{zys}) = \frac{1}{B}\sum_b \hat{p}_{zysb} - p_{zys}$$

and average the absolute value over category combinations.

4. the RMSE by

$$\widehat{RMSE}(\hat{p}_{zys}) = \sqrt{\frac{1}{B}\sum_{b}(\hat{p}_{zysb} - p_{zys})^2}$$

and average over category combinations

5. derive the 95% confidence interval by ordering all $\hat{p}_{zysb}$ and taking the 2.5th and 97.5th percentiles as lower and upper limits, respectively

This procedure resulted in $S = 10$ estimates of the bootstrap means, standard deviations, bias and RMSE, which were averaged. These bootstrap estimates should be close to the estimates based on the R simulated samples. Appendix B includes a justification for the number of bootstraps.

3.2.6 Assessing uncertainty

In statistical matching it is possible to use the Fréchet property to provide absolute limits on the range of possible probabilities within a joint distribution. This will result in lower and upper bounds that can be used to assess the uncertainty of a statistical matching procedure (D'Orazio, 2019). The Fréchet property can be used to identify the following interval:

$$\max(0; P(\mathrm{Y} = y) + P(\mathrm{Z} = z) - 1) \leq P(\mathrm{Y} = y, \mathrm{Z} = z) \leq \min\big(P(\mathrm{Y} = y); P(\mathrm{Z} = z)\big)$$

$$y = 1, \dots, Y; z = 1, \dots, Z$$

where y and z are the categories of non-overlapping variables Y and Z, respectively.

A numerical example: say variable Y takes value $y = 1$ with a probability of .429 and variable Z takes value $z = 1$ with a probability of .652. The joint probability $p_{1,1}$ would then be within the limits

$$\max(0; .429 + .652 - 1) \leq P(\mathrm{Y} = 1, \mathrm{Z} = 1) \leq \min(.429; .652), \text{ i.e.}$$
$$.081 \leq P(\mathrm{Y} = 1, \mathrm{Z} = 1) \leq .429$$

In other words, the joint probability $P(\mathrm{Y} = 1, \mathrm{Z} = 1)$ is bounded from below by .081 and bounded from above by .429.

Calculating conditional bounds, i.e. conditioning variables Y and Z on another variable, has been shown to result in tighter intervals (Conti et al., 2012). Formally, it entails taking the expectation of the conditional bounds

$$\underline{P(\mathrm{Y} = y, \mathrm{Z} = z)} \leq P(\mathrm{Y} = y, Z = z) \leq \overline{P(\mathrm{Y} = y, \mathrm{Z} = z)}$$

where

$$\underline{P(\mathrm{Y}=y,\mathrm{Z}=z)} = \sum_{x} \max(0, P(\mathrm{Y}=y \mid \mathrm{X}=x) + P(\mathrm{Z}=z \mid \mathrm{X}=x) - 1)\, P(\mathrm{X}=x)$$

$$\overline{P(\mathrm{Y}=y,\mathrm{Z}=z)} = \sum_{x} \min(P(\mathrm{Y}=y \mid \mathrm{X}=x); P(\mathrm{Z}=z \mid \mathrm{X}=x))\, P(\mathrm{X}=x)$$

with X representing the overlapping variable(s) in the data used for matching and x being the categories of X. When there are several overlapping variables, a single variable X can be constructed by crossing the overlapping variables (D'Orazio, 2019).

In this project, marginal and conditional Fréchet bounds were calculated for all sample pairs, in each situation, for each population. Conditional bounds were used to assess the magnitude of the bias and to assess the uncertainty around the estimates, in the form of confidence intervals.

3.2.7 IVA versus CIA

To assess whether the IVA is a viable alternative to using the CIA in statistical matching scenarios, all analyses were done using both approaches. The bias and RMSE of both approaches were compared to assess accuracy of both methods. Confidence intervals (and Fréchet bounds) and standard deviations were compared to assess uncertainty and precision.

Chapter 4

Results

4.1 Evaluation IVA

To evaluate whether the IVA is a suitable assumption to use in a statistical matching context, the RMSE, bias and standard deviation calculated from the simulation were assessed. These measures were compared over the three matching scenarios for each of the five populations where the IVA is violated to varying degrees. In Figure 4.1 the bias, RMSE and standard deviations for all situations are shown. In this section the notation $Bias_{kj}$, $RMSE_{kj}$ and SD_{kj} is used where $k = 1, \dots 5$ and $j = 1, \dots, 3$ indicate the five populations and three matching scenarios (see Table 3.1). A measure of uncertainty is the 95% confidence interval around the estimates for the category combinations. These intervals are assessed in contrast to the conditional Fréchet bounds. For a full overview of all point estimates, standard deviations, confidence intervals, bias estimates and RMSEs, the reader is referred to Appendix C.

4.1.1 Bias

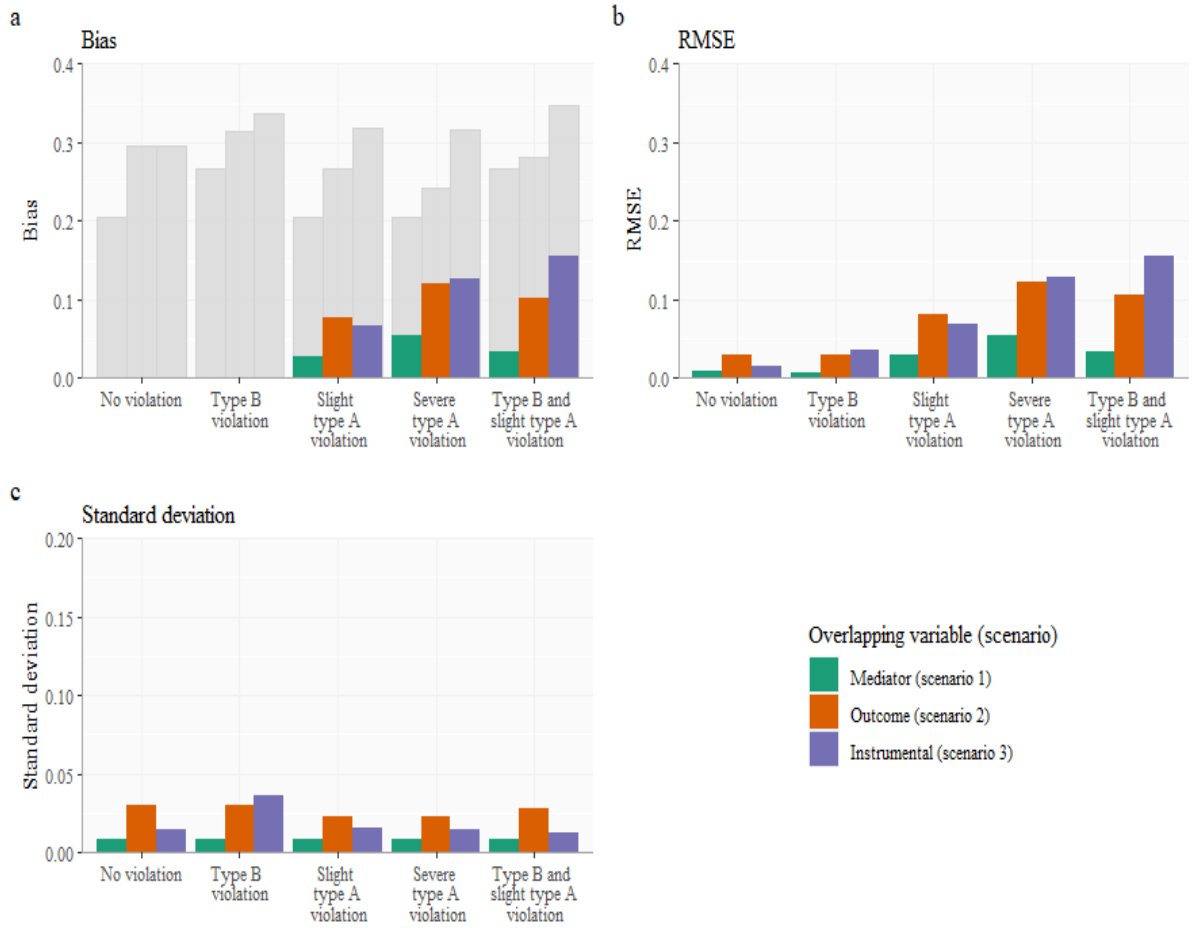
The bias was used to assess the systemic error in the estimation in the various simulated situations. Bias for all populations and scenarios can be found in Figure 4.1a. The grey bars in Figure 4.1a represent the width of the conditional Fréchet bounds for that specific situation. These bounds give an indication of the magnitude of the realized bias as they represent the maximum amount of bias possible. The biases will be compared to the conditional Fréchet bounds by reporting the proportion of the bounds that is covered by the bias.

In general, the bias increases with more violation of the IVA except when there is solely a type B violation, where the bias is almost non-existent for all scenarios. When the IVA exactly holds or when there is solely a type B violation, the IVA approach to statistical matching is unbiased. Overall, the bias is smallest when the mediator is the overlapping variable (scenario 1).

With a slight type A violation (population 3), the bias for scenario one is not very large, $Bias_{31} = .0278$ taking up 13.5% of the Fréchet bound. A more severe type A violation (population 4) results in a larger but still moderate bias, $Bias_{41} = .0537$ covering 26.1% of the Fréchet bound. For both degrees of type A violation, the bias for scenarios two and three is comparable in absolute sense, $Bias_{32} = .0767$ versus $Bias_{33} = .0664$; $Bias_{42} = .1198$ versus $Bias_{43} = .1272$. However, if the bias is taken relative to the conditional Fréchet bound, it is larger when the outcome variable overlaps,

Figure 4.1

Simulated bias, RMSE and standard deviation of five populations using the instrumental variable approach to statistical matching



Note: Grey areas in plot *a* indicate conditional Fréchet bounds for that specific scenario.

where the bias covers 28.7% and 49.7% of the bound. When the instrumental variable overlaps the bias covers 20.9% and 40.3% of the bound.

In the case of both a type A and type B violation (population 5), the bias for scenario one, $Bias_{51} = .0328$, in absolute sense is larger compared to only a type A violation (population 3) (where both type A violations were of the same magnitude, see Table 3.1). Taking the bias relative to the Fréchet bound indicates it to be smaller, covering 12.3% of the bound. For the other two scenarios, the biases are larger when both a type A and type B violation occur (population 5) relative to only a type A violation occurring (population 3). This difference was observed in both the absolute and relative sense. The highest absolute bias observed overall is found in scenario three, $Bias_{53} = .1543$ which covers 44.3% of the Fréchet bound.

4.1.2 RMSE

The RMSE was used to assess the accuracy of the estimation in the various simulated situations. Figure 4.1b shows the RMSE for each population and matching scenario. In general, the estimation becomes less precise as the IVA is violated, particularly when there is a type A violation. When the mediator overlaps accuracy is generally highest.

When the IVA is not violated at all, the RMSE is highest for scenario two, $RMSE_{12} = .030$. Overall, when the IVA is not violated, accuracy of the statistical matching is high for all scenarios. When there is solely a type B violation (population 2), the RMSE for scenarios one and two is comparable to the situation where the IVA is not violated. The RMSE for scenario three doubles in magnitude, $RMSE_{23} = .036$ versus $RMSE_{13} = .015$. Taken together, the overall accuracy for a type B violation is high for all scenarios.

When there is solely a type A violation, the RMSE for scenarios two and three are comparable. A slight violation (population 3) results in a higher RMSE for scenario two compared to scenario three, $RMSE_{32} = .081$ versus $RMSE_{33} = .068$. A severe violation (population 4) results in a slightly higher RMSE for scenario three compared to two, $RMSE_{43} = .128$ versus $RMSE_{42} = .122$. Overall, a severe type A violation results in low accuracy of the estimation where a slight violation results in a moderate accuracy of the estimation. When both types of violation are present (population 5) the RMSE is higher compared to only a slight type A violation, for all scenarios. The highest RMSE is observed in scenario three, $RMSE_{53} = .155$. Overall accuracy of the estimation for both types of violation is low.

4.1.3 Precision and uncertainty

4.1.3a Standard deviation

The standard deviation was used as a measure of precision of the estimations in the various simulated situations. In Figure 4.1c the standard deviations for each population and scenario are shown. In general, the estimates are quite precise as none of the measures exceed .036. For scenario one (mediator as overlapping variable), precision is highest and constant over populations with all standard deviations hovering around .008.

Scenario two has the lowest precision compared to the other scenarios, except for the situation with two types of violation. When there is a type A violation, precision increases. Standard deviations are constant for no violation ($SD_{12} = .030$), type B violation ($SD_{22} = .030$) and both violations ($SD_{52} = .028$). Precision of the estimates for scenario three are constant for no violation ($SD_{13} = .015$), type A violation (slight: $SD_{33} = .015$; severe: $SD_{43} = .015$) and both types of violation ($SD_{53} = .012$). The exception is the precision in the case of only a type B violation (population 2), which is also the overall highest standard deviation with $SD_{23} = .036$.

4.1.3b Confidence intervals

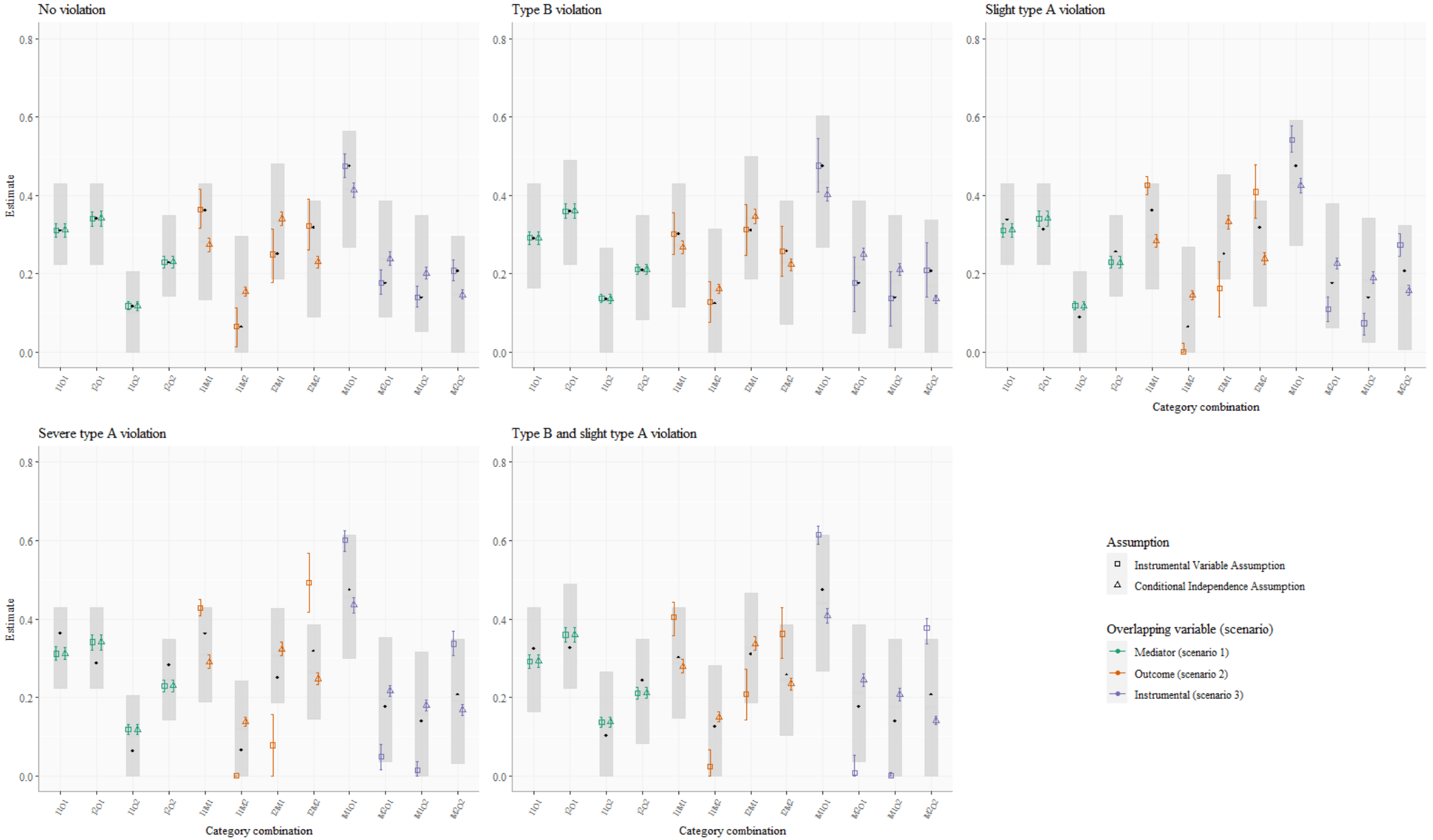
95% confidence intervals were used to assess uncertainty around the estimates. Figure 4.2 shows the true population value and the estimates and their confidence intervals. For the populations where the IVA is not violated or only a type B violation is present, the population value lies within the confidence interval (as expected since the bias was virtually zero). For populations where the IVA is violated, the true value often lies outside of the confidence intervals. Confidence intervals for scenario one are narrow and never fall outside the conditional Fréchet bounds. For scenario two and three the confidence intervals are generally much wider and also differ per category combination. In populations where a slight or severe type A violation of the IVA is present, the confidence intervals are partly or fully outside of the Fréchet bounds.

4.1.4 Bootstrap

To assess how close bootstrap estimation comes to simulation results in the IV approach to statistical matching, the RMSE, bias and standard deviation bootstrap estimates were compared to the simulation results. Figure 4.3 displays the results for all three measures side by side. Overall, no large deviations were observed between the bootstrap and simulation estimates.

Figure 4.2

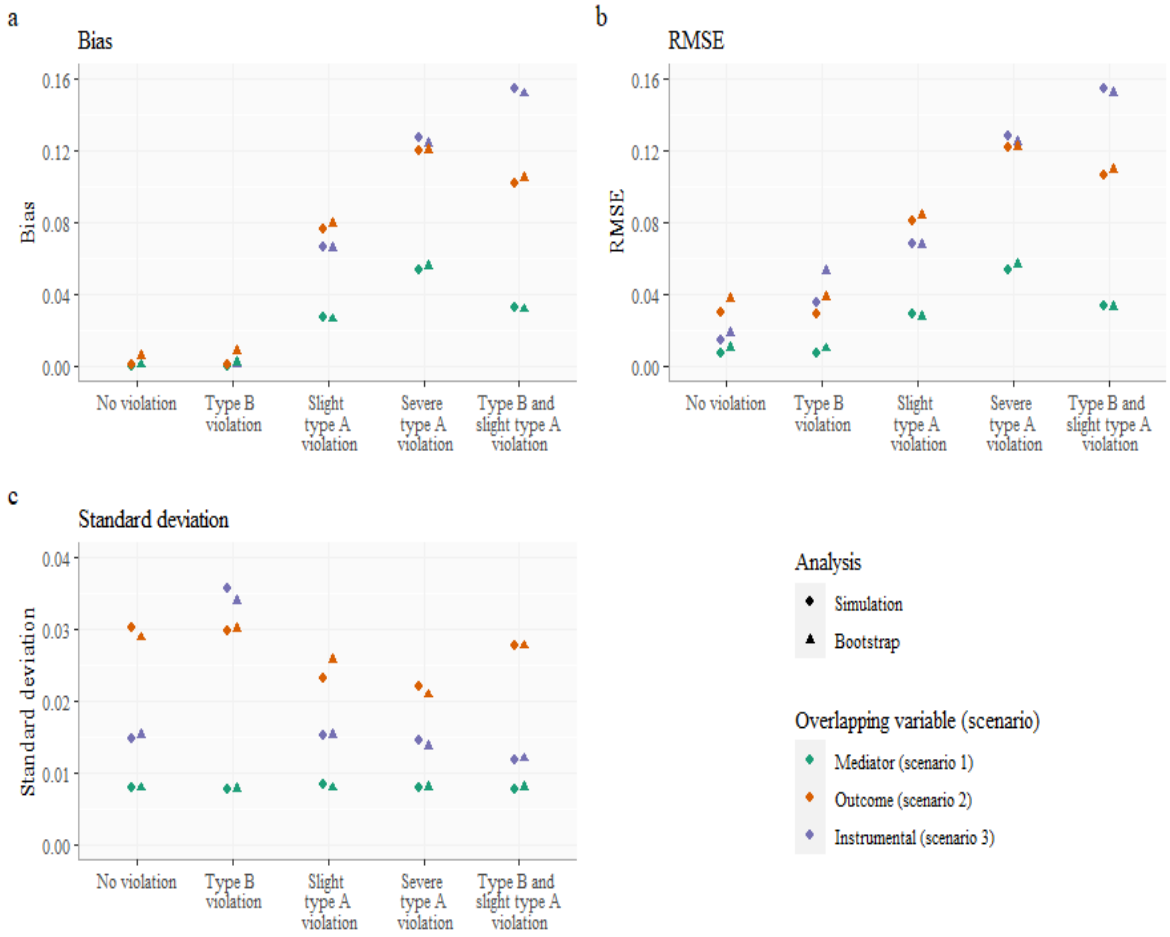
Simulated estimated probabilities and their 95% confidence intervals for each category combination for all populations, each method and scenario



Note: A type B violation indicates the IVA to be violated with respect to the relationship between the instrumental and mediator variable, a type A violation indicates the IVA to be violated with respect to the relationship between the instrumental and outcome variables. Along the x-axes the category combinations can be found, here *I* stands for the instrumental variable, *O* for the outcome variable and *M* for the mediating variable. Numbers indicate the value of that variable, for instance if the category combination is *I2O1*, the instrumental variable has value 2 and the outcome variable has value 1. Grey areas indicate the conditional Fréchet bounds for that specific category combination. Black diamonds demarcate the population value for that specific category combination.

Figure 4.3

Simulated and bootstrapped bias, RMSE and standard deviation of five populations using the instrumental variable approach to statistical matching



Note: A type B violation indicates the IVA to be violated with respect to the relationship between the instrumental and mediator variable, a type A violation indicates the IVA to be violated with respect to the relationship between the instrumental and outcome variables.

4.2 IVA versus CIA

To evaluate whether the IVA is a viable alternative to the CIA, the estimation results for both methods are compared with regards to bias, RMSE, standard deviations and confidence intervals.

4.2.1 Bias

The bias was used to compare the systematic error between the IVA and CIA approaches to statistical matching. In Figure 4.4a both methods' bias is contrasted. For scenario one, where the mediator is the overlapping variable, all values align because in that situation both approaches give the exact same answer.

Table 4.1

Violation of the CIA when the IVA is violated to different degrees and the observed bias and RMSE, for two statistical matching scenarios.

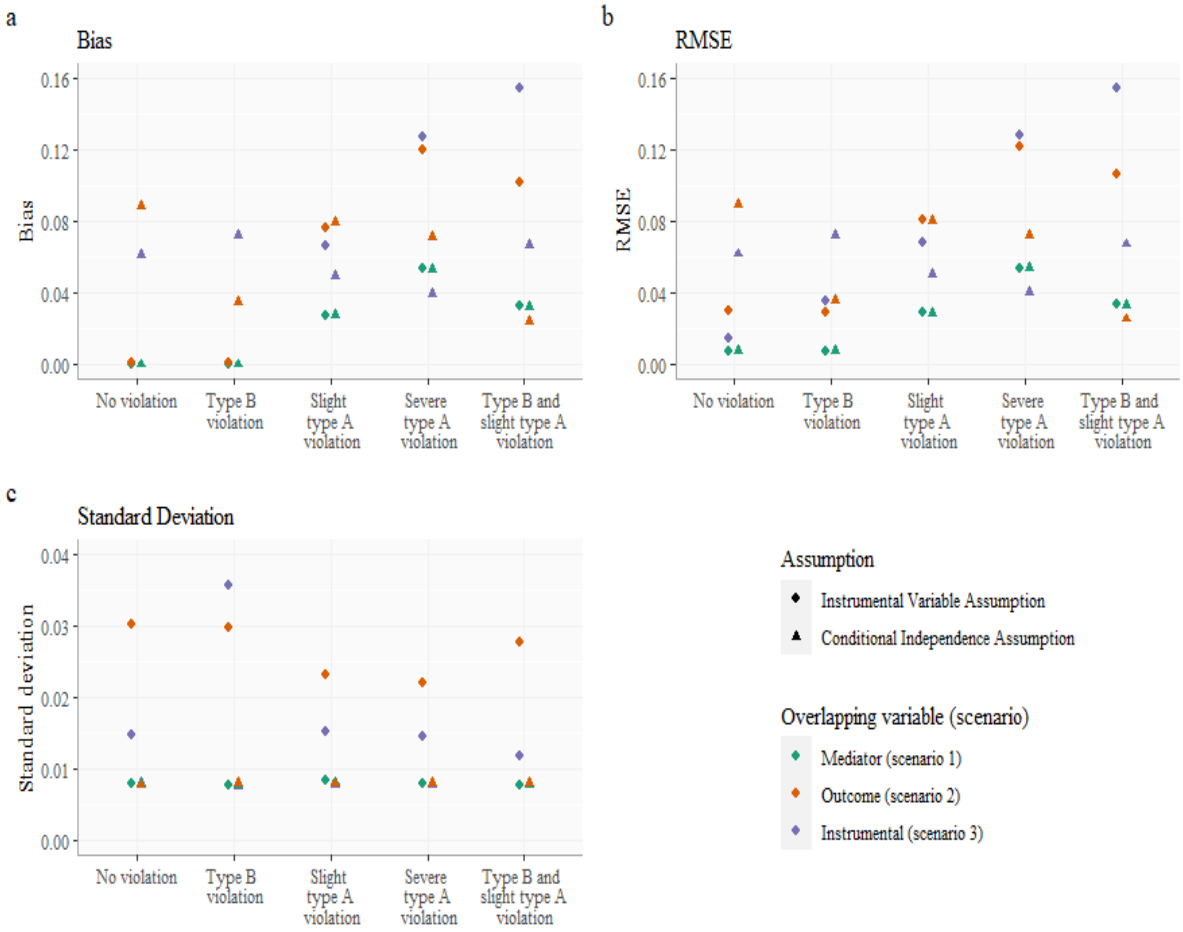
	Overlapping variable									
	X = O (scenario 2)					X = I (scenario 3)				
IVA violation	-	+	*	**	+*	-	+	*	**	+*
CIA violation	**	*	**	**	*	**	**	*	*	*
$OR(\mathrm{Z},\mathrm{Y}\,\vert\mathrm{X}=x)$	7	2	5.9	5.4	1.6	4	4	3.1	2.5	3.7
	Bias									
CIA	.0892	.0353	.0801	.0718	.0245	.0616	.0724	.0501	.0398	.0670
IVA	.0013	.0014	.0767	.1198	.1025	.0004	.0010	.0664	.1272	.1543
	% conditional Fréchet bound covered by bias									
CIA	30.2	11.3	30.0	29.8	8.7	20.8	21.5	15.8	12.6	19.3
IVA	0.4	0.45	28.7	49.7	36.5	0.1	0.3	20.9	40.3	44.3
	RMSE									
CIA	.089	.036	.080	.072	.026	.062	.073	.051	.041	.067
IVA	.030	.030	.081	.122	.106	.015	.036	.068	.128	.155

Note: - indicates no violation, + indicates a weak association between the instrumental and mediator variables, * indicates a slight violation, ** indicates a severe violation and +* indicates a weak association between the instrumental and mediator variables combined with a slight violation. O and I represent the outcome and instrumental variables. X represents the overlapping variable, Y and Z represent the non-overlapping variables. CIA represents the Conditional Independence Assumption, IVA the Instrumental Variable Assumption.

For the other scenarios results are quite different since the CIA is violated to different degrees compared to the IVA. Table 4.1 shows to what extent the CIA is violated in the different populations for scenarios two and three. For instance, when the IVA is not violated, the CIA is severely violated in both scenarios. Table 4.1 also contrasts the bias and its magnitude (as measured by the proportion of the conditional Fréchet bound covered by the bias) observed under the IVA and the CIA. Of note is the fact that for severe violations of the CIA, the bias is smaller compared to severe violations of the IVA. For instance, when the IVA is severely violated the bias under the IVA is .1198. In the same situation the CIA is also severely violated with a bias .0718.

Figure 4.4

Bias, RMSE and standard deviation of five populations using the Instrumental Variable and Conditional Independence approaches to statistical matching



Note: A type B violation indicates the IVA to be violated with respect to the relationship between the instrumental and mediator variable, a type A violation indicates the IVA to be violated with respect to the relationship between the instrumental and outcome variables.

4.2.2 RMSE

The RMSE was used to compare the accuracy of the estimation between the IVA and the CIA approaches to statistical matching. Figure 4.4b shows the RMSEs of both methods side by side. As with the bias, the RMSEs for scenario one are identical since estimation is exactly the same. For the other scenarios the conclusions are similar to the conclusions for the bias, since low variance accounts for the bias to be very close to the RMSE. Overall, when the IVA is violated, accuracy is higher when using the CIA irrespective of to what degree the CIA is violated. When the IVA is not violated, accuracy is higher when using the IVA for estimation.

4.3.3 Precision and uncertainty

4.1.3a Standard deviation

To compare precision of the estimates for various situations, the standard deviations for the IVA and CIA approach to statistical matching were compared. Figure 4.4c shows these standard deviations side by side. The standard deviations when estimating under the CIA are constant across populations and matching scenarios. Consequently, they all align with the standard deviations of scenario one when statistically matching under the IVA.

4.1.3b Confidence intervals

In addition to 95% confidence intervals for the IVA estimates, Figure 4.2 also includes these intervals for the CIA estimates. The confidence intervals for all scenarios and populations under the CIA coincide with the confidence intervals for scenario one under the IVA. This can be directly related to the similar standard deviations. The estimations and their confidence intervals under the CIA also never venture outside the conditional Fréchet bounds.

Chapter 5

Discussion

This project aimed to evaluate a new assumption – the IVA – as the basis for statistical matching and compare it to the classically used CIA. The main difference between these methods is that the CIA always assumes the non-overlapping variables to be independent given the overlapping variable, while the IVA always assumes the instrumental and outcome variables to be independent given the mediating variable, irrespective of which of these variables overlaps in the data. As such, the methods were compared for scenarios where either the mediating, the instrumental, or the outcome variable was the overlapping variable. This was done for situations where the IVA was violated to various degrees. The bias, accuracy, precision, and uncertainty were assessed and compared for both methods.

The results indicate that a slight violation of the IVA results in bias and moderate loss of accuracy, while a severe violation of the IVA results in a substantial bias and loss of accuracy. This effect is strongest when the outcome variable is also the overlapping variable. Having a strong relationship between the instrumental and mediating variables is a protective factor against a slight violation of the IVA. Additionally, the results show that when the IVA is not violated, the IVA outperforms the CIA when the outcome or instrumental variable is the overlapping variable. When the mediator is the overlapping variable, either the IVA or the CIA can be used as they give identical results. Interestingly, in this specific study, the CIA seems to be more robust against any type of violation, particularly a severe violation. Overall, the CIA seems to be a good method that is quite robust against violations of the assumption.

The study by Kim et al. (2016) also found a small violation of the IVA to not result in large biases. This is the only study that was found that uses an IVA approach to statistical matching. However, it is hard to compare the current study to theirs since they used a vastly different method for continuous data and only assessed the situation where the instrumental variable overlaps in the data. The importance of further exploration of the IVA approach has recently been emphasized by D'Orazio (2024).

The current study highlights that, in some situations, the IVA approach might be more appropriate than relying on the CIA, particularly when the IVA is expected to hold well. Expert knowledge in the area of the variables under investigation can indicate whether this expectation is justified. In their 2017 paper, D'Orazio et al. highlighted the importance of using expert knowledge when selecting matching variables. In the context of the IVA approach to statistical matching, this can be crucial.

This study provides a clear and direct comparison between the IVA and the CIA approach to statistical matching in identical situations. This way, insight is gained into which method is most suited

for which conditions. Furthermore, this study highlights to what degree the CIA is violated for different violations of the IVA. This offers a deeper understanding of the limitations but also of the strengths of both methods. Comparison of the approaches is more straightforward when one knows the exact violation of the assumptions. An additional strength of this study is the use of odds ratios. By defining the population distribution according to a certain odds ratio, total control over the associations between the variables is possible. This level of control enhances reliability and validity of the analysis because it is possible to have exact control over the degree to which the assumption is violated.

A number of potential points of improvement can be identified. The confidence intervals and estimates obtained using the IVA approach sometimes exceed the conditional Fréchet bounds, which in principle should be impossible. This might be the case because conditional Fréchet bounds are calculated conditioning on the overlapping variable, while the IVA approach always conditions on the mediating variable. In scenarios two and three, the mediating variable is not the overlapping variable. In the simulation study the Fréchet bounds were only exceeded in populations where the IVA was violated. Finding a way to calculate Fréchet bounds specifically for the IVA approach might be valuable in future research as they can then be used as a diagnostic tool to see whether the IVA is violated in a given sample. An additional cause for exceeding Fréchet bounds might be that the probabilities are not restricted to these bounds during estimation. In future simulations this might be valuable to get a more accurate estimate of the bias.

In addition, this study only considered categorical variables with two categories, which may limit generalizability of the results. Exploring situations with more categories could provide a more comprehensive understanding of the IVA approach to statistical matching. For the situation where the mediator is the overlapping variable, this is not a problem since all probabilities can be directly estimated. For the other two scenarios, this is less straightforward.

Say the instrumental, mediating and outcome variables have C_I, C_M, and C_O categories, respectively. When the mediating variable overlaps, it is always possible to directly calculate estimates for $P(O = o \mid I = i)$, $(o = 1, \dots, C_O; i = 1, \dots, C_I)$. When the outcome variable overlaps, there are $C_I C_O$ equations in the form of Equation (2) and C_I equations of the form $\sum_{m=1}^{C_M} P(M = m \mid I = i) = 1$. In total that gives $C_I C_O + C_I$ equations (of which some might be redundant) and $C_I C_M$ unknowns. Similarly, when the instrumental variable overlaps, there is a total of $C_I C_O + C_M$ equations and $C_O C_M$ unknowns. If the number of (non-redundant) equations exceeds the number of unknowns, the system of equations is overdetermined and has no exact solution. If the number of unknowns exceeds the number of (non-redundant) equations the system is underdetermined and has several solutions.

An example of a situation where the system is underdetermined is when the instrumental and outcome variables have two categories but the mediating variable has three categories. In that case there are six equations when the outcome variable overlaps and seven equations when the instrumental variable overlaps. In both situations there are two redundant equations (for the same reason mentioned

in Appendix A), and six unknowns. It is then impossible to find a unique solution for $P(O = o \,|M = m)$ (if the instrumental variable overlaps) or $P(M = m \,|I = i)$ (if the outcome variable overlaps). A unique solution exists if and only if the number of unknowns is equal to the number of non-redundant equations. This poses an extra challenge when studying variables with more than two categories.

Future endeavors into the use of the IVA for statistical matching can explore several different aspects. It would be worthwhile to study the accuracy and precision of statistical matching under the IVA and the CIA when the distribution of one or more of the variables in the population is more extreme. An interesting question here would be whether the CIA would still result in an overall lower bias compared to the IVA approach. Studying situations where one or more variables have a highly skewed distribution could provide further insights into the robustness and real-world applicability of the IVA method and how it compares to the CIA. Moreover, studying the IVA approach when there are more categories or when one or more variables are continuous might be valuable. Investigating different variable types can deepen understanding of the IVA approach and shine a light on the specific situations where the IVA might be more suitable than the CIA. Finally, the current study assessed how the CIA performs under different IVA violations. It would also be interesting to assess how the IVA performs under different violations of the CIA.

In conclusion, the IVA approach to statistical matching might be a good alternative in certain situations since it offers more flexibility with regard to the overlapping variable. When a researcher is certain about an IV mechanism being present, the IVA approach, rather than the CIA approach, can be especially useful when a variable other than the mediator overlaps. When the IVA holds, the IVA approach provides more accurate results in situations where the mediator is not the overlapping variable compared to the CIA approach. It might also be useful when there are a number of overlapping variables, with one variable possibly being part of an IV mechanism with the non-overlapping variables. The researcher can then choose to use these variables for the statistical matching procedure. When it is uncertain whether an IV mechanism exists in the population, using the CIA might be more advisable as it is more robust against violations in general. All in all, this research offers a first insight into a new approach to statistical matching and can serve as a stepping stone for future endeavors.

References

CBS (2023). [Homepage of CBS's Statline]. Retrieved November 27th, 2023, from https://opendata.cbs.nl/statline/#/CBS/nl/

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. Communications in Statistics - Simulation and Computation, 39(4), 860–864. https://doi.org/10.1080/03610911003650383

Conti, P. L., Marella, D., & Scanu, M. (2012). Uncertainty analysis in statistical matching. *Journal of Official Statistics*. *28*(1), 69–88.

Conti, P. L., Marella, D., & Scanu, M. (2017). How far from identifiability? A systematic overview of the statistical matching problem in a non parametric framework. *Communications in Statistics-Theory and Methods*, *46*(2), 967–994. https://doi.org/10.1080/03610926.2015.1010005

D'Orazio, M. (2019). Statistical learning in official statistics: The case of statistical matching. *Statistical Journal of IAOS*. *35*(3), 435–441. https://doi.org/10.3233/SJI-190518

D'Orazio, M., Di Zio, M., & Scanu, M. (2017). The use of uncertainty to choose matching variables in statistical matching. *International Journal of Approximate Reasoning. 90*. 433–440. https://doi.org/10.1016/j.ijar.2017.08.015

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons.

D'Orazio, M., Di Zio, M., & Scanu, M. (2024). What is the state of play on statistical matching with a focus on auxiliary information, complex survey designs and quality issues? *The Survey Statistician. 89*, 47–58.

Instrumental variable estimation (2024, January, 4). In *Wikipedia*. Retrieved January 24, 2024, from https://en.wikipedia.org/wiki/Instrumental_variables_estimation.

Kim, J., Berg, E., & Park, T. (2016). Statistical matching using fractional imputation. *Survey methodology*. *42*(1), 19–40.

Leulescu, A., and Agafitei, M. (2013). Statistical matching: A model based approach for data integration. *Eurostat Methodologies and Working Papers.*

Newhouse, J. P., & McClellan, M. (1998). Econometrics in outcomes research: the use of instrumental variables. *Annual review of public health*, *19*(1), 17–34. https://doi.org/10.1146/annurev.publhealth.19.1.17

R Core Team (2023). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/.

Appendix

Appendix A

Equations to be solved when the mediator is not the overlapping variable

Below the equations are given for calculating the conditional probabilities in an IV situation where each variable has two categories. In the second scenario where the outcome variable is overlapping in both datasets, this entails solving the equations for $P(M = m \mid I = i)$. In the third scenario where the instrumental variable is overlapping in both datasets, this entails solving the equations for $P(O = o \mid M = m)$.

$$P(O = 1 \mid I = 1) =$$
$$= P(O = 1 \mid M = 1)P(M = 1 \mid I = 1) + P(O = 1 \mid M = 2)P(M = 2 \mid I = 1)$$

$$P(O = 2 \mid I = 1) =$$
$$= P(O = 2 \mid M = 1)P(M = 1 \mid I = 1) + P(O = 2 \mid M = 2)P(M = 2 \mid I = 1)$$

$$P(O = 1 \mid I = 2) =$$
$$= P(O = 1 \mid M = 1)P(M = 1 \mid I = 2) + P(O = 1 \mid M = 2)P(M = 2 \mid I = 2)$$

$$P(O = 2 \mid I = 2) =$$
$$= P(O = 2 \mid M = 1)P(M = 1 \mid I = 2) + P(O = 2 \mid M = 2)P(M = 2 \mid I = 2)$$

Two of these equations are in fact redundant since

$$P(O = 1 \mid I = 1) + P(O = 2 \mid I = 1) = 1$$

and

$$P(O = 1 \mid I = 2) + P(O = 2 \mid I = 2) = 1$$

Appendix B

Determining the number of simulations and bootstrap samples

To determine how many simulation replications (R) and bootstrap samples (B) were necessary, one population was used to run a test simulation and subsequent bootstrap analysis. From the test simulation, ten sample pairs of $n = 2000$ were saved and used for the bootstrap analysis. The simulated and bootstrapped bias was calculated in the same way as described in Chapter 3 of this thesis. Table B.1 describes the settings used to generate the population to sample from.

In Figure B.1 the bias for the test simulation for several *R* replications is shown. The differences in the bias are not very large in general. With computational time and efficiency in mind it was decided that $R = 500$ would be sufficient. Figure B.2 depicts the bias for the test bootstrap for several B bootstrap samples. The differences here are small and again with computation time in mind, it was determined that $B = 100$ bootstrap samples (per simulated sample) would be sufficient. The results highlighted in Chapter 4 confirm that 100 bootstrap samples was sufficient.

Table B.1

Population used to do a test simulation in order to determine the number of simulation replications and bootstrap samples

	Test population
N	1 000 000
P(Mediator = 1)	.850
P(Outcome = 1)	.652
P(Instrumental = 1)	.900
OR(Mediator, Outcome)	4
OR(Instrumental, Mediator)	7
OR(Outcome, Instrumental | Mediator)	1

Figure B.1

Development of the estimated simulated bias in a statistical matching procedure using the IVA approach, for an increasing number of Monte Carlo simulations

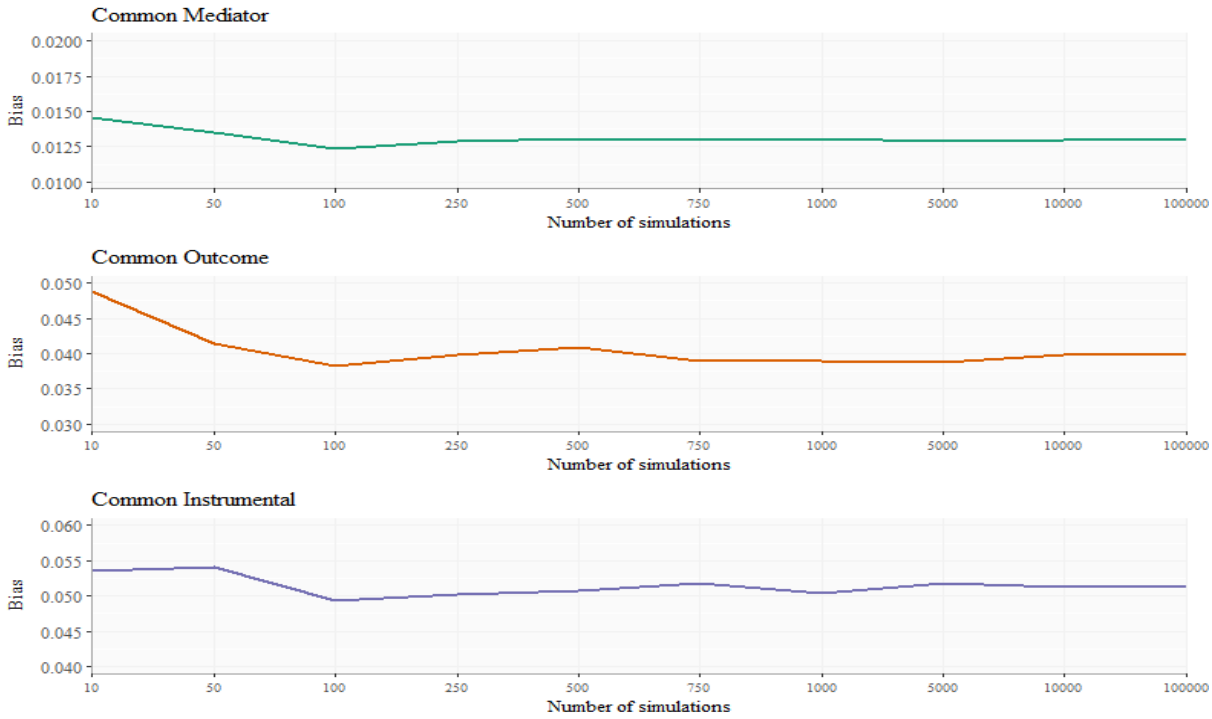
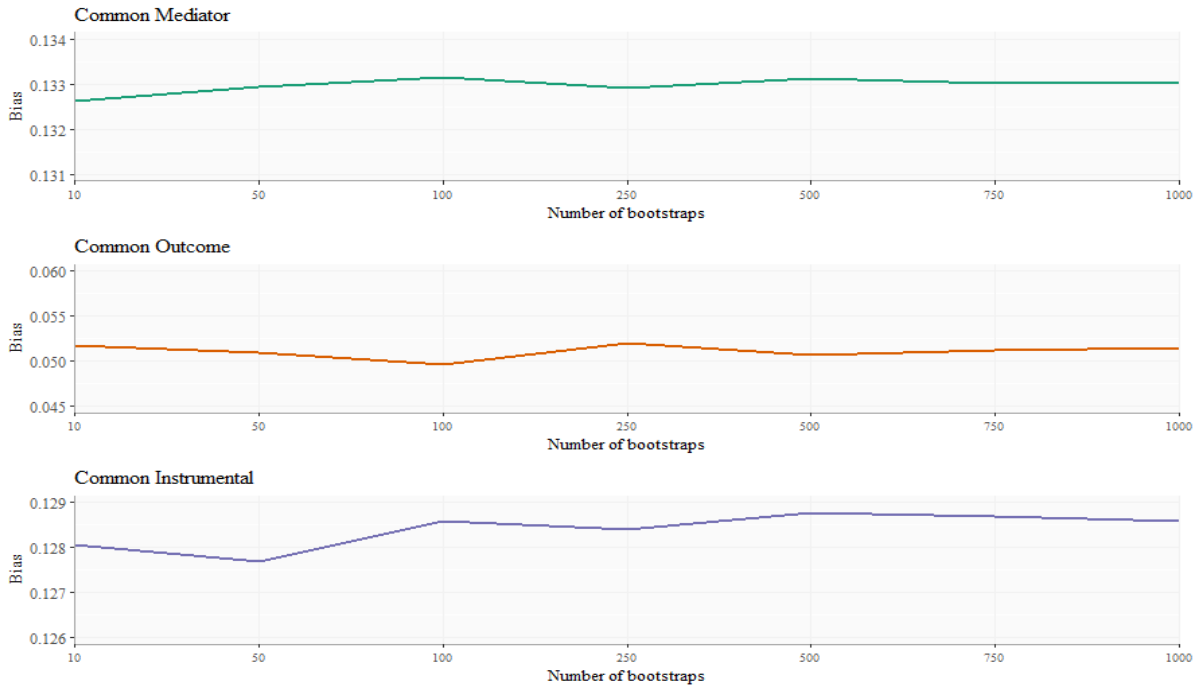


Figure B.2

Development of the estimated bootstrap bias in a statistical matching procedure using the IVA approach, for an increasing number of Bootstrap samples



Note: bootstrap analysis is based on ten simulated samples, number of bootstraps indicate the number of bootstrap samples taken from each simulated sample.

Appendix C

Estimates of the simulation study with bootstrap analysis

Table C.1

Population value, marginal and conditional Fréchet bounds, simulated and bootstrap estimates, standard deviations and 95% confidence intervals for the IVA and CIA approach to statistical matching, for five populations, three scenarios and four category combinations

						IVA				CIA			
						Simulation		Bootstrap		Simulation		Bootstrap	
Population - IVA violation	Overlapping variable	Category combination	Population value	Marginal Frechet bounds	Conditional Frechet bounds	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI
1 - no violation	Mediator	I1O1	.31	[.08, .43]	[.22, .43]	.31 (.009)	[.29, .33]	.31 (.009)	[.29, .33]	.31 (.009)	[.29, .33]	0.31 (.009)	[.29, .33]
		I2O1	.34	[.22, .57]	[.22, .43]	.34 (.010)	[.32, .36]	.35 (.010)	[.33, .36]	.34 (.010)	[.32, .36]	0.34 (.010)	[.33, .36]
		I1O2	.12	[.00, .35]	[.00, .21]	.12 (.006)	[.11, .13]	.12 (.005)	[.11, .13]	.12 (.006)	[.11, .13]	0.12 (.005)	[.11, .13]
		I2O2	.23	[.00, .35]	[.14, .35]	.23 (.008)	[.22, .25]	.23 (.008)	[.21, .24]	.23 (.008)	[.22, .25]	0.23 (.008)	[.22, .24]
	Outcome	I1M1	.36	[.04, .43]	[.13, .43]	.36 (.027)	[.32, .42]	.37 (.024)	[.32, .41]	.27 (.009)	[.26, .29]	0.28 (.009)	[.26, .29]
		I1M2	.07	[.00, .39]	[.00, .29]	.07 (.026)	[.01, .11]	.06 (.023)	[.02, .11]	.16 (.006)	[.14, .17]	0.34 (.009)	[.32, .36]

						IVA				CIA			
						Simulation		Bootstrap		Simulation		Bootstrap	
Population - IVA violation	Overlapping variable	Category combination	Population value	Marginal Frechet bounds	Conditional Frechet bounds	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI
2 – type B violation	Instrumental	I2M1	.25	[.19, .57]	[.19, .48]	.25 (.035)	[.18, .31]	.26 (.035)	[.19, .32]	.34 (.009)	[.32, .36]	0.16 (.006)	[.14, .17]
		I2M2	.32	[.00, .39]	[.09, .39]	.32 (.033)	[.26, .39]	.31 (.034)	[.25, .38]	.23 (.007)	[.22, .25]	0.23 (.008)	[.22, .25]
		M1O1	.48	[.27, .61]	[.27, .56]	.47 (.016)	[.45, .51]	.47 (.016)	[.44, .50]	.41 (.010)	[.39, .43]	0.42 (.010)	[.40, .43]
		M2O1	.18	[.04, .39]	[.09, .39]	.18 (.016)	[.15, .21]	.18 (.016)	[.15, .21]	.24 (.009)	[.22, .26]	0.24 (.008)	[.22, .25]
		M1O2	.14	[.00, .35]	[.05, .35]	.14 (.014)	[.11, .17]	.14 (.015)	[.12, .17]	.20 (.008)	[.19, .22]	0.20 (.008)	[.19, .22]
		M2O2	.21	[.00, .35]	[.00, .30]	.21 (.014)	[.18, .24]	.20 (.015)	[.17, .23]	.15 (.006)	[.14, .16]	0.14 (.006)	[.13, .15]
	Mediator	I1O1	.29	[.08, .43]	[.16, .43]	.29 (.008)	[.28, .31]	.29 (.009)	[.28, .31]	.29 (.008)	[.28, .31]	0.29 (.009)	[.28, .31]
		I2O1	.36	[.22, .57]	[.22, .49]	.36 (.010)	[.34, .38]	.36 (.010)	[.34, .37]	.36 (.010)	[.34, .38]	0.36 (.010)	[.34, .37]
		I1O2	.14	[.00, .35]	[.00, .27]	.14 (.006)	[.13, .15]	.14 (.006)	[.13, .15]	.14 (.006)	[.13, .15]	0.14 (.006)	[.13, .15]
		I2O2	.21	[.00, .35]	[.08, .35]	.21 (.007)	[.20, .22]	.21 (.007)	[.20, .23]	.21 (.007)	[.20, .22]	0.21 (.007)	[.20, .23]
	Outcome	I1M1	.30	[.04, .43]	[.12, .43]	.30 (.027)	[.25, .35]	.30 (.027)	[.25, .35]	.27 (.008)	[.25, .28]	0.26 (.009)	[.25, .28]

						IVA				CIA			
						Simulation		Bootstrap		Simulation		Bootstrap	
Population - IVA violation	Overlapping variable	Category combination	Population value	Marginal Frechet bounds	Conditional Frechet bounds	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI
		I1M2	.13	[.00, .38]	[.00, .31]	.13 (.027)	[.08, .18]	.13 (.026)	[.08, .18]	.16 (.007)	[.15, .17]	0.35 (.009)	[.33, .36]
		I2M1	.31	[.19, .57]	[.19, .50]	.31 (.032)	[.25, .38]	.33 (.034)	[.26, .39]	.35 (.010)	[.33, .37]	0.16 (.006)	[.15, .17]
		I2M2	.26	[.00, .38]	[.07, .38]	.26 (.032)	[.19, .32]	.25 (.033)	[.18, .31]	.22 (.008)	[.21, .24]	0.23 (.008)	[.21, .24]
	Instrumental	M1O1	.48	[.27, .62]	[.27, .60]	.48 (.035)	[.41, .54]	.48 (.034)	[.42, .55]	.40 (.009)	[.39, .42]	0.41 (.009)	[.39, .42]
		M2O1	.18	[.04, .38]	[.05, .38]	.18 (.035)	[.10, .24]	.18 (.035)	[.11, .24]	.25 (.008)	[.23, .27]	0.25 (.008)	[.23, .26]
		M1O2	.14	[.00, .35]	[.01, .35]	.14 (.036)	[.07, .21]	.14 (.033)	[.07, .20]	.21 (.008)	[.20, .23]	0.21 (.007)	[.20, .23]
		M2O2	.21	[.00, .35]	[.00, .34]	.21 (.036)	[.14, .28]	.21 (.034)	[.15, .28]	.14 (.005)	[.12, .15]	0.13 (.005)	[.12, .14]
3 – slight type A violation	Mediator	I1O1	.34	[.08, .43]	[.22, .43]	.31 (.009)	[.29, .33]	.31 (.009)	[.29, .33]	.31 (.009)	[.29, .33]	0.31 (.009)	[.29, .33]
		I2O1	.31	[.22, .57]	[.22, .43]	.34 (.010)	[.32, .36]	.34 (.010)	[.32, .36]	.34 (.010)	[.32, .36]	0.34 (.009)	[.32, .36]
		I1O2	.09	[.00, .35]	[.00, .21]	.12 (.006)	[.11, .13]	.12 (.006)	[.11, .13]	.12 (.006)	[.11, .13]	0.12 (.006)	[.11, .13]
		I2O2	.26	[.00, .35]	[.14, .35]	.23 (.008)	[.21, .25]	.23 (.008)	[.22, .24]	.23 (.008)	[.21, .24]	0.23 (.008)	[.22, .25]

						IVA				CIA			
						Simulation		Bootstrap		Simulation		Bootstrap	
Population - IVA violation	Overlapping variable	Category combination	Population value	Marginal Frechet bounds	Conditional Frechet bounds	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI
	Outcome	I1M1	.36	[.04, .43]	[.16, .43]	.43 (.012)	[.40, .45]	.42 (.015)	[.39, .45]	.28 (.008)	[.27, .30]	0.28 (.009)	[.27, .30]
		I1M2	.07	[.00, .39]	[.00, .27]	.00 (.006)	[.00, .02]	.01 (.010)	[.00, .03]	.15 (.006)	[.13, .16]	0.33 (.009)	[.32, .35]
		I2M1	.25	[.19, .57]	[.19, .45]	.16 (.038)	[.09, .23]	.15 (.039)	[.08, .22]	.33 (.009)	[.31, .35]	0.15 (.006)	[.14, .16]
		I2M2	.32	[.00, .39]	[.12, .39]	.41 (.037)	[.34, .48]	.42 (.038)	[.35, .49]	.24 (.008)	[.22, .25]	0.24 (.008)	[.22, .25]
	Instrumental	M1O1	.48	[.27, .61]	[.27, .59]	.54 (.016)	[.51, .58]	.54 (.016)	[.51, .57]	.43 (.010)	[.41, .44]	0.43 (.009)	[.41, .44]
		M2O1	.18	[.04, .39]	[.06, .38]	.11 (.016)	[.08, .14]	.11 (.016)	[.08, .14]	.23 (.008)	[.21, .24]	0.23 (.008)	[.21, .24]
		M1O2	.14	[.00, .35]	[.02, .34]	.07 (.014)	[.04, .10]	.07 (.014)	[.05, .10]	.19 (.007)	[.18, .20]	0.19 (.007)	[.18, .20]
		M2O2	.21	[.00, .35]	[.01, .32]	.27 (.015)	[.24, .30]	.28 (.015)	[.25, .31]	.16 (.006)	[.15, .17]	0.16 (.006)	[.15, .17]
4 – severe type A violation	Mediator	I1O1	.36	[.08, .43]	[.22, .43]	.31 (.008)	[.30, .33]	.31 (.009)	[.29, .33]	.31 (.008)	[.30, .33]	0.31 (.009)	[.29, .33]
		I2O1	.29	[.22, .57]	[.22, .43]	.34 (.010)	[.32, .36]	.34 (.010)	[.33, .36]	.34 (.010)	[.32, .36]	0.34 (.009)	[.33, .36]
		I1O2	.06	[.00, .35]	[.00, .21]	.12 (.006)	[.11, .13]	.12 (.006)	[.11, .13]	.12 (.006)	[.11, .13]	0.12 (.006)	[.11, .13]

						IVA				CIA			
						Simulation		Bootstrap		Simulation		Bootstrap	
Population - IVA violation	Overlapping variable	Category combination	Population value	Marginal Frechet bounds	Conditional Frechet bounds	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI
		I2O2	.28	[.00, .35]	[.14, .35]	.23 (.008)	[.21, .24]	.23 (.008)	[.21, .24]	.23 (.008)	[.21, .24]	0.23 (.008)	[.22, .24]
	Outcome	I1M1	.36	[.04, .43]	[.19, .43]	.43 (.011)	[.41, .45]	.42 (.011)	[.40, .45]	.29 (.009)	[.27, .31]	0.29 (.009)	[.27, .31]
		I1M2	.07	[.00, .38]	[.00, .24]	.00 (.000)	[.00, .00]	.00 (.000)	[.00, .00]	.14 (.006)	[.13, .15]	0.33 (.009)	[.31, .34]
		I2M1	.25	[.19, .57]	[.19, .43]	.08 (.040)	[.00, .16]	.08 (.037)	[.02, .15]	.32 (.010)	[.31, .34]	0.14 (.006)	[.12, .15]
		I2M2	.32	[.00, .38]	[.14, .39]	.49 (.038)	[.42, .57]	.49 (.036)	[.42, .56]	.25 (.008)	[.23, .26]	0.25 (.008)	[.23, .26]
	Instrumental	M1O1	.48	[.27, .62]	[.30, .62]	.60 (.014)	[.57, .62]	.60 (.014)	[.57, .62]	.44 (.009)	[.42, .45]	0.43 (.009)	[.42, .45]
		M2O1	.18	[.04, .38]	[.04, .35]	.05 (.017)	[.01, .08]	.05 (.016)	[.02, .08]	.22 (.008)	[.20, .23]	0.22 (.008)	[.20, .23]
		M1O2	.14	[.00, .35]	[.00, .32]	.01 (.011)	[.00, .04]	.01 (.010)	[.00, .04]	.18 (.007)	[.17, .19]	0.18 (.007)	[.17, .19]
		M2O2	.21	[.00, .35]	[.03, .35]	.34 (.016)	[.31, .37]	.34 (.015)	[.31, .37]	.17 (.007)	[.16, .18]	0.17 (.007)	[.16, .18]
5 – type A and type B violation	Mediator	I1O1	.32	[.08, .43]	[.16, .43]	.29 (.008)	[.27, .31]	.29 (.009)	[.27, .31]	.29 (.008)	[.28, .31]	0.29 (.009)	[.27, .31]
		I2O1	.33	[.22, .57]	[.22, .49]	.36 (.010)	[.34, .38]	.36 (.010)	[.34, .38]	.36 (.010)	[.34, .38]	0.36 (.010)	[.34, .38]

						IVA				CIA			
						Simulation		Bootstrap		Simulation		Bootstrap	
Population - IVA violation	Overlapping variable	Category combination	Population value	Marginal Frechet bounds	Conditional Frechet bounds	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI	Estimate (SD)	95% CI
		I1O2	.10	[.00, .35]	[.00, .27]	.14 (.006)	[.13, .15]	.14 (.006)	[.13, .15]	.14 (.006)	[.13, .15]	0.14 (.006)	[.13, .15]
		I2O2	.24	[.00, .35]	[.08, .35]	.21 (.007)	[.20, .22]	.21 (.008)	[.20, .23]	.21 (.007)	[.20, .23]	0.21 (.008)	[.20, .23]
	Outcome	I1M1	.30	[.04, .43]	[.15, .43]	.41 (.022)	[.36, .44]	.41 (.020)	[.36, .44]	.28 (.008)	[.26, .30]	0.28 (.009)	[.26, .29]
		I1M2	.13	[.00, .38]	[.00, .28]	.02 (.021)	[.00, .07]	.02 (.017)	[.00, .06]	.15 (.006)	[.14, .16]	0.34 (.009)	[.32, .36]
		I2M1	.31	[.19, .57]	[.19, .47]	.21 (.034)	[.14, .27]	.21 (.037)	[.13, .27]	.34 (.009)	[.32, .36]	0.15 (.006)	[.14, .16]
		I2M2	.26	[.00, .38]	[.10, .38]	.36 (.034)	[.30, .43]	.37 (.036)	[.30, .43]	.23 (.008)	[.22, .25]	0.23 (.008)	[.22, .25]
	Instrumental	M1O1	.48	[.27, .62]	[.27, .62]	.61 (.011)	[.59, .64]	.61 (.013)	[.58, .63]	.41 (.009)	[.39, .43]	0.40 (.010)	[.39, .42]
		M2O1	.18	[.04, .38]	[.04, .38]	.01 (.015)	[.00, .05]	.01 (.014)	[.00, .04]	.24 (.009)	[.23, .26]	0.25 (.008)	[.23, .26]
		M1O2	.14	[.00, .35]	[.00, .35]	.00 (.004)	[.00, .01]	.00 (.004)	[.00, .01]	.21 (.008)	[.19, .22]	0.21 (.007)	[.19, .22]
		M2O2	.21	[.00, .35]	[.00, .35]	.38 (.017)	[.34, .40]	.38 (.017)	[.34, .41]	.14 (.006)	[.13, .15]	0.14 (.006)	[.13, .16]

Note: numbers between brackets denote the standard deviation of the estimate. Category combinations are coded, here *I* stands for the instrumental variable, *O* for the outcome variable and *M* for the mediating variable. Numbers indicate the value of that variable, for instance if the category combination is *I2O1*, the instrumental variable has value 2 and the outcome variable has value 1.

Table C.2

Marginal and conditional Fréchet bound widths, simulated and bootstrap bias, RMSE and standard deviation estimates for the IVA and CIA approaches to statistical matching, for five populations and three scenarios

				IVA						CIA					
				Simulation			Bootstrap			Simulation			Bootstrap		
Population	Overlapping variable	Width marginal Fréchet bound	Width conditional Fréchet bound	Bias	RMSE	SD	Bias	RMSE	SD	Bias	RMSE	SD	Bias	RMSE	SD
1	Instrumental	.348	.298	.0004	.015	.015	.0059	.019	.015	.0616	.062	.008	.0629	.063	.008
	Mediator	.350	.210	.0002	.008	.008	.0012	.010	.008	.0002	.008	.008	.0011	.010	.008
	Outcome	.388	.295	.0013	.030	.030	.0059	.038	.029	.0892	.089	.008	.0886	.089	.008
2	Instrumental	.348	.318	.0664	.068	.015	.0660	.068	.015	.0501	.051	.008	.0510	.052	.008
	Mediator	.350	.210	.0278	.029	.008	.0267	.028	.008	.0278	.029	.008	.0264	.028	.008
	Outcome	.388	.268	.0767	.081	.023	.0796	.084	.026	.0801	.080	.008	.0801	.080	.008
3	Instrumental	.348	.318	.1272	.128	.015	.1244	.125	.014	.0398	.041	.008	.0424	.043	.008
	Mediator	.350	.210	.0537	.054	.008	.0561	.057	.008	.0537	.054	.008	.0560	.057	.008
	Outcome	.382	.242	.1198	.122	.022	.1203	.123	.021	.0718	.072	.008	.0692	.070	.008

				IVA						CIA					
				Simulation			Bootstrap			Simulation			Bootstrap		
Population	Overlapping variable	Width marginal Fréchet bound	Width conditional Fréchet bound	Bias	RMSE	SD	Bias	RMSE	SD	Bias	RMSE	SD	Bias	RMSE	SD
4	Instrumental	.348	.335	.0010	.036	.036	.0015	.053	.034	.0724	.073	.008	.0714	.072	.008
	Mediator	.350	.270	.0003	.008	.008	.0025	.010	.008	.0003	.008	.008	.0026	.010	.008
	Outcome	.382	.310	.0014	.030	.030	.0092	.039	.030	.0353	.036	.008	.0331	.034	.008
5	Instrumental	.348	.348	.1543	.155	.012	.1520	.153	.012	.0670	.067	.008	.0690	.069	.008
	Mediator	.350	.270	.0328	.034	.008	.0321	.033	.008	.0328	.034	.008	.0322	.033	.008
	Outcome	.382	.280	.1025	.106	.028	.1054	.110	.028	.0245	.026	.008	.0247	.026	.008

Appendix D

Reference to online repository

https://github.com/avdmerbel/AEvdMerbel_Thesis_StatisticalMatching