Mediation Analysis of Nurse Practitioner Impact on Renal Outcome in CKD Patients

Author: Luna Elise Schernthaner  
s4703928

First supervisor/assessor:  
dr. ir. T. Claassen  
tomc@cs.ru.nl

Second supervisor:  
dr. A.J.G. van den Brand  
jan.vandenbrand@radboudumc.nl

Second assessor:  
prof. dr. T.M. Heskes  
t.heskes@science.ru.nl

June 27, 2020
Abstract

Previous efforts have shown a positive association between chronic kidney disease (CKD) progression and blood pressure, proteinuria and salt intake management. However, the exact extent of these associations is still unknown. With this knowledge, guidelines for CKD treatment could be refined to advance CKD care, which in turn could improve outcomes for CKD patients. This research aims to find the extent of these effects using mediation analysis on the MASTERPLAN dataset. The MASTERPLAN trial studied outcomes and guideline adherence in 788 CKD patients, allocated to receive standard care or additional support of a nurse practitioner to achieve 11 treatment goals over a median follow-up time of 5.7 years. We present four hypothetical mediation models to find if the effect of the nurse practitioner intervention on renal outcome is mediated by the individual blood pressure, proteinuria, ACE inhibitor or ARB use, and salt intake treatment goals individually. From this analysis follows that none of these four treatment goals significantly mediate the impact of the intervention on the renal endpoint. However, the dataset suffered from significant contamination bias, which means that the found effects may be smaller than their true values. Based on these findings, we conclude that the nurse practitioner intervention may not improve renal outcome in CKD patients in general, nor through the individual blood pressure, proteinuria, ACE inhibitor or ARB use, or salt intake treatment goals.
Acknowledgements

I would like to thank my supervisor from the Radboud University Medical Center, Jan van den Brand for the idea for this thesis, giving me access to the MASTERPLAN dataset, and his guidance. I would also like to thank my supervisor from the Institute for Computing and Information Sciences, Tom Claassen for his input and advice. I am very thankful for their time, even with all of the events going on.

Finally, I would like to thank everyone who helped me throughout this semester, making me believe in myself and supporting me through difficult times.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>38</td>
</tr>
<tr>
<td>5.1</td>
<td>38</td>
</tr>
<tr>
<td>5.2</td>
<td>39</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
</tr>
<tr>
<td>6.1</td>
<td>43</td>
</tr>
<tr>
<td>A</td>
<td>56</td>
</tr>
<tr>
<td>B</td>
<td>70</td>
</tr>
<tr>
<td>C</td>
<td>94</td>
</tr>
<tr>
<td>D</td>
<td>97</td>
</tr>
<tr>
<td>E</td>
<td>102</td>
</tr>
<tr>
<td>F</td>
<td>109</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Chronic kidney disease (CKD) is a significant public health problem, estimated to affect 843.6 million people, or 11.1% of the world population, in 2017 [20]. The two main causes are diabetes and high blood pressure [39]. It has many symptoms that decrease quality of life, which are shown in figure 1. CKD is associated with substantial morbidity, mortality and significant healthcare costs. It also increases risk of cardiovascular disease (CVD) and progression to end-stage renal disease (ESRD) [35, 63, 14]. Research has verified that prevention, early detection and treatment of complications can improve quality of life and survival, even if there is no effect on progression of CKD [30].

As a result, multiple national and international guidelines for the diagnosis and treatment of CKD have been created, however, adherence to these guidelines is unsatisfactory. Part of this may be explained by the fact that these guidelines focus on more than 10 goals simultaneously, which can be overwhelming to patients. Besides this, an average consultation by a nephrologist takes 10 minutes, which means it is infeasible to properly address each of the numerous goals. Recent trials suggest that additional coaching by nurse practitioners to improve treatment goal adherence could be beneficial for the life expectancy and quality of CKD patients [71].

The goal of this thesis is to answer the question:

In patients with CKD, does increased effort to eliminate high blood pressure, decrease proteinuria, increase ACE inhibitor or ARB use, and lessen salt intake by means of coaching by a nurse practitioner have a significant positive impact on the long term progression of kidney function?

To answer this question, we will use the MASTERPLAN dataset and a hypothetical mediation model to look at the effect of the nurse practitioner intervention on the occurrence of a composite renal endpoint a median follow-up of 5.7 years. The analysis will be limited to adult patients with category...
3-4 estimated glomerular filtration rate (eGFR) who reached the treatment goals of:

- blood pressure 130/85 mm Hg or lower,
- proteinuria 0.5 g per day or lower,
- ACE inhibitor or ARB e.g. enalapril 5 mg twice daily (or comparable dose of other ACE inhibitor) or irbesartan 75-150 mg (or comparable dose of other ARB) daily, and
- sodium intake 2g per day (or sodium excretion 100 mmol per day) or lower

Figure 1.1: Symptoms and signs of CKD [63].
within one year, versus patients who did not reach this goal. This composite renal endpoint consists of dialysis, transplant, 50% decrease of eGFR relative to baseline, and death.

We will construct four causal models in which the total effect of the nurse practitioner intervention is partitioned into the effect through the aforementioned treatment goals. Such models describe causation as opposed to mere association because of the assumptions made about the directions of the effects. To extract the size and significance of the mediators of the effect of the intervention on the composite endpoint and the effect itself, we will use mediation analysis.

It is important to note that we have chosen to construct these models with a combination of domain knowledge and the ‘guessing-and-testing’ method described by Shalizi [56], as opposed to the possibly more accurate method of using a discovery algorithm in the interest of time. Thus, the mediation analysis relies on the models we constructed which may be incorrect.

The partitioning of the effect into these treatment goals is done because previous scientific research has shown that control of these variables has a beneficial effect on CKD patients [22]. This is especially relevant because other research has merely shown that nurse practitioner support may benefit CKD patients, but not through which mechanisms this benefit is achieved. That is, the effect on endpoints and treatment goals is often measured but not the effect on endpoints through these separate treatment goals. The MASTERPLAN trial used several treatment goals and guidelines, but the effect of the intervention on outcomes was not broken down to analyse the influence of these individual goals and guidelines [72, 75, 47].

The mediation analysis performed in this thesis serves to shed more light on the mechanism through which this nurse practitioner intervention affects renal outcome in CKD patients. With this information, future applications of this intervention and even usual care can be improved by prioritising treatment goals that are found to contribute significantly to the effect of the nurse practitioner intervention on renal outcome.

Because this analysis relies on the causal models constructed in this thesis, which is partially based on hypotheses, these results may be incorrect despite our best efforts to build valid models. The method of model creation was employed because of time constraints. For the same reason we have chosen to limit the variables in the models to the blood pressure, proteinuria, ACE inhibitor or ARB use, and salt intake treatment goals, and to divide the DAG into four models, each containing only one mediator. Despite these limitations, this analysis still contributes to the body of knowledge, at the least as a stepping stone for a more comprehensive model.
1.1 Chapter overview

In chapter 2 we provide related background information on CKD, the MASTERPLAN study, mediation analysis and causal path diagrams. Chapter 3 comprises the methods we used for data preprocessing, data visualisation, generating causal path diagrams and their validation and statistical analyses using those diagrams. Then, in chapter 4 we describe the results of the main mediation analyses of the DAGs described in chapter 3. Chapter 5 includes the limitations and related work to this research. Finally, chapter 6 contains the conclusions of our work and speculation on future efforts that will improve and expand on our research.
Chapter 2

Preliminaries

In this chapter we will explain a number of concepts relevant to this research. We will give the definition of chronic kidney disease (CKD) in section 2.1. In section 2.2 we will describe the treatment goals and results of the MASTERPPLAN study. After this, in section 2.3 we will give a brief description of directed acyclic graphs (DAGs) and their implications. Then, in section 2.4 we will give an overview of the statistical analysis used in this thesis. Finally, in section 2.5 we will explain why and how we use causal mediation analysis.

2.1 Chronic Kidney Disease

In this paper, we will mainly consult the guideline published in 2012 by the Kidney Disease: Improving Global Outcomes (KDIGO) organization, which contains the definition, treatment recommendations and further information on CKD.

2.1.1 CKD Definition

CKD is defined as the presence of decreased GFR or a marker of kidney damage (or both) for longer than three months, according to the 2012 KDIGO guideline \[22\]. Table 2.1 shows some more details on this definition. In section 2.1.2 and 2.1.3 we will further explain glomerular filtration rate and albuminuria, which are the main criteria used in the classification of CKD. To give a short description, glomerular filtration rate gives an indication of the speed at which the kidneys filter the blood, while albuminuria gives an indication of how damaged the kidneys are.

2.1.2 Glomerular Filtration Rate

The glomerular filtration rate (GFR) is used as a measure for kidney function. It is the total filtration rate of the kidneys. The estimated GFR
Markers of kidney damage (one or more)

- Albuminuria (ACR ≥30 mg/24 hours; ACR ≥30 mg/g [≥3 mg/mmol])
- Urine sediment abnormalities
- Electrolyte and other abnormalities due to tubular disorders
- Abnormalities detected by histology
- Structural abnormalities detected by imaging
- History of kidney transplantation

Decreased GFR

GFR <60 ml/min/1.73 m² (GFR categories G3a–G5)

Table 2.1: Criteria for CKD (either decreased GFR or marker(s) present for >3 months) [22]

(eGFR) is, as the name says, an estimation of this filtration rate. It is calculated using the concentration of creatinine in the blood and is given in ml/min/1.73 m², where the 1.73 m² is a correction for body surface area, which is important for e.g. obese patients. Creatinine is created by the muscles and is a waste product that is normally filtered out of the blood by the kidneys. When there is a large amount of malfunction in the kidneys, a rise in blood creatinine levels can be observed. This happens with age and can vary among gender and other factors [40].

Calculating the eGFR with creatinine levels in the blood is preferred because it is not very invasive [40]. For this computation, the Modification of Diet in Renal Disease (MDRD) formula was used in the MASTERPLAN trial. However, research has shown that the eGFR becomes more inaccurate when the kidneys are more damaged [33]. In an attempt to minimize the effect of this inaccuracy, we use a relative decrease of 50% of eGFR as one of the endpoints instead of a set value. Because rapid progression is defined as a sustained decline in eGFR of more than 5 ml/min/1.73 m² per year, the relative decrease we use should capture this category for patients with GFR stage 3-5 with some room for the inaccuracy of the measurements.

2.1.3 Albuminuria

Albuminuria is also an indicator of kidney disease. When the kidneys malfunction, large proteins that should stay in the blood leak into the urine, which is called proteinuria. One of these proteins is albumin, which prevents fluid leakage from blood vessels among other things [61]. Albuminuria, abnormal leakage of albumin into the urine is used in the diagnosis of CKD because it is the most common protein lost in the urine in most cases of CKD. The albumin excretion rate (AER) is a measurement of mg albumin/24 hours in the urine. The 2012 KDIGO guideline recommends mea-
suring the albumin-to-creatinine ratio (ACR) in mg/g, which is the chosen unit for proteinuria in this thesis. [22].

### 2.1.4 CKD Classification

CKD classification is based on the GFR and ACR of a patient. A visual representation of the CKD classification is shown in the table “Prognosis of CKD by GFR and albuminuria category” in figure 2.1. The prognosis, or risk of CKD outcome such as progression or cardiovascular disease is indicated by the colours.

#### Figure 2.1: Prognosis of CKD by GFR and albuminuria category. Green: low risk (if no other markers of kidney disease, no CKD); Yellow: moderately increased risk; Orange: high risk; Red, very high risk of CKD outcome [22].

<table>
<thead>
<tr>
<th>GFR categories (ml/min/1.73 m²)</th>
<th>Persistent albuminuria categories Description and range</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1 Normal or high</td>
<td>&gt;=90</td>
</tr>
<tr>
<td>G2 Mildly decreased</td>
<td>60-89</td>
</tr>
<tr>
<td>G3a Mildly to moderately decreased</td>
<td>45-59</td>
</tr>
<tr>
<td>G3b Moderately to severely decreased</td>
<td>30-44</td>
</tr>
<tr>
<td>G4 Severely decreased</td>
<td>15-29</td>
</tr>
<tr>
<td>G5 Kidney failure</td>
<td>&lt;15</td>
</tr>
<tr>
<td>A1 Normal to mildly increased</td>
<td>Moderately increased</td>
</tr>
<tr>
<td>A2 Moderately increased</td>
<td>Severely increased</td>
</tr>
<tr>
<td>A3 &gt;30 mg/g &lt;3 mg/mmol</td>
<td>30-300 mg/g</td>
</tr>
<tr>
<td></td>
<td>3-30 mg/mmol</td>
</tr>
<tr>
<td></td>
<td>&gt;30 mg/g &gt;30 mg/mmol</td>
</tr>
</tbody>
</table>

The 2012 KDIGO guideline specifies a number of measures addressing cardiovascular health and CKD to prevent the progression of CKD. Control of blood pressure, reduction of proteinuria and lifestyle interventions including reduced daily sodium intake are critical in preventing CKD progression. Besides their impact on CKD progression, these three variables are also closely

9
related to each other [22]. Despite years of effort, the exact extent of the individual effects on CKD progression is still unknown and mostly consists of hypotheses. For this research we limit our scope to these three variables to stay within the timeframe given. Besides this, as mentioned before, they are known to be closely related and to have an impact on CKD progression. Next follows a brief overview of these individual variables and their relation to CKD.

CKD and elevated blood pressure are strongly associated and each can cause or aggravate the other [23]. Lowering blood pressure in CKD patients helps retard CKD progression. It is notable that this effect is only observable after long-term follow-up [47]. Specifically, meta-analysis showed that blood pressure control decreased the risk of end stage renal disease (ESRD) by 21% in CKD patients with proteinuria [63].

Proteinuria is also found in many CKD patients. Actively lowering proteinuria may decrease cardiovascular risk and acute kidney injury (AKI) risk for these patients [22]. The presence of proteinuria is also associated with increased risk of CKD progression, mortality and possibly increased risk of cancer [22]. The MDRD study in 1995 showed that reduction of proteinuria and decrease in rate of GFR decline were correlated [1]. Other research has identified proteinuria not just as a marker of CKD, but a contributor to the progression of the disease as well [9, 12].

Angiotensin-converting enzyme (ACE) inhibitors and Angiotensin II receptor blockers (ARBs) are important agents in lowering blood pressure and may even decrease proteinuria. Because of their effectiveness, using one or both of these medications is recommended in nearly all cases of CKD [23]. However, research has indicated that not many patients adhere to their prescriptions, which may be improved by means of additional coaching [72]. We refrain from explaining the effect of these medications in more detail in the interest of keeping this paragraph short, as this is less relevant to our thesis.

Impaired excretion of sodium is common in CKD patients as well. Lowering sodium intake for these patients decreases blood pressure and proteinuria [22]. The KDOQI 2002 and 2012 CKD guidelines therefore suggest lowering sodium intake to <100 mmol (<2 g) per day [31, 22]. Dutch guidelines on CKD have also recommended this or a lower maximum since 2004 [64, 15, 43, 44]. Despite these recommendations, many CKD patients exceed this daily limit. Italian, Dutch and Turkish studies showed that only 19%, 17% and 14.7%, respectively, of CKD patients met the target daily sodium intake [25].

2.1.6 CKD Complications
Chronically decreased kidney function can lead to a lot of problems, some of which are shown in figure 1.1. It also increases risk of cancer and cardiovascular disease (CVD), including stroke and heart attack [63]. The risk of
CVD is so severe that CKD patients are more likely to experience a heart attack than to progress to end-stage renal disease (ESRD). The kidneys of a patient with CKD that develops into ESRD function at 10-15% of their normal capacity. At this point, a patient would need dialysis or a kidney transplant to stay alive [41].

Even when CKD does not progress to such a severe state, patients are often in a lot of discomfort. The kidneys perform many important functions in the human body so when their performance decreases, multiple systems start to fail. Because of the multitude of different systems failing, CKD patients often get a multiple prescriptions, each with their own negative side effects. A restrictive diet is also required to limit the consumption of salt, protein, potassium, phosphorous, fat and fluids. Furthermore, many patients develop uremia, of which symptoms include nausea, vomiting, loss of appetite, tiredness and nerve problems [4]. On top of all of this, patients are advised to visit their doctor one to four times each year to get blood and urine measurements done to evaluate the progression of the disease [22].

2.2 The MASTERPLAN Study

In 2004, the Multifactorial approach and superior treatment efficacy in renal patients with the aid of nurse practitioners (MASTERPLAN) study was started in the Netherlands with 793 CKD patients. It was a multicenter randomized controlled clinical trial with patients of nine Dutch hospitals with a median follow-up of 5.7 years. The study was designed to evaluate whether a multifactorial approach with the aid of nurse practitioners reduces cardiovascular risk in patients with CKD [74].

The goal of the study was to compare the results of treatment executed by physicians, who were mostly nephrologists, with execution of the treatment regimen by a specialized nurse under the supervision of, and in collaboration with a nephrologist. For both of these categories, the same set of guidelines and treatment goals applied. These guidelines and treatment goals were based on several existing national and international guidelines on the treatment of CKD and were focused on reduction of vascular risk factors. It is notable that most of the national and international guidelines were updated or replaced during the study.

In total, there were 11 treatment goals in the MASTERPLAN study: blood pressure, proteinuria, LDL cholesterol, glycemic control, hemoglobin, serum phosphate, serum PTH, sodium excretion, BMI, and adherence to guidelines of healthy physical activity and not smoking. Besides these goals, there were also 4 medication types that were recommended by default: statins, acetylsalicyclic acid, ACE inhibitors or ARBs, and active vitamin D [73]. Because this thesis is confined to the effect of blood pressure, proteinuria, ACE inhibitor or ARB use, and sodium intake, we will solely consider
### Risk factors Goal

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood pressure</td>
<td>≤ 130/85 mm Hg*</td>
</tr>
<tr>
<td>Proteinuria</td>
<td>&lt;0.5 g/day</td>
</tr>
<tr>
<td>Sodium excretion</td>
<td>100 mmol/day</td>
</tr>
<tr>
<td>ACE-I or ARB use</td>
<td>e.g. enalapril 5 mg twice daily (or comparable dose of other ACE-I) or irbesartan 75-150 mg (or comparable dose of other ARB) daily</td>
</tr>
</tbody>
</table>

*In case of proteinuria > 1 g/day: 125/75 mm Hg

ACE-I = Angiotensin-converting enzyme inhibitor

ARB = Angiotensin II receptor blocker

---

Table 2.2: Treatment goals of the MASTERPLAN study relevant to this thesis [74].

<table>
<thead>
<tr>
<th>Renal endpoint</th>
<th>Control N</th>
<th>Control %</th>
<th>Intervention N</th>
<th>Intervention %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialysis</td>
<td>9</td>
<td>2.5</td>
<td>17</td>
<td>4.5</td>
</tr>
<tr>
<td>Kidney transplant</td>
<td>6</td>
<td>1.6</td>
<td>6</td>
<td>1.6</td>
</tr>
<tr>
<td>50% reduction in eGFR*</td>
<td>110</td>
<td>30.1</td>
<td>91</td>
<td>24.1</td>
</tr>
<tr>
<td>Death</td>
<td>36</td>
<td>9.9</td>
<td>29</td>
<td>7.7</td>
</tr>
<tr>
<td>Censored</td>
<td>204</td>
<td>55.9</td>
<td>234</td>
<td>62.1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>365</strong></td>
<td><strong>55.9</strong></td>
<td><strong>377</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Renal endpoint occurrences per treatment group in the context of this thesis.

The treatment goals related to these subjects. The relevant treatment goals are shown in table [2.2].

After two years, treatment goal adherence improved in both the control and intervention group, with some improvements more pronounced in the intervention group. The study showed that the aid of nurse practitioners in adhering to treatment goals benefits everyday care of CKD patients. Furthermore, during the first two years, the number of visits of the intervention group was higher than that of the control group, but the number of physician visits was lower. Another found benefit of the nurse practitioners is that the differences in the achievement of treatment goals between hospitals were reduced because adherence to guidelines was improved [72].

The original intention of the trial was to assess the effect of nurse practitioner intervention on cardiovascular outcomes, consisting of non-fatal stroke, non-fatal heart attack and fatal heart attack. The intervention did
not reduce the rate of the composite cardiovascular endpoint in the intervention group compared to the control group. A non-significant hazard ratio of 0.90 was found, which means that the intervention group had 0.9 times as much risk of a cardiovascular outcome as the control group. However, this hazard ratio was not significant, which means that it is likely invalid [75].

Another analysis was performed with the trial data that looked at a composite renal endpoint instead, which consisted of 50% increase in serum creatinine, end stage renal disease, and death. This study found a significant hazard ratio of 0.80, which means the intervention reduced incidence of the composite renal endpoint by 20%. The hazard ratio was significant, which means that this found decrease is likely valid [47].

The composition of the composite renal endpoint relevant to our analysis and the occurrence of each endpoint per treatment group are presented in table 2.3 as amounts and proportions. A composite endpoint was chosen to simplify the analysis and to emphasize the complexity of CKD progression.

2.3 Directed Acyclic Graphs

To perform mediation analysis, a number of assumptions are required about the relationships between variables in the context of the analysis. These assumptions can be visualised with a directed acyclic graph (DAG). These graphs are a type of causal path diagram and are used in mediation analysis to give a clear overview of the context, specifically they make the assumptions explicit [46]. To construct an accurate DAG, it is important to know about confounders, which cause a difference in the risk of the outcome in the exposed and unexposed populations that cannot be explained by the exposure [59]. Figure 2.2 is a simple example of a DAG with variables $X$, $M$, $Z$ and $Y$, where $M$ is a mediator of the relationship of $X$ on $Y$, and $Z$ is a confounder of the effect of $X$ on $Y$, $X$ on $M$, and $M$ on $Y$.

The layout of a DAG is as follows:

- Nodes represent variables existing of the exposure, outcome and other variables, or covariates.
- Directed edges represent relationships between variables; a direct path indicates the possibility of causal relevance, however, the absence of such a path indicates the absence of any causal relationship.

For a more detailed explanation of the composition of DAGs, we recommend chapter 2 and 3 of Tennant et al. [59].

Because DAGs are mere conceptual models built on hypotheses, a researcher can accidentally omit important confounders or interactions, which would mean that the analysis based on their DAG is invalid. Thus, the validity of our mediation analysis also depends strongly on the accuracy of our DAG.
Figure 2.2: A simple DAG where $M$ mediates the effect of $X$ on $Y$ and $Z$ is a confounder of the effect of $X$ on $Y$, $X$ on $M$, and $M$ on $Y$.

As mentioned before, two variables in a DAG that are not directly connected imply an independence. In figure 2.3, $A$ and $B$ are independent, when controlling for $M$. This means that if $M$ is held constant, there is no relationship between $A$ and $B$, which can be expressed as $A \perp \perp B \mid M$ (or $B \perp \perp A \mid M$). This independence can be falsified if the multiple regression formula $A \sim B + M$ produces a non-null parameter estimate. This indicates that $A$ and $B$ are related, even when controlling for $M$, thus the absence of the arrow is not justified. Regression will be explained in more detail in the next section.

Figure 2.3: A simple example DAG to illustrate conditional independency.

Thus a DAG produces a list of conditional independencies which can be falsified using regression and data. More specifically, this list is obtained using $d$-separation. For a short introduction to this theory, we suggest reading \textit{d-Separation Without Tears}. On a final note, the inability to falsify any implied conditional independency of a DAG does not mean it is valid, however, it does strengthen its plausibility.

In general, there are three ways to formulate a DAG: prior knowledge, guessing-and-testing, and discovery algorithms. The first is simply visualising known causal processes, however, it is still important to check this knowledge, which is done by the guess-and-test method. First build an initial DAG based on guesses, then find all implied conditional independencies using $d$-separation, and finally reject the DAG if not all conditional independencies hold. The final method requires an algorithm, such as the very basic
Spirtes, Glymour, and Scheines algorithm, which infers the DAG from data \[56\]. For our research, we will adhere to the guessing-and-testing method due to time constraints.

### 2.4 Edge Estimation

The next step after designing a DAG is calculating the edge weights and significance. This can be done with different types of regression, depending on the variables that are being tested.

#### 2.4.1 Multiple Imputation

Before moving on to edge estimation, it is necessary to find a solution for missing data. In trial data, it is nearly impossible to perform all measurements like blood pressure and albumin-creatinine ratio at every visit of every patient, which leads to incomplete data. One method to deal with this is to simply omit incomplete data, though this could introduce significant bias. For example, one could lose most data on patients who are more ill because they have a higher rate of missing variables than patients who are less ill. This introduction of bias is more severe depending on the type of missingness, often classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In MCAR, there are no systematic differences between the missing and the observed values, whereas in MAR there are systematic differences, however, they can be explained with other observed measurements. In MNAR, there are systematic differences as well, but these cannot be explained with other observed measurements. MAR is most frequently observed, though it should be noted that it is not possible to prove the presence of the MNAR pattern using data \[57\].

One solution to missing data is multiple imputation (MI), with which multiple plausible complete datasets are produced and the results of these datasets are combined using specific methods, taking the uncertainty of the imputed values into account. First, to produce the plausible datasets, missing variables are replaced by sampling from the predictive distribution of the observed data \[57\]. Common methods for imputation of numeric values and binomial values are predictive mean matching and logistic regression, respectively. Furthermore, the imputed value is also based on other parameters, or predictors. For further explanation of these methods, we recommend studying chapter 3 of Buuren \[6\]. When the imputed datasets are produced, it is possible to fit a chosen model to each dataset and find the overall parameter and variance estimations. This is where Rubin’s rules come into play to properly handle the variance introduced by the imputations \[53\].

To combine the results of several imputed datasets for measures like the regression coefficient and standard deviation, the average of the parameter...
estimates of all imputed datasets is taken, or in the form of a formula:
$$\hat{\theta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m$$
where $M$ is the number of imputed datasets and $\hat{\theta}_m$ is the parameter estimate of the regression for imputed dataset $m$. Obtaining the combined variance estimate of several imputed datasets is somewhat more complex, with the formula
$$\hat{\text{Var}}(\hat{\theta}_{MI}) = \text{within imputation variance} + \frac{1}{M} \text{between imputation variance}$$
In which the within imputation variance is expressed as
$$\frac{1}{M} \sum_{m=1}^{M} \hat{\text{Var}}(\hat{\theta}_m) = \frac{1}{M} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta}_m)^2$$
and the between imputation variance is expressed as
$$\frac{1}{M-1} \sum_{m=1}^{M} (\theta_m - \bar{\theta}_{MI})^2$$
where $\bar{\theta}_{MI}$ is the average of the parameter estimate over all $M$ imputed datasets [34, 55].

### 2.4.2 Linear Regression

A commonly used method to estimate edges pointing towards continuous variables is multiple linear regression, which attempts to express the effect of predictor variables on a response variable. This effect can be expressed in form of the mean function with $p$ predictor variables $(x_p)$: $E(y|x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ [65].

The linear regression intercept ($\beta_0$) indicates the value of $y$ when all predictor variables $(x_i)$ are equal to 0. Furthermore, when all other $x$ are held constant, a change of one unit in $x_i$ causes a change of $\beta_i$ (the logistic regression slope for $x_i$) in $y$ [65].

To compute the coefficients $\beta_i$ in this model, the ordinary least squares method can be used. This method minimises the difference between the true value of $y$ (in the dataset) and the estimated value of $y$ ($E(y|x_1, \ldots, x_p)$) as calculated by the model), where this difference for datapoint $i$ is expressed as $\epsilon_i = y_i - E(y_i|x_{i1}, \ldots, x_{ip})$. This minimising is often done with the ordinary least squares method, which finds the values $\hat{\beta}_0, \ldots, \hat{\beta}_p$ of $(\beta_0, \ldots, \beta_p)$ that minimise the residual sum of squares function, or $\text{RSS}(\beta_0, \ldots, \beta_p)$, given by $\text{RSS}(\beta_0, \ldots, \beta_1) = \sum_{i=1}^{n} (y_i - E(y_i|x_{i1}, \ldots, x_{ip}))^2 = \sum_{i=1}^{n} \epsilon_i^2$.

These calculations can be translated into R commands. Essentially, the computation of the edges of a DAG can be done by regressing a variable on all variables which have a direct outgoing arrow to it. For the simple graph in figure 2.2 the edges $X \rightarrow Y$, $M \rightarrow Y$, and $Z \rightarrow Y$ can be calculated by
regressing $Y$ on $X$, $M$ and $Z$. For a continuous $Y$ this can be done in R with the command `lm(Y ~ X + M + Z, data)`.

Because of the different units of most continuous variables, the obtained coefficients are not comparable. To enable comparing these coefficients, the independent continuous variables in a regression formula should be standardised. Standardising results in a variable that has a mean of 0 and a standard deviation of 1. This is achieved with a simple transformation of subtracting the mean of the variable and then dividing this by the standard deviation of the variable, and should be done for all three regression types mentioned in this chapter [19]. In R, the function `scale()` is used to perform this transformation.

### 2.4.3 Logistic Regression

Whereas linear regression can be used if the dependent variable is continuous, binary logistic regression can be used if the dependent variable is binomial. Binary logistic regression seeks to describe the effect of predictor variables ($x_i$) on the probability that the outcome variable ($Y$) takes on the value 1, also indicated as $\pi = P(Y = 1)$. Usually, a nonlinear transformation is applied to $\pi$ which maps its range from $(0, 1)$ to $(-\infty, \infty)$. A popular choice for this transformation is the logit function: $g(\pi) = \log(\frac{\pi}{1-\pi})$. After applying this transformation, the probabilities are related linearly to the predictor variables, such that: $\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

This is also known as the logistic regression model [12].

Because $\pi$ is known as the probability of success, $\frac{\pi}{1-\pi}$ is known as the odds of success. The logistic regression model can also be expressed as $\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}$, which emphasises the S-shaped relationship of $\pi$ with increasing values of $\beta_i x_i$ [12].

The logistic regression intercept ($\beta_0$) indicates the log odds when all predictor variables ($x_i$) are equal to 0. Additionally, the logistic regression slope for variable $x_i$ ($\beta_i$), indicates the amount the log odds changes when $x_i$ changes by one unit, given that all other variables remain constant. Note that these interpretations are analogue to the linear regression intercept and slope [12]. Furthermore, if $x_i$ increases by one unit while all other variables are held constant, the odds of success multiplicatively changes by a factor of $\exp(\beta_i)$.

To compute the coefficients $\beta_i$ in the model, the iterative reweighted least squares method is utilised. To dissect this method, first, the weighted least squares method minimises the sum of squared errors like the ordinary least squares method, however, each squared error is appointed a weight such that not all errors are of equal importance in this minimising problem. Then, the iterative reweighted least squares method finds the weights for which the sum of squared errors is minimised. A more mathematical description has been left out in the interest of keeping this section short, but can be found [17].
Performing this calculation in R requires a slightly different command from linear regression. Again, looking at figure 2.2 calculating the edges $X \rightarrow Y$, $M \rightarrow Y$, and $Z \rightarrow Y$ can be done by regressing $Y$ on $X$, $M$ and $Z$. Regressing a binomial $Y$ requires the command `glm(Y ~ X + M + Z, family=binomial, data)` in R.

In the context of a medical trial, the effects of an intervention or a mediator on the survival time can also be calculated with Cox regression instead of linear or logistic regression. We explain this concept in the next section.

2.4.4 Cox Models

A Cox proportional hazards regression model is a type of survival model which relates several exposures, like age or an intervention, considered simultaneously to survival time. This survival time is the time a patient lives until an event, such as death or disease recurrence, occurs. Such a model results in a hazard rate, which is the risk of a patient suffering from an event, given that they survived until a given time. Thus, the hazard represents the expected number of events per unit of time. To illustrate, a survival model could be used to inspect the effect of a new medicine on the lifespan of patients of a certain disease. We will utilise Cox regression to find the effect of the nurse practitioner intervention and the four aforementioned individual treatment goals on the time it takes for the composite renal endpoint to occur. Often, the hazard ratio (HR) is the measure of interest. It is the ratio of the total number of observed to expected events in two different groups. This could be the treatment and the control groups in a trial, for example 26.

The Cox regression model thus calculates the expected hazard at time $t$, by taking the product of the baseline hazard and the exponential function of the linear combination of the exposures. The baseline hazard is the hazard when all exposures are equal to 0. The model can be written as:

$$h(t) = h_0(t) \exp(b_1X_1 + b_2X_2 + \cdots + b_pX_p)$$

which means that the exposures have a multiplicative effect on the expected hazard. These coefficients for the exposures are in units of log-HR, which makes them difficult to interpret. This is why they are exponentiated to then obtain the HR of the individual exposures. These, in turn, can be interpreted as the expected hazard increasing by a factor of this coefficient for every single unit increase of the corresponding exposure, while keeping all other exposures constant 26. The method used to calculate these coefficients was described by Cox and Oakes 8 and has been excluded from this thesis to conserve the readability of this paragraph. For an applied example of this theory, we suggest looking at LaMorte 26.
As with the other regression types, in the context of figure 2.2, the edges $X \rightarrow Y$, $M \rightarrow Y$, and $Z \rightarrow Y$ can be computed by performing Cox proportional hazards regression with the R command `coxph(Surv(time,status) ~ X + M + Z, data)` from the survival library. This `Surv(time,status)` command creates a survival object. The status is often either 1 or 2, where 1 means the subject was lost to follow-up or did not experience an event before there was no more data recorded, also known as censored, and 2 means a predefined event happened. The time is the last time that data was recorded for the subject, where recording stops after an event has occurred.

For the final model in this thesis, it is important to note that logistic regression estimates the odds ratio (association estimate), while Cox regression estimates the hazard ratio (time to event estimate), where probability is not the same thing as hazard. In more informal terms, logistic and linear regression on survival data assesses a proportion, while Cox regression estimates a rate [29].

### 2.4.5 Edge Significance

Because these coefficients are mere estimations, a measure of significance is desired. This is where the confidence interval (CI) comes into play. A 95% CI indicates the upper and lower limits between which the estimated coefficient lies with 95% certainty. If this CI includes 0, there is not a strong indication that the coefficient is significant [49]. To illustrate, imagine that the estimated edge coefficient, or parameter estimate, of the edge from $X$ to $Y$ in figure 2.2 is 0.38 with a 95% CI lower bound of 0.30 and an upper bound of 0.48, also indicated as (0.30 − 0.48). This would indicate that the effect of $X$ on $Y$ is significant (0 is not included in the CI) and is estimated to be of size 0.38.

To obtain this 95% CI, the bootstrap estimate of the standard error (SE) is required, with which the 95% CI can be expressed as

$$\theta = (\bar{T} - \hat{B}^*) \pm 1.96 \times \hat{SE}^*(T^*)$$

where the coefficient 1.96 specifically belongs to a 95% CI. To dissect this formula, first the average of the bootstrapped parameter estimate is

$$\bar{T}^* = \hat{E}^*(T^*) = \frac{1}{B} \sum_{b=1}^{B} T_b^*$$

where $B$ is the number of bootstrap samples. Then $\hat{B}^* = \bar{T}^* - T$ is an estimate of the bias of $T$. The next step is the estimated bootstrap variance of $T^*$:

$$\hat{Var}^*(T^*) = \frac{1}{B - 1} \sum_{b=1}^{B} (T_b^* - \bar{T}^*)^2$$
Finally, the bootstrap estimate of the SE of \( T \) is expressed as
\[
\hat{SE}^*(T^*) = \sqrt{\hat{Var}^*(T^*)}
\]
In the notation above, \( Y^* \) is used to indicate bootstrap variables and \( Y \) is used to denote original sample variables \cite{13}. For this thesis, we use a simplified version of the formula for the confidence interval mentioned above, where the bias is left out of the calculation, resulting in the formula \( \theta = T \pm 1.96 \times \hat{SE}^*(T^*) \).

The different samples for the bootstrap are usually chosen “with replacement”, which means that in a dataset with 30 datapoints, the 10th datapoint may occur 7 times within a bootstrap sample, while the 29th datapoint may not be used at all in that same sample \cite{68}.

For this thesis, we use bootstrapping within the imputed datasets, to obtain a variance estimate for each estimated coefficient. For this, we use the “MI Boot” method, as described by Schomaker and Heumann \cite{55}. First, multiple imputation is used to generate \( M \) imputed datasets, where for each dataset the parameter estimates are made, or in our case, the regression coefficients are computed which can then be used to retrieve the overall parameter estimate using Rubin’s rules. Then, from each dataset \( B \) bootstrap samples are drawn which are then used to calculate the variance estimate of the regression for each imputed dataset, which in turn can be used to obtain the overall variance estimate using Rubin’s rules. The variances of the regression coefficients for imputed dataset \( m \) can be obtained with
\[
\hat{Var}(\hat{\theta}_m) = \frac{1}{B-1} \sum_{b=1}^B B(\hat{\theta}_{m,b} - \hat{\bar{\theta}}_m)^2
\]
where \( \hat{\bar{\theta}}_m = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_{m,b} \) \cite{55}.

2.5 Mediation Analysis

Mediation analysis is a method to make causal inferences about the effect of an exposure on an outcome through mediators. We will use this method to study the mediation by the four treatment goals mentioned earlier of the effect of the nurse practitioner intervention on the incidence of the renal endpoint.

Often, a visual representation in the form of a directed acyclic graph (DAG) is used to show the assumptions a researcher made in the analysis \cite{48}. To give a simple example of how this method would work in practice, imagine a researcher wants to know the effect of an antidepressant on depression and how this effect is mediated by insomnia, a common side effect of antidepressants. Thus, the antidepressant would be the exposure, insomnia the mediator and depression the outcome. The DAG would look like the
one in figure 2.4. For an example DAG that is more similar to ours with repeated measures, we refer to section 14.7 of Hayes [17] in the interest of keeping this explanation short.

Figure 2.4: A simple mediation model for the relation between an antidepressant, insomnia and depression.

Because of the assumptions made in the construction of the mediation model, the researcher can then uncover the effect of the antidepressant on depression through insomnia (indirect effect) and not through insomnia (direct effect). With these indirect and direct effects, they could calculate the total effect of the antidepressant on depression [52].

As mentioned before, we use mediation analysis to partition the effect of the nurse practitioner intervention on the time it takes for the composite renal endpoint to occur in CKD patients through the four treatment goals mentioned earlier. A concept version of this mediation model is shown in figure 2.5

2.5.1 Potential Outcomes Framework

Mediation analysis is mainly about estimating the direct and indirect effects. One of the methods to obtain these estimates makes use of the potential outcomes (PO) framework, which builds upon the concept of “opposite worlds” in which the observations of the experiment are different from the true observations. In this framework, $Y_i(a,m)$ is the counterfactual outcome for person $i$ if exposure was set to $a$ and mediator to $m$, and $M_i(a)$ is the value assumed by the mediator for person $i$ if exposure had been set to $a$. If $a$ is the exposure, $a^*$ is the opposite exposure. Furthermore, $E[Y(a,m)]$ is the
average counterfactual outcome if each individual had been introduced to exposure \( a \) and mediator \( m \) \[28\].

For an example application, imagine a trial where depressed individuals are randomly appointed to receive \((A = 1)\) or not receive \((A = 0)\) an antidepressant over a time period and at the end, a score of the severity of their depression \((Y)\) and presence of insomnia \((M = \{0, 1\})\) is measured. The suspected relationships between the variables is shown in figure 2.4. If we apply the theory above within the context of this example, \( E[Y(1, 1)] \) denotes the average depression score if all patients would have received the antidepressant and had experienced insomnia.

With this theory, it is possible to express the different indirect and direct effects shown in figure 2.4. In these formulas, \( X \) is the exposure, \( M \) is the mediator and \( Y \) is the outcome \[28\].

- The natural direct effect is defined as the difference between the counterfactual outcome if \( i \) were exposed to \( A = a \) and the counterfactual outcome if \( i \) were exposed to \( A = a^* \), while keeping \( M \) equal to \( M = M(a^*) \).
  
  \[
  \text{natural direct effect} = E[Y(1, M(0))] - E[Y(0, M(0))]
  \]

- The natural indirect effect is defined as the difference between the counterfactual outcome if \( i \) were exposed to \( M = M(a) \) and the coun-
terfactual outcome if \( i \) were exposed to \( M = M(a*) \), while keeping \( A \) equal to \( A = a \).

natural indirect effect = \( E[Y(1, M(1))] - E[Y(1, M(0))] \)

- The total effect is defined as the difference between the counterfactual outcome if \( i \) were exposed to \( A = a \) and \( M = M(a) \) and the counterfactual outcome if \( i \) were exposed to \( A = a^* \) and \( M = M(a^*) \).

\[
\text{total effect} = E[Y(1, M(1))] - E[Y(0, M(0))] = (E[Y(1, M(1))] - E[Y(1, M(0))]) + (E[Y(1, M(0))] - E[Y(0, M(0))]) = \text{natural indirect effect} + \text{natural direct effect}
\]

It is of importance that some outcomes cannot be observed, like \( E[Y(1, M(0))] \), which is the average outcome when exposure \( X \) is applied, but when mediator \( M \) takes on the value that would have manifested if the exposure had not been applied. Moreover, \( Y(1, M(1)) \) cannot be observed in subjects that do not receive exposure \( X \). This means that mediation analysis in the PO framework requires imputation of unobserved outcomes as described by Lange et al. \[28\], of which we give a brief overview in the next section.

In practice, DAGs for mediation analysis are much more complex than the simple three-variable model from our illustration. The calculations get increasingly large very quickly when more variables and interactions are added, which is discussed in Taguri, Featherstone, and Cheng \[58\] and Miocenski et al. \[38\], to name just a few examples.

In general, the direct effect of the exposure on the outcome is the combination of all its indirect effects through mediators that are not explicitly modeled. In the context of our model, this means that the direct effect of the nurse practitioner intervention on the survival time is the combination of all indirect effects mediated by the other treatment goals and guidelines that are not present in our model.

### 2.5.2 Effect Estimation

In essence, mediation analysis based on the PO framework is a problem of regression with missing data. Because there is only observed data, the counterfactual data are missing. This means that finding the indirect and direct effects with this method requires imputation, which was used as well in section 2.4.1, albeit with a few extra steps. For a mediation model with a single mediator (see figure 2.2, a copy of figure 2.2 in section 2.3), the method of Lange et al. \[28\] can be applied. It consists of 6 steps:

- **Step 1**: Fit a survival model to the survival times using the treatment group, mediators and confounders.

- **Step 2**: Duplicate the original dataset twice and add a new counterfactual variable for the intervention. In the first duplication, this new
variable assumes the observed value, while in the second duplication, it assumes the opposite variable (in the case of a dichotomous exposure).

- **Step 3**: Impute the unknown survival times for the new counterfactuals created in the previous step.

- **Step 4**: Fit a Cox model to the extended dataset by regressing the imputed survival times on the confounders and newly produced counterfactual exposures.

- **Step 5**: Repeat step 3 and 4 10 times and pool the obtained coefficients.

- **Step 6**: Repeat step 1-5 1,000 times while taking a new bootstrap sample each time to obtain the significance of the obtained coefficients.

After performing these steps, the coefficient for the actual exposure will estimate the natural indirect effect, while the coefficient for the created exposure variable will estimate the natural direct effect. Furthermore, as these are results of a Cox proportional hazards regression, the effects will be expressed in log-HR. To make these more readable, the effect estimates are exponentiated, so that the expected hazard for an effect increases by a factor of the found amount for every one unit increase in the effect, while holding all other effects constant [26]. After this exponentiation, a HR of >1 means that the effect increases risk of an event happening, while a HR of <1 indicates the effect reduces the chance of an event happening.

Of note is that this method does not impute the counterfactual mediator. This solution is achieved based on four assumptions about the relationships between the outcome, intervention, mediator(s), and confounders and a mathematical proof based on these assumptions, which are described by Lange et al. [28].

As with the edge estimations, in the context of our thesis, there are 34 imputed datasets to begin with to handle the missing data. This means that an extra step of pooling is required after finishing the procedure described above, which can be done with the help of Rubin’s rules again.
Chapter 3

Method

In this chapter we will describe the method of our research. First, we will report the data preprocessing performed on the MASTERPLAN dataset and visualise this dataset for some insight into the data it contains. Then, we will show our efforts to validate the final DAG. Finally, we will describe the statistical analyses with which the edge weights and significance were calculated in our final DAG.

The code used for this thesis can be found at https://github.com/LESchernthaner/ThesisCode.

3.1 Preprocessing

The preprocessing was performed in R, using the Foreign (import .dta), Dplyr (data manipulation), Ggplot2 (plots), GridExtra (plot arrangements), Rmarkdown (reporting), Knitr (reporting), Mice (imputation), Dagitty (DAG handling), Writexl (xlsx reading and writing), Lubridate (time casting), Boot (bootstrapping), Survival (Cox regression), Survminer (Cox plots) and Amelia (imputation pooling) packages.

We performed a short analysis of the variables in the MASTERPLAN dataset relevant to this thesis to gain insight in their distributions and extreme values. We noticed the values of albumin-creatinine ratio on the first visit were very high compared to the other visits in both treatment groups. This may be explained by the fact that the start of the trial gave incentive to the doctors to properly manage it. There were also two extreme BMI measurements; one below 15 and one above 50. They were not excluded from the dataset, because comparison with the waist-hip ratio implied that these values were plausible. Finally, there was one patient with a low diastolic blood pressure (<50), which was also not seen as an outlier, as this was consistently present in this patient. The full analysis can be found in appendix A.
After this, we imputed the missing values of the baseline and one-year check-up visits. This imputation was done to obtain more complete data for the final mediation analysis. The one-year check-ups were set to be the fifth visit of each patient. However, this was not always realised, as some patients had this check-up on their fourth visit, for example. We tried to reduce this effect with a special selection procedure. If the fifth visit of a patient contained more than 8 missing variables, we selected the fourth visit if it contained less than 8 missing variables. If this also did not hold, we selected the sixth visit if it contained less than 8 missing variables, and if this also did not hold, the fifth visit was selected after all to prevent including much earlier or later visits.

The data was transformed to a long format with the baseline measurements, first visit measurements and fifth visit measurements, obtaining a dataset of 742 variables. The most common pattern was no missing variables (343 times), and after that, missing proteinuria on visit 5 (118 times).

The variables that were missing most often were proteinuria (34%), urine sodium (10%) and albumin-creatinine ratio (8%), all at visit 5. We generated 34 imputed datasets, as the most frequently missing variable (proteinuria on visit 5) was missing in 34% of the data. This variable could not be dropped, as it is part of our mediation model, which is why we took the precaution of producing an equal amount of imputed datasets. Usually, a cutoff value of 40% missingness is given, and a recent publication even suggests that the amount of missingness can be higher [32]. In a short analysis of the imputation, we found good convergence and reasonable distributions of the imputed variables. Throughout our analysis, we used the same 34 imputed datasets by using the same seed, imputation methods and predictor matrix. The full missing data analysis, the imputation and the analysis of the imputed data are shown in appendix B.

### 3.2 Data Visualisation

To provide some insight into the MASTERPLAN dataset, two tables were constructed. Table 3.1 contains an overview of the baseline characteristics of the 788 randomised patients and table 3.2 gives a summary of the variables of the altered dataset mentioned in the previous section, which contains data of the first and fifth visits of the 742 patients who were still participating during the fifth visit. In the final analyses, we further filter the dataset to only contain patients with stage 3-4 GFR, obtaining data with 340 patients in the control group and 334 in the intervention group.
<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>393</td>
<td>395</td>
</tr>
<tr>
<td>Age (years)</td>
<td>59.3 (12.8)</td>
<td>58.9 (13.1)</td>
</tr>
<tr>
<td>Gender (male) (%)</td>
<td>68</td>
<td>67</td>
</tr>
<tr>
<td>Race (Caucasian) (%)</td>
<td>93</td>
<td>91</td>
</tr>
<tr>
<td>Kidney transplantation (%)</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Prior CVD (%)</td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td>History of DM (%)</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>Smoking (%)</td>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>27.2 (4.9)</td>
<td>27.0 (4.6)</td>
</tr>
<tr>
<td>eGFR (ml/min/1.73 m²)</td>
<td>35.6 (13.3)</td>
<td>36.3 (14.5)</td>
</tr>
<tr>
<td>Systolic BP (mm Hg)</td>
<td>136 (21)</td>
<td>135 (20)</td>
</tr>
<tr>
<td>Diastolic BP (mm Hg)</td>
<td>79 (11)</td>
<td>78 (11)</td>
</tr>
<tr>
<td>Proteinuria (g/24 hr)</td>
<td>0.3 [0.1-0.8]</td>
<td>0.2 [0.1-0.8]</td>
</tr>
<tr>
<td>Albumin-creatinine ratio (mg/g)</td>
<td>138.8 [28.8-496.6]</td>
<td>111.4 [30.8-416.3]</td>
</tr>
<tr>
<td>Sodium excretion (mmol/24 hr)</td>
<td>150 [113-189]</td>
<td>148 [116-194]</td>
</tr>
<tr>
<td>ACE-I or ARB use (%)</td>
<td>78</td>
<td>81</td>
</tr>
</tbody>
</table>

Data are given as proportions, means with corresponding standard deviation, or median with 25th and 75th quartiles.

CVD = cardiovascular disease
DM = diabetes mellitus
BMI = body mass index
eGFR = estimated glomerular filtration rate
BP = blood pressure
ACE-I = angiotensin-converting enzyme inhibitors
ARB = Angiotensin II receptor blockers

Table 3.1: Baseline characteristics of the MASTERPLAN dataset.
<table>
<thead>
<tr>
<th></th>
<th>Baseline Control</th>
<th>Baseline Intervention</th>
<th>Year 1 Control</th>
<th>Year 1 Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>365</td>
<td>377</td>
<td>365</td>
<td>377</td>
</tr>
<tr>
<td>eGFR (ml/min/1.73m²)</td>
<td>36.1 (13.1)</td>
<td>36.7 (14.5)</td>
<td>34.2 (14.1)</td>
<td>35.0 (14.8)</td>
</tr>
<tr>
<td>Systolic BP (mm Hg)</td>
<td>135 (20)</td>
<td>135 (20)</td>
<td>133 (20)</td>
<td>130 (19)</td>
</tr>
<tr>
<td>Diastolic BP (mm Hg)</td>
<td>79 (11)</td>
<td>78 (11)</td>
<td>78 (11)</td>
<td>76 (10)</td>
</tr>
<tr>
<td>Proteinuria (g/24 hr)</td>
<td>0.3 [0.1-0.8]</td>
<td>0.2 [0.1-0.8]</td>
<td>0.3 [0.1-1.0]</td>
<td>0.3 [0.1-0.9]</td>
</tr>
<tr>
<td>Albumin-creatinine ratio (mg/g)</td>
<td>133.5 [28.0-439.7]</td>
<td>110.1 [30.0-406.5]</td>
<td>24.4 [8.6-77.6]</td>
<td>20.2 [6.0-57.3]</td>
</tr>
<tr>
<td>ACE-I or ARB use (%)</td>
<td>78</td>
<td>81</td>
<td>84</td>
<td>92</td>
</tr>
</tbody>
</table>

Data are given as proportions, means with corresponding standard deviation, or median with 25th and 75th quartiles.

eGFR = estimated glomerular filtration rate
BP = blood pressure
ACE-I = angiotensin-converting enzyme inhibitors
ARB = Angiotensin II receptor blockers

Table 3.2: Characteristics of the dataset constructed in the previous section
Figure 3.1: The DAG upon which all other DAGs were built. The numbers with two digits are the estimated edge weights and the numbers between brackets are the lower and upper bounds of the 95% confidence interval.

3.3 Final DAGs

We constructed the final DAGs for the main analysis of this thesis with the help of the 2012 KDIGO guideline [22] and the MASTERPLAN thesis [72]. These works contain information about the relations of the variables used in our model. Some are hypothesised, while others are supported with empirical proof. This means that we utilised the method of ‘guessing-and-testing’ to construct a DAG based on these hypotheses [56]. Furthermore, we split the DAG with the treatment goals on T1 up into four smaller DAGs, which does not allow us to capture the interactions between the variables. However, this choice was necessary to perform the final mediation analysis, as it would have been beyond the scope of this thesis to carry out a mediation analysis with four interacting variables. To build the individual DAGs, the graph in figure 3.1 was used as a base. To clarify, common variable names and their corresponding methods of measurement and units are listed below. Besides this, all continuous independent variables were standardised before performing this analysis, which means that the edge weights show their size relative to other edge weights, as opposed to the increase in the dependent variable based on an increase of one unit of the independent variable.

- **Blood pressure**: systolic blood pressure measured over 30 minutes (mm Hg)
- **Proteinuria**: urinary protein concentration (mg/24 hr)
- **ACE inhibitor or ARB use**: use of angiotensin-converting-enzyme inhibitors or angiotensin II receptor blockers (true/false)
- **Salt intake**: urinary sodium excretion (mmol/24 hr)
• **eGFR**: estimated glomerular filtration rate with the MDRD<sub>175</sub> formula (ml/min/1.73m<sup>2</sup>)

• **Intervention**: addition of a nurse practitioner to the care team or standard treatment (true/false)

• **(Renal) composite endpoint / outcomes**: composite renal endpoint of death, 50% decrease in eGFR relative to baseline, dialysis, and kidney transplant (true/false)

To validate the DAGs that we made, we first tested the implications of each DAG. This test was performed by extracting the conditional independencies from a DAG with the `impliedConditionalIndependencies` function of the Dagitty package, and using the MASTERPLAN dataset to validate these independencies. It is important to perform these tests properly when using imputed data, as is the case for this thesis. We used the `pool` function of the Mice package to do this. This function first tests a conditional independency, that is, it estimates a regression coefficient, with each of the 34 imputed datasets and then finds the average estimate and calculates the total variance over the repeated analyses, according to Rubin’s rules. The script for these tests is presented in appendix D and the model codes of each DAG can be found in appendix C.

After this, we used our combined knowledge about mediation analysis and CKD to perform a face validation of the DAG. Because a lot of the mechanisms are hypothesized but not known, it was only possible to identify relationships in our DAG that have been disproven, or missing relationships that have been proven. Thus, this step was another attempt at falsification, rather than verification. Because of these uncertainties, it is possible that our model is incorrect. However, our falsification attempts failed, which increases the plausibility of our model. Figure 3.2 contains the hypothetical DAGs we constructed.

### 3.4 Statistical Analyses

To compute the edge weights, we used linear, logistic or Cox proportional hazards regression depending on the dependent variable. For the outcome variable (ESRD), Cox regression was utilised, while binary dependent variables required logistic regression. All other edges were computed by means of linear regression. Like with the independency tests, it is important to use the correct method for these estimations and their significance with imputed datasets. We employed the ‘MI Boot’ method described by Schomaker and Heumann [55] in which the missing data is first imputed and after this, each imputed dataset is bootstrapped, where all regressions are performed on each bootstrap sample. For correctly pooling the results, a manual implementation of Rubin’s rules was built. We generated 34 imputed datasets
and used 5,000 bootstrap samples for each imputed dataset. The script for these estimations can be found in appendix E and the results are shown in figures 3.1 and 3.2.

From these figures, it is clear that not all hypothesised edges are significant, yet leaving them out resulted in invalidation of the implications. This indicates that there are other mechanisms through which the effects of the variables are established. Because of the limited time available for this research, we were not able to find the optimal DAGs in which all edges are significant. Despite this limitation, mediation analysis could still be performed as the DAGs do show at least a limited direct or indirect effect of each of the baseline variables on the occurrence of ESRD. Since the method we use for the final mediation analyses includes the baseline values of blood pressure, proteinuria, ACE inhibitor or ARB use, and sodium excretion as possible confounders for each of the four analyses, the insignificance of some of the edges in the aforementioned figures is not detrimental to the validity of the results of the analyses. This insignificance does, however, show that our hypotheses are not completely correct.
Figure 3.2: The DAGs we constructed in the context of this thesis. In order top to bottom, the DAGs contain the blood pressure, proteinuria, ACE inhibitor or ARB use, and sodium excretion treatment goals as mediators. The red arrows indicate the indirect effect, while the blue arrows indicate the direct effect. The numbers with two digits are the estimated edge weights and the numbers between brackets are the lower and upper bounds of the 95% confidence interval. The estimates of the edges in the base DAG can be found in figure 3.1.
Figure 3.2: (cont.) The DAGs we constructed in the context of this thesis. In order from top to bottom, the DAGs contain the blood pressure, proteinuria, ACE inhibitor or ARB use, and sodium excretion treatment goals as mediators. The red arrows indicate the indirect effect, while the blue arrows indicate the direct effect. The numbers with two digits are the estimated edge weights and the numbers between brackets are the lower and upper bounds of the 95% confidence interval. The estimates of the edges in the base DAG can be found in figure 3.1.
Chapter 4

Mediation Analysis

The final part of our analysis exists of the computation of the direct and individual indirect effects of the nurse practitioner intervention on CKD patient survival time. As explained before, using the potential outcomes (PO) framework, mediation analysis can be reduced to a multiple regression problem utilising imputation.

To obtain the size and significance of the indirect, direct and total effects of each DAG, we followed the method of Lange et al. \[28\] applied to the 34 imputed datasets. This means that each of the 34 imputed datasets was bootstrapped 1,000 times and for each bootstrap sample, the counterfactual outcomes and survival times were imputed. Like the other analyses on these imputed datasets, the results should be pooled in accordance with Rubin’s rules. As mentioned before, the script for this process is based on the code written by Lange et al. \[28\] and can be found in appendix \[F\]. The result of the mediation analyses of the individual DAGs can be found in 4.1. We now shortly discuss our observations of this end result.

First, blood pressure does not appear to have a significant mediating role (HR 0.95, 95% CI 0.78, 1.17) on the impact of the intervention, which may seem unlikely. However, this could be explained by the fact that all doctors were informed of the treatment goals, thus adherence increased in both the treatment and control group (see table 3.2). Despite this insignificant mediating effect, it may still be beneficial to pursue adherence to the blood pressure treatment goal, as high blood pressure has been closely related to cardiovascular events and CKD progression \[23\].

Besides this, proteinuria, does not act as a significant mediator either (HR 0.99, 95% CI 0.81, 1.22). This might be explained by the fact that many patients already met the relevant treatment goal at the start of the trial, and by the small difference between the treatment and intervention group after one year of follow-up (see table 3.2 \[72\]. However, similar to blood pressure, increased effort to decrease proteinuria could still be valuable to CKD patients, as it is strongly associated with CKD progression \[22\] and
<table>
<thead>
<tr>
<th>Effect</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower limit</td>
</tr>
<tr>
<td>Natural indirect</td>
<td>0.95</td>
<td>0.78</td>
</tr>
<tr>
<td>Natural direct</td>
<td>0.82</td>
<td>0.67</td>
</tr>
<tr>
<td>Total</td>
<td>0.79</td>
<td>0.57</td>
</tr>
<tr>
<td>Mediated proportion</td>
<td>0.35</td>
<td>-1.92</td>
</tr>
</tbody>
</table>

(a) Blood pressure

<table>
<thead>
<tr>
<th>Effect</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower limit</td>
</tr>
<tr>
<td>Natural indirect</td>
<td>0.99</td>
<td>0.81</td>
</tr>
<tr>
<td>Natural direct</td>
<td>0.79</td>
<td>0.65</td>
</tr>
<tr>
<td>Total</td>
<td>0.79</td>
<td>0.57</td>
</tr>
<tr>
<td>Mediated proportion</td>
<td>0.11</td>
<td>-1.96</td>
</tr>
</tbody>
</table>

(b) Proteinuria

Table 4.1: Results of the mediation analysis for each of the four individual DAGs. In order from left to right, the mediators are the blood pressure, proteinuria, ACE inhibitor or ARB use, and sodium excretion treatment goals.

HR = hazard ratio  
CI = confidence interval

even plays an active role in this process [10].

Closely related to proteinuria and blood pressure management is ACE inhibitor or ARB use. However, the mediation analysis does not show a significant effect (HR 0.98, 95% CI 0.81, 1.20) for this treatment. This absence of an effect may again be due to the design of the MASTERPLAN study, as all doctors were made aware of the recommendation to prescribe these medications, which resulted in a significant increase in usage in both treatment groups, such that almost every patient received a prescription (see table 3.2). Thus, it may still be beneficial to recommend ACE-inhibitor and ARB usage to CKD patients, as the difference between the groups was small but nearly all patients used them.

Like the other treatment goals discussed above, salt intake, or sodium excretion, is an insignificant mediator (HR 1.01, 95% CI 0.83, 1.23) of the effect of the nurse practitioner intervention on the survival time, and unlike the others has an average HR of above 1. This could have been anticipated,
### Table 4.1: (cont.) Results of the mediation analysis for each of the four individual DAGs. In order from left to right, the mediators are the blood pressure, proteinuria, ACE inhibitor or ARB use, and sodium excretion treatment goals.

**HR = hazard ratio**  
**CI = confidence interval**

<table>
<thead>
<tr>
<th>Effect</th>
<th>HR (95% CI)</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural indirect</td>
<td>0.98 (0.81, 1.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural direct</td>
<td>0.80 (0.66, 0.98)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.79 (0.57, 1.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mediated proportion</td>
<td>0.14 (-1.92, 2.21)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**(c) ACE inhibitor or ARB use**

<table>
<thead>
<tr>
<th>Effect</th>
<th>HR (95% CI)</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural indirect</td>
<td>1.01 (0.83, 1.23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural direct</td>
<td>0.78 (0.64, 0.96)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.79 (0.57, 1.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mediated proportion</td>
<td>0.02 (-2.00, 2.05)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**(d) Sodium excretion**

As there was barely any reduction, or rather even an increase in the measured sodium excretion in both treatment groups (see table 3.2). This absence of change may be caused by the great difficulty found in restricting dietary sodium intake and lack of appropriate support strategies [37]. However, research has indicated the severe effect of excessive salt consumption, which means that it may be worthwhile to find alternative, more effective coaching methods to reduce salt intake [24].

Of course, there are more treatment goals and guidelines than these four, which are aggregated into the direct effect in each graph. Because these direct effects may include mediators that interact with the explicitly modelled mediator, it is not possible to conclude anything on these treatment goals.

Each subtable of table 4.1 not only shows the effects, but also the mediated proportion of the outcome by the individual mediators. Normally, these would be summable, resulting in a mediated proportion of 0.6 by the
blood pressure, proteinuria, and ACE inhibitor or ARB use treatment goals. However, as described in chapter 2, there is interaction between these treatment goals, which means that the mediated proportions found are higher than they are in reality.

Finally, the total effect of the nurse practitioner intervention does not appear to be significant either (HR 0.79, 95% CI 0.58, 1.08), which does not agree with the results of Peeters et al. [47], who found a HR for the intervention on renal outcomes of 0.80 (95% CI 0.66, 0.98), while Zuilen et al. [75] found a HR of 0.83 (95% CI 0.57, 1.20) for the intervention on the incidence of ESRD, though this was found with data of a shorter follow-up period of 4.6 years, as opposed to a median follow-up of 5.7 years. Besides this, we limited our dataset to only include patients with stage 3-4 GFR and adjusted for baseline systolic blood pressure, proteinuria, ACE inhibitor or ARB use, sodium excretion, and eGFR, while Peeters et al. [47] did not filter the dataset and only adjusted for baseline serum creatinine. Another dissimilarity that could explain the difference in results is the fact that we used multiple imputation to handle missing data and utilised mediation analysis, while the two previously mentioned papers do not employ either of these methods. Further research based on improved models should be performed to analyse these differences.
Chapter 5

Discussion

In this chapter, we will first discuss the limitations of the data and methods used in this thesis. After this we will describe previous research related to ours and demonstrate that this thesis has resulted in new knowledge in the field of CKD.

5.1 Limitations

Mediation analysis relies on the assumption of a complete model. Any unmeasured or unaccounted for confounders will decrease the accuracy of a model. Despite our efforts to increase the plausibility of our model, it is still possible that our model is incomplete. The method we used for the construction of the model was chosen to allow the completion of this thesis within the given time frame. Another obstacle in the use of DAGs is that the functional relationship of the nodes is unclear from a DAG alone. For example, moderation is not modelled separately, it is shown as two parent variables having a separate relation (arrow) with the common child node. This makes it harder to interpret results in the applied context. Furthermore, we limited our DAGs to only include one mediator, which means we were not able to capture the relationships between these mediators. Finally, because of the method used for construction of the models, the results included some edges that were insignificant. Despite this, it was still possible to perform the mediation analyses, as it was clear that all baseline values of the four treatment goals were relevant to the survival time, directly or indirectly.

Another limitation of our analyses is our assumption of linear dose-response relationships, which was chosen to simplify the calculations, while in reality, far from all relationships are linear. Finally, we did not account for measurement errors. Besides confounders that are not accounted for, measurement errors in confounding variables can result in residual confounding, disturbing the true outcomes of the mediation analysis [27].
There are also some limitations to the MASTERPLAN dataset that was used for our analysis. Because patients were selected from nine Dutch hospitals, the population was mostly Dutch and Caucasian. Since multiple variables in our analysis are affected by race, this decreases its generalization. Besides this, the population was also relatively young and healthy. In addition, the trial took place in nine different hospitals, which may have given rise to selection bias. More importantly, many patients already met some goals, like ACE inhibitor or ARB use, at the start of the trial, which may have reduced the potential of the nurse practitioner intervention. Finally, there was a significant contamination bias, which means that all practitioners were informed about the treatment goals, thus the difference in treatment between the two groups was decreased, which can be seen, for example in the significant increase in ACE inhibitor and ARB prescription after the first visit in both groups, as opposed to just the intervention group.

The field of mediation analysis is developing rapidly and increasingly more scientists use this method in their research. In this paper, we have tried to use proper methodology and avoid common issues as described by Memon et al. As a first step, we made an effort to lay a good foundation of knowledge about mediation analysis and build our model based on previously established relationships or at least hypotheses. The research question was constructed using the PICOT format, which allowed us to formulate the mediation hypotheses in a clinical context. Baron and Kenny’s (1986) approach, which has been heavily critiqued, was avoided and instead, Hayes was consulted. This also means that presence of mediation and therefore the continuation of the analysis was not based on the presence of a direct effect. Besides this, we also refrained from labelling the mediation as ”partial” or ”complete” and inspected the specific indirect effects instead of the total indirect effect. Finally, we inspected both the two-tailed p-value and confidence intervals and utilized bootstrapping as recommended in Hayes.

5.2 Related Work

In 2002, NKF-KDOQI developed a conceptual model of CKD, which was revised and adapted by KDIGO in 2005, shown in figure. Some arrows are bidirectional because progression of CKD does not transpire in all patients and early interventions may prevent progression. The left side of the arrows is dotted to point out that remission is less frequent than progression.

This conceptual model is illustrative on a conceptual level, but does not lend itself to effect computations like our models do.

In the effect of the nurse practitioner intervention on a composite...
cardiovascular endpoint was analysed using the MASTERPLAN data. The effect of the intervention on the treatment goals and on the renal outcomes was evaluated. However, the effect of the intervention through the individual treatment was not calculated. Furthermore, the effect of the intervention on the secondary outcome of end-stage renal disease (ESRD) after a follow-up of 2 years was found to be insignificant.

This somewhat contradicts the conclusion of Peeters et al. [47] where the effect of the intervention on a composite renal endpoint including ESRD, death and 50% decrease in serum creatinine using the MASTERPLAN data as well. This paper reports that the effect of the intervention on this composite endpoint and the individual parts except for death after 2 years of follow-up is significant. Like the previous paper, this one only examined the total effect of the nurse practitioner intervention, not the effects mediated by the different treatment goals.

Two other trials were conducted with the same type of intervention for CKD patients, one in Canada and another in The Netherlands. Both trials were of shorter duration (24 and 12 months, respectively) than the MAS-
TERPLAN trial and also had different outcomes from the previously mentioned trials. The Canadian trial found little to no improvement in treatment goals or eGFR decrease between the two treatment groups after two years, while the Dutch trial found a decrease in systolic blood pressure after just one year. Furthermore, both papers included analysis of the effect of the intervention on the individual treatment goals, the analysis of the Canadian trial included estimation of the effect of the intervention on eGFR decline and neither included the effect on an endpoint [3, 54].

In summary, the MASTERPLAN dataset has been analysed before to investigate the effect of the addition of a nurse practitioner to the care team for CKD patients. Specifically the effect of this intervention on 11 different treatment goals and on a composite renal or cardiovascular endpoint. However, this total effect on the endpoints has not yet been split into the effect through the individual treatment goals. This thesis presents a first effort to do so on a subset of these goals, together with four models incorporating this subset, which has also not been done before, to our knowledge.
Chapter 6

Conclusions

Chronic kidney disease (CKD) is a heavy burden on patients and is very complex. Sadly, many questions still remain about its underlying mechanisms and the best treatment to prevent disease progression. Previous trials have shown that the addition of a nurse practitioner to the care team of CKD patients to enhance guideline adherence may improve quality of life and even renal outcomes. The MASTERPLAN trial is such a trial where 788 Dutch CKD patients were examined over a median follow-up time of 5.7 years while they were being treated with several treatment goals and guidelines under standard supervision or with the addition of a nurse practitioner to their care team. However, no analysis has been done yet to analyse through which of the treatment goals this intervention may exert its effect.

In this thesis we answer the research question

In patients with CKD, does increased effort to eliminate high blood pressure, decrease proteinuria, increase ACE inhibitor or ARB usage, and lessen salt intake by means of coaching by a nurse practitioner have a significant positive impact on the long term progression of kidney function?

To answer this question, we performed a mediation analysis on the MASTERPLAN data, limited to patients with stage 3-4 estimated glomerular filtration rate (eGFR). Patients who reached the treatment goals of

- blood pressure 130/85 mm Hg or lower,
- proteinuria 0.5 g per day or lower,
- ACE inhibitor or ARB e.g. enalapril 5 mg twice daily (or comparable dose of other ACE inhibitor) or irbesartan 75-150 mg (or comparable dose of other ARB) daily, and
- sodium intake 2g per day (or sodium excretion 100 mmol per day) or lower
within one year were compared to patients who did not reach this goal within
one year by means of survival time until a composite renal endpoint. This
renal endpoint consists of dialysis, transplant, 50% decrease of eGFR relative
to baseline, and death. It is important to note that our mediation analysis
is limited by the mediation graphs we constructed within the context of this
thesis. Because they are partially based on hypotheses that have not yet
been verified, they may be incorrect, which could reduce the validity of our
results.

From our mediation analysis we conclude that the effect of the nurse
practitioner intervention is not mediated by the aforementioned four treat-
ment goals. However, this does not mean that these goals do not improve
survival time of CKD patients. Their mediating effect may have been re-
duced by pre-existing good adherence for the ACE inhibitor or ARB use
goal, and general difficulty in achieving it for the salt intake goal. Besides
this, contamination bias was present, which means that all doctors were in-
formed of the treatment goals, which led to increase in adherence for some
of the goals in both treatment groups. Furthermore, the effect of the inter-
vention on the survival time of CKD patients itself is insignificant as well,
which is not in line with previous analyses of the dataset. These analyses,
however, were performed without imputation of incomplete data, used a
different renal endpoint, and did not employ mediation analysis. Despite
these differences, our results should be interpreted with caution and further
analysis, preferably based on an improved model, should be performed to
explain these differences. We conclude that the nurse practitioner interven-
tion is not effective in improving renal outcome in CKD patients with stage
3-4 eGFR, although the treatment goals, both discussed and not discussed
in this thesis, should not be neglected, despite the outcome of our analysis.

In practice, these conclusions indicate that nurse practitioners should
not focus on blood pressure, proteinuria, ACE inhibitor or ARB use, or salt
intake of CKD patients, as their effect is not mediated significantly by the
treatment goals relevant to these variables. Furthermore, the addition of a
nurse practitioner to the care teams of CKD patients will not improve renal
outcome.

6.1 Future Work

For this thesis, we limited our models to only include variables relevant to
blood pressure, proteinuria, ACE inhibitor or ARB use, and sodium ex-
cretion in the context of CKD progression. In reality, there are far more
variables of influence in this context. This means that more expansive mod-
els could provide more insight into the mediation of CKD progression.

Furthermore, for the construction of our models we utilised the ‘guessing-
and-testing’ method, which led to the modelling of some insignificant rela-
tionships. Another method to construct such a graph utilises discovery algorithms, where a DAG is constructed with actual data, while the exact relationships are still unknown, which could have resulted in more accurate models [56]. Thus, with future efforts, improved models could be made by adding more variables and constructing it from data with an algorithm instead of merely checking it with data.

Besides this, we divided the DAG into separate DAGs, each with one mediator, to keep the mediation analysis within the scope of this thesis. By building a model that includes all mediators and their interactions, more insight could be gained into their relationships. Finally, we assumed linear dose-response relationships to simplify the analyses. By employing non-linear models to more closely mimic the true relationships of the variables, the accuracy of the results of further analyses could be improved further.
Bibliography


45

DOI: 10.1111/bcp.12104.
URL: [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3791975/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3791975/) (visited on 03/15/2020).

“Proteinuria should be used as a surrogate in CKD”. In: *Nature Reviews Nephrology* 8.5 (May 2012). Number: 5 Publisher: Nature Publishing Group, pp. 301–306. ISSN: 1759-507X.
DOI: 10.1038/nrneph.2012.42.

URL: [http://statwww.epfl.ch/davison/BMA/](http://statwww.epfl.ch/davison/BMA/).


“Bootstrapping Regression Models in R”. In: (Sept. 21, 2018), p. 18.


[50] R Core Team. foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, … 2016. url: https://CRAN.R-project.org/package=foreign
[51] R Core Team. 


Appendix A

Exploratory Data Analysis

In this appendix, we present the results of the exploratory data analysis performed on the MASTERPLAN dataset.
Appendix A

MASTERPLAN Exploratory Data Analysis

The MASTERPLAN dataset contains CKD patient data from 788 subjects of the trial. 393 patients received standard care, while 395 patients received additional support from a nurse practitioner in achieving the 11 treatment goals set for the trial.

To summarise, the trial started with 788 patients of which only one patient was diagnosed with GFR stage 1. After the 20th visit, less than 25% of patients were still participating in the trial. About 2/3 of the patients were male and the majority was caucasian. Sex, race, diagnosis of diabetes and previous cardiovascular disease were fairly equally distributed over the control and intervention groups. RAS-inhibitor and ARB use increased significantly after the first visit in both treatment groups. Moreover, there were two extreme BMI measurements, and some uncommonly low or high eGFR measurements, neither of which could be identified as outliers. A few waist measurements were entered as 0 instead of NA, which was corrected. Furthermore, the albumin-creatinine ratio variance decreased extremely after the first visit, and one patient was found to have persistently low blood pressure, both of which were also not identified as outliers.

Treatment groups

The plot Distribution of patients over CKD stages shows that the patients of each GFR stage are about equally distributed over the control and intervention groups. There are relatively few patients in the stage 1 and stage 5 categories which makes sense, as these are the extreme cases.

In stage 1, there are usually few symptoms as the renal function is only slightly impaired, which means this diagnosis is often missed. This dataset only contains one patient with stage 1 GFR, which means that generalisations for this stage with this dataset could yield highly inaccurate conclusions. In stage 5, there is a larger chance of a patient satisfying one of the exclusion criteria for the MASTERPLAN like receiving a renal transplant less than a year before inclusion. Analyses on these two groups should be done with caution, as there is only 1 patient with GFR stage 1 and only 27 patients with GFR stage 5.

Visits distribution

The trial started with 788 patients. The plot Numer of patients participating over visits shows the progression of patient decline per visit. The horizontal lines indicate 75%, 50% and 25% of the number of patients at baseline.
Age

Looking at the age distribution, it is clear that most patients fall between 50-70 years and the differences between the GFR stages are not very large (see Distribution of age over CKD stages). GFR stage 1 only contains one data point, so there is no proper distribution.

The risk of CKD increases with damage of the kidneys. Kidney damage grows with age and so do conditions like diabetes and coronary heart disease, which in turn cause damage to the kidneys. This explains the age distribution. There are some outliers between 18 and 32 which are not unreasonable as CKD can be present at a younger age, though it is less likely.

Gender

The proportion of male to female patients is nearly equal in both treatment groups (0.68 to 0.32 for control versus 0.67 to 0.33 for treatment). Inspecting the distribution of sex over the CKD GFR stages (see Distribution of patients over CKD stages), there are less women in each category except for stage 5 in the intervention group.
Race
The majority of the trial population is caucasian, as the study was performed in Dutch hospitals (see Distribution of race over GFR stages).

Diabetes
In general, the proportion of diabetic patients is about the same in the intervention and control groups (0.25 and 0.22, respectively), however the distribution over the GFR stages differs more (see Distribution of diabetic patients over GFR stages).
**Cardiovascular disease**

The proportion of patients with cardiovascular disease differs somewhat between the treatment groups (0.25 in the control group and 0.33 in the intervention group), and differs more between the different GFR stages (see Distribution of CVD over GFR stages).

In general, the proportion of diabetic patients is about the same in the intervention and control groups (0.25 and 0.22, respectively), however the distribution over the GFR stages differs more (see Distribution of diabetic patients over GFR stages).

**Cholesterol**

Because low-density lipoprotein (LDL) is associated with increased risk of cardiovascular disease, which in turn is associated with CKD progression, it is important to treat high levels of LDL in CKD patients [3]. This means that a level of 100 mg/dL LDL or below is recommended for CKD patients.

The plot LDL over GFR stages for visit 1 and 5 by CVD diagnosis shows that realization of this goal is improved after one year for all GFR stages and CVD diagnoses. It is also noteworthy that LDL levels are lower for CKD patients with CKD for all GFR stages and both visits, except on visit 5 for GFR stage 2.
RAS-inhibitors and ARBs

The number of patients decreases every visit, so both use and non-use of RAS-inhibitors and ARBs decreases over time. For some visits, RAS-inhibitor use was not recorded, which also contributes to this decrease (See RAS-I & ARB use throughout visits).

It is clear, however that after the first visit, the increase in RAS-inhibitor and ARB usage relative to non-usage can be explained by a sudden increase in prescription due to the start of the trial. Doctors of both the intervention and control group were informed of the blood pressure treatment goal. It is also clear that this sudden increase is present in both groups, albeit less so in the control group.

BMI

BMI is calculated with the formula weight(kg) / (height(m))^2, where obesity is defined as BMI > 30 and underweight is defined as BMI < 18.5. there are 2 extreme BMI values, one below 15 and one above 50 (see BMI distribution).
The plot Waist circ and BMI shows that these two values (red and green for low and high BMI, respectively) are probably accurate, as the extremely high BMI is associated with an extremely high waist circumference and the extremely low BMI is associated with an extremely low waist circumference. This is further supported by plot Waist circ and hip circ.

The plot Waist circ and BMI also shows two other patients (purple and orange) with extreme waist circumference values. The patient with a waist circumference of 0 cm (purple) has simply not been recorded and should be a NA datapoint instead of 0, which is why it is not present in the plot Waist circ and hip circ. The other patient with a BMI of around 25 (orange) has a relatively small waist circumference. This patient also has a relatively small hip circumference (see Waist circ and hip circ), so it is possible that the high BMI can be explained by a small length.

Estimated glomerular filtration rate

As a sanity check, in the next plot the distribution of eGFR over the different GFR stages as measured at baseline is shown. This distribution fits with the definition of the stages. One outlier is shown for stage 2, but this eGFR is well within the bounds of 60 and 90.

There are two patients in the dataset with only one eGFR datapoint available, however these are not extreme values and do not need to be considered outliers (not shown). Furthermore, there are 14 patients with only two eGFR datapoints available, however, these are also not extreme thus do not need to be considered outliers either (not shown).

The GFR stages are shown in the table GFR stages.

Table 1: GFR stages
The next plot illustrates a number of abnormal measurements of eGFR throughout the MASTERPLAN trial, excluding the baseline. Progression is not requisite in CKD but remission is less frequent. Progression explains the decrease of eGFR in the different stages, however, the eGFR increase displayed by some datapoints is less realistic. There are, for example, two measurements of patients in the stage 3b group with eGFR of around 130, which is higher than the normal value of people without CKD [1].

In the plot, values that are more than two GFR categories larger than a patient’s diagnosis are marked as “High” and values that are more than three GFR categories smaller than a patient’s diagnosis have been marked “small”. Remission through renal replacement therapy (RRT) could explain the remission for patients with a higher GFR stage. The next plots provide a more detailed view of these extreme values.
Because there are only 14 patients with extreme “High” values, it is possible to look at their data individually to distinguish errors from plausible values. If the largest eGFR value of a patient is twice as large as the largest value before that one, and the patient did not receive RRT that could explain this difference, it can be treated as an outlier (based on the upper bound in the table eGFR outliers per stage).

There is a large difference between pre- and post-transplant eGFR values, which is visible as two separate clusters on the right in the graph Distribution of eGFR with High extremes (confirmation of dates before and after transplants was done but is not shown).

On the left in the same graph it is clear that there are some patients who have one eGFR value that is much greater than others which cannot be explained by RRT. These extreme values are somewhat odd but could still be accurate values.

There are 10 patients with extreme “Low” eGFR measurements, so it is possible to look at these individually. On the right in the plot Distribution of eGFR with Low extremes it is clear that the extreme “Low” values can be explained by progression of CKD. In the left graph of the same plot the extreme “Low” values are not part of such a trend, but might be explained by episodes of acute kidney injury.
Albumin-creatinine ratio

The plot below shows the distribution of the albumin-creatinine ratio over ckd patients at baseline. Albuminuria is not necessarily present in conjunction with CKD, but as CKD gets worse, there is more kidney damage, thus more albuminuria. This explains the upwards trend in the boxes in the plot. However, there are multiple extreme values in this graph, of up to 20 times the lower bound of severe albuminuria. These extreme values call for a more detailed analysis.

There are 31 patients with only one ACR measurement, two of which were performed at the second visit, while the rest was performed at the first visit (not shown).

The albuminuria stages are shown in the table Albuminuria stages.

Table 3: Albuminuria stages

<table>
<thead>
<tr>
<th>Albuminuria stage</th>
<th>ACR range</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>&gt;=90</td>
</tr>
<tr>
<td>G2</td>
<td>60-89</td>
</tr>
<tr>
<td>G3a</td>
<td>45-59</td>
</tr>
<tr>
<td>G3b</td>
<td>30-44</td>
</tr>
<tr>
<td>G4</td>
<td>15-29</td>
</tr>
<tr>
<td>G5</td>
<td>&lt;15</td>
</tr>
</tbody>
</table>
By plotting the acr distribution for every visit, it becomes clear that the baseline visit has significantly larger extreme values, up to more than twice as big (6201) as the largest followup value (2750). The variance is significantly decreased after the first visit and does not change nearly as much after (see Distribution of albumin-creatinine ratio for each visit), except for visits later than visit 22 as there are few data entries for those visits.

One theory to explain this variance is that the start of the trial changed treatment of the patients, which in turn drastically changed the albumin-creatinine ratio. This could be the case, as not all patients were treatment preceding the trial.

Blood pressure and heart rate

Systolic and diastolic blood pressure were measured two different ways; once yearly with a more accurate method (30 minute long procedure) and twice yearly with a less accurate method (oscillometric). The heart rate was measured with the once yearly more accurate method. There is a clear pattern in the twice yearly oscillometric blood pressure measurements due to rounding (see Distribution of sbp and dbp for both methods).
For patients with CKD, blood pressure higher than 130/80 mm Hg is considered high. A clear correlation exists between systolic and diastolic blood pressure, which is visible in the plot Systolic bp over diastolic bp per GFR stage, specifically for GFR stages 3a and 3b. In the other stages, there are a few datapoints that distort this correlation. These datapoints (marked in blue) are from 5 patients in total.

Specifically in the graph of GFR stage 2, the blue datapoints all belong to one patient with a rather low diastolic blood pressure. Looking at the medians of the systolic- and diastolic blood pressure measurements with both methods, it appears that these datapoints are real measured values and not measurement errors.
The heart rate was also measured with the 30 minute long procedure once a year. The normal resting heart rate for children older than 10 through seniors is 60 - 100 bpm [2]. A large amount of the measurements are lower or higher than this range, but they do not appear to be observational errors.

<table>
<thead>
<tr>
<th>Median SBP (30 min)</th>
<th>Median DBP (30 min)</th>
<th>Median SBP (osc)</th>
<th>Median DBP (osc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>146.5</td>
<td>48.8</td>
<td>161</td>
<td>52</td>
</tr>
</tbody>
</table>

Sodium excretion

The mean for sodium excretion lies around 150 mmol/24hr for all GFR stages (except stage 1, as there are insufficient datapoints to properly indicate a mean), and there is not much variation in this between the first and fifth visit (baseline and 1 year checkup, respectively).

The treatment goal for sodium excretion is a maximum of 100 mmol/24hr, so there are many values that are much larger than this limit, some even more than 4 times as large.
References


Appendix B

Missing Data Analysis and Imputation

This appendix contains the results of the missing data analysis on the MASTERPLAN dataset and the settings used for imputation and an analysis of its results.
MASTERPLAN Imputation Analysis

This analysis was constructed with the help of the MICE vignettes by Gerko Vink and Stef van Buuren (see here [https://www.gerkovink.com/miceVignettes/]).

To summarise, the data was transformed into long format containing baseline variables, measures on the first visit and measures on the fifth visit, creating a dataset of 742 entries. The most common pattern was no missing variables (343), and after that missing proteinuria measurement at visit 5 (118). Furthermore, we concluded that the data was probably subject to a MAR pattern. Because visit 5 proteinuria was the most frequently missing variable (34%), we produced 34 imputed datasets using the Mice library. We found good convergence for the imputations and performed an analysis on the imputed datasets. We concluded that the imputations were successful.

Structure of the dataset

The MASTERPLAN dataset contains measurements of 788 patients over the course of the trial. These measurements include systolic blood pressure, heart rate, albumin-creatinine ratio, estimated glomerular filtration rate, RAS-inhibitor usage, etc. Visits 1, 5, 9, 13, 17 and 21 were the designated yearly checkup visits for both the intervention and control groups.

The base dataset contains 11708 entries of yearly and non-yearly visits of the 788 patients until censoring or a composite renal endpoint of 50% decrease in eGFR, dialysis, kidney transplant, or death. The dataset used for imputation contains only the baseline measurements (visit 1) and the first yearly checkup visit (visit 5). Of the initial 788 patients, 742 were still participating at visit 5.

Some of the more important variables for our research are outcome, date of outcome, systolic blood pressure, eGFR, albumin-creatinine ratio, RAS-inhibitor and ARB use, proteinuria and sodium excretion.

Furthermore, the original dataset contained separate rows for each visit with an identifying patient id and visit number. For the imputation analysis this data was transformed from long data into wide data, resulting in a dataset of 742 rows instead of the original 1530.

The missingness in this dataset is mostly dependent on the type of visit; only the yearly checkups required measurement of most variables. This is illustrated by the low median of amount of missing values of the yearly visits. However, this effect decreases after a number of visits in the dataset, as the alignment of yearly visits with the visit number gets worse.

The pattern of yearly visits is recognisable in the plot and table below, but it is also clear that it gets less pronounced over time. To illustrate this with numbers, only 448 patients who had a check up visit between 12 and 18 months after enrolling had less than 8 missing values. At that point, 726 patients were still participating in the trial.

The plot also shows that the trial was closed in July 2010, 75 months after the starting date in April 2004. After that, only data on mortality and renal outcome were collected, which explains the number of missing variables.

Median of amount of missing values per visit

<table>
<thead>
<tr>
<th>Visit</th>
<th>Missing Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>21</td>
<td>10.5</td>
</tr>
</tbody>
</table>
To increase number of more complete entries, some records of the first and fifth visit were replaced. For visits with more than 8 missing values, one visit earlier or later is selected, on the condition that such a visit contains less than 8 missing values. The resulting set of datapoints still does not capture all true “checkup” visits, however, which is illustrated in the plot below.

Missingness of variables is also influenced by them usually being measured together, like systolic and diastolic blood pressure, waist and hip circumference, or drug use.

**Missing data patterns**

The missing data pattern analysis shows that the most common pattern is no missing values in both visits and after that, missing proteinuria on visit 5. These two patterns are printed below together with the total number of missing entries per variable.
The most common missing data pattern has all values that were recorded at baseline and did not change (sex, race, etc), while all other values are missing (blood pressure, sodium excretion, etc) except for eGFR (mdrd_175). Part of this is due to the fact that not all variables are measured every visit and the incidence of an event or censoring is not known beforehand. Another important factor is that at least 200 patients were still included during the time of extended follow-up, which means that a lot of their values are missing by default.
In the fluxplot below of the data used for the imputation, a higher outflux indicates possibly more influence as predictor, while a higher influx indicates a stronger dependence on the imputation model. Like the missing data pattern indicated; in the upper left are variables with more known values, while in the bottom right are variables with more missing values.

A lot of the variables are in the top left, meaning they are probably significantly influential as predictors in the imputation model, because a lot of their values are known. The inverse is true for the variables in the bottom right.

**Type of missingness**

Incomplete variables can come about by different mechanisms; missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MCAR means that there is no systematic pattern for missing variables, MAR means that systemic patterns for missing variables can be explained completely with other observed variables, and MNAR means that these patterns cannot be explained by observed variables.

By plotting an observed variable against the missingness of other variables, MAR mechanisms can be uncovered. This can be seen as right-tailed or left-tailed MAR missingness. When the plot of the observed data is further to the right when the other variable is observed as well and thus is further to the left when the other variable is not observed, this indicates a right-tailed MAR missingness. When there is not a clear horizontal shift in the plots, there is a MCAR or possible MNAR missingness.

From the 5 plots below, it is clear that the missingness of most suspected variables is not dependent on eGFR. This is somewhat notable, as a lower eGFR indicates greater illness in a patient, which might cause a doctor to refrain from measurements to minimize the burden on the patient. This implies that the missingness of the plotted variables is MCAR or possibly MNAR.

Furthermore, SBP and RAS-I use on visit 5 have a very limited amount of missing values (17 and 11, respectively of 742 total), where the largest amount occurs with low eGFR values of the same visit. Thus, this missingness could be at least somewhat right-tailed.
To get some assurance of the data being MAR, one can perform a sensitivity analysis. By adding or subtracting some amount (delta) to an imputed value, the imputations will change. If the adjustment has little effect on the analysis of the imputations, the data is stable to that type of MNAR-mechanism, thus it is more probable that the data is MAR.

**Imputation**

First, a ‘dry run’ is performed with the mice function to initialize the imputation methods (meth) and predictor matrix (pred) for all variables.

```r
set.seed(1235)
ini <- mice(m.re.imp,maxit=0)
meth <- ini$meth
pred <- ini$pred
```

By default, mice uses predictive mean matching for numericals, logistic regression for factors with 2 levels, polytomous regression for unordered factors of 2+ levels, and a proportional odds model for ordered factors of 2+ levels. For this analysis only predictive mean matching and logistic regression are used because of the types of variables that are missing.

To further improve the imputation results, passive imputation was used for the waist-hip ratio (waist/hip). Because the initial predictor matrix uses all variables except itself for imputation, this would cause a feedback loop. Thus the waist-hip ratio should be excluded from the set of predictors for waist and hip circumference.
There are also some variables that should be chosen as predictors with caution because they are highly correlated, for example bmi and obesity (defined as bmi>30). The predictor matrix was computed with the quickpred function with a minimal correlation of 0.1 and a minimum proportion of usable cases of 0.25. Age and sex were included as predictor for every variable, while pid was excluded. Finally, some variables were manually excluded.

The imputation method for each variable is shown below. The predictor matrix is too large to properly display, so only the commands for the construction and the count of variables (row 2) using x amount of predictor variables (row 1) is shown.

```r
meth["whr"] <- "-I(waist/hip)"
incl = c("age","sex")
excl = c("pid","waist","hip","obesity","visit_type.5","age_fu.1","age_fu.5","date_base","time.1","time.5")
pred <- quickpred(data=m.re.imp, mincor=0.1, minpuc=0.25, include=incl, exclude=excl)
pred[c("waist","hip","whr"),] <- 0
pred[c("waist","hip","bmi"),] <- 1
pred["whr","c("waist","hip")]
pred[c("ldl.1","hdl.1"),"bmi"] <- 1
pred[grepl("\.5",rownames(pred),invert=T),grepl("\.5",rownames(pred))]
``` 

The number of imputations is set to the percentage of missingness of the variable that is missing most often; proteinuria on visit 5 is missing 252 of the 742 times, so the number of imputations is set to 34. For each imputation, 5 iterations are performed.

```r
imp <- mice(m.re.imp, meth=meth, pred=pred, print=FALSE, m=34)
```

**Imputation analysis**

To analyze the imputations, it is important to look at the convergence and distribution of the imputations. The graphs below show the mean and standard deviation of the imputed variables per imputation over the iterations.
The next plots will show the range of the observed and imputed values for 6 randomly chosen imputations of the 34 of waist versus hip, whr versus waist, sbp versus dbp at visit 5, sbp_dnmp and dbp_dnmp at visit 5, album versus sbp_dnmp at visit 1, ldl versus hdl at visit 1, and ldl versus hdl at visit 5. The imputed values are within the range of the observed values and look reasonable.
Next, the stripplots of album.1, mdrd_175.5, creat.5 and hb.5 are shown for 6 randomly chosen imputations. These variables are plotted separately because they contain very few missing values. The distribution of the imputed variables seems reasonable.
Below are the density plots of some other imputed variables. The distributions of the imputed values are mostly close to the original values.
Furthermore, smoke and stroke contain 9 and 56 missing values, so the ratio of the original data remains about the same, regardless of the imputed values.

Thus, after consideration of the imputed values, it seems that these are reasonable and can be used for further analyses.
Appendix C

Individual DAGs

This appendix contains the dagitty code for the base DAG and the four individual DAGs.

---Base---

dag {
  ACEI_ARB_T0 [pos="0.000,2.000"]
  BP_T0 [pos="0.000,1.000"]
  Intervention [exposure,pos="0.000,5.000"]
  PU_T0 [pos="0.000,3.000"]
  SE_T0 [pos="0.000,0.000"]
  eGFR_T0 [pos="0.000,4.000"]
  PU_T0 -> BP_T0 [pos="0.315,2.016"]
  SE_T0 -> BP_T0
  SE_T0 -> PU_T0 [pos="0.315,1.365"]
  eGFR_T0 -> ACEI_ARB_T0 [pos="0.284,2.971"]
  eGFR_T0 -> PU_T0
  eGFR_T0 -> SE_T0 [pos="0.729,2.018"]
}

---Blood Pressure---

dag {
  ACEI_ARB_T0 [pos="0.000,2.000"]
  BP_T0 [pos="0.000,1.000"]
  BP_T1 [pos="1.000,0.000"]
  ESRD [outcome,pos="2.000,3.000"]
  Intervention [exposure,pos="0.000,5.000"]
  PU_T0 [pos="0.000,3.000"]
  SE_T0 [pos="0.000,0.000"]
  eGFR_T0 [pos="0.000,4.000"]
  ACEI_ARB_T0 -> BP_T1
  ACEI_ARB_T0 -> ESRD
  BP_T0 -> BP_T1
  BP_T1 -> ESRD
  Intervention -> BP_T1
  Intervention -> ESRD

94
PU, T0 -> BP, T0  \[ pos = "0.315, 2.016" \]
PU, T0 -> BP, T1
PU, T0 -> ESRD
SE, T0 -> BP, T0
SE, T0 -> BP, T1
SE, T0 -> ESRD
SE, T0 -> PU, T0  \[ pos = "0.315, 1.365" \]
eGFR, T0 -> ACE,ARB, T0  \[ pos = "0.284, 2.971" \]
eGFR, T0 -> BP, T1
eGFR, T0 -> ESRD
eGFR, T0 -> PU, T0
eGFR, T0 -> SE, T0  \[ pos = "0.729, 2.018" \]

---Proteinuria---

\[
\begin{align*}
\text{ACEI,ARB, T0} & \rightarrow \text{ESRD} \\
\text{ACEI,ARB, T0} & \rightarrow \text{PU, T1} \\
\text{BP, T0} & \rightarrow \text{ESRD} \\
\text{BP, T0} & \rightarrow \text{PU, T1} \\
\text{Intervention} & \rightarrow \text{ESRD} \\
\text{Intervention} & \rightarrow \text{PU, T1} \\
\text{PU, T0} & \rightarrow \text{BP, T0}  \[ pos = "0.315, 2.016" \] \\
\text{PU, T0} & \rightarrow \text{PU, T1} \\
\text{PU, T1} & \rightarrow \text{ESRD} \\
\text{SE, T0} & \rightarrow \text{BP, T0} \\
\text{SE, T0} & \rightarrow \text{ESRD} \\
\text{SE, T0} & \rightarrow \text{PU, T0}  \[ pos = "0.315, 1.365" \] \\
eGFR, T0 & \rightarrow \text{ACE,ARB, T0}  \[ pos = "0.284, 2.971" \] \\
eGFR, T0 & \rightarrow \text{ESRD} \\
eGFR, T0 & \rightarrow \text{PU, T0} \\
eGFR, T0 & \rightarrow \text{PU, T1} \\
eGFR, T0 & \rightarrow \text{SE, T0}  \[ pos = "0.729, 2.018" \]
\end{align*}
\]

---ACE Inhibitor or ARB use---

\[
\begin{align*}
\text{ACEI,ARB, T0} & \rightarrow \text{ESRD} \\
\text{ACEI,ARB, T1} & \rightarrow \text{ESRD} \\
\text{BP, T0} & \rightarrow \text{ESRD} \\
\text{BP, T0} & \rightarrow \text{PU, T1} \\
\text{Intervention} & \rightarrow \text{ESRD} \\
\text{Intervention} & \rightarrow \text{PU, T1} \\
\text{PU, T0} & \rightarrow \text{BP, T0}  \[ pos = "0.315, 2.016" \] \\
\text{PU, T0} & \rightarrow \text{PU, T1} \\
\text{PU, T1} & \rightarrow \text{ESRD} \\
\text{SE, T0} & \rightarrow \text{BP, T0} \\
\text{SE, T0} & \rightarrow \text{ESRD} \\
\text{SE, T0} & \rightarrow \text{PU, T0}  \[ pos = "0.315, 1.365" \] \\
eGFR, T0 & \rightarrow \text{ACE,ARB, T0}  \[ pos = "0.284, 2.971" \] \\
eGFR, T0 & \rightarrow \text{ESRD} \\
eGFR, T0 & \rightarrow \text{PU, T0} \\
eGFR, T0 & \rightarrow \text{PU, T1} \\
eGFR, T0 & \rightarrow \text{SE, T0}  \[ pos = "0.729, 2.018" \]
\end{align*}
\]
Appendix D

Implication Testing Script

This appendix contains the script for testing the conditional independencies of our DAGs, specifically for the DAG containing the blood pressure treatment goal as mediator.

```
# Load required packages
library(dplyr)
library(mice)
library(dagitty)
library(writexl)
filter <- dplyr::filter

# Impute missing data

# Read imputation data
# This CSV file contains the baseline and fifth visit values of 742 patients in long format
imputation.data <- read.csv("Z:/data//imputationdata.csv")

# Transform date and factor columns
factor.vars <- c("treatment", "sex", "caucasian", "diabetes", "cvd", "chd", "stroke", "smoke", "ntx_bas", "obesity", "med_rasi.1", "med_diur.1", "med_rasi.5", "med_diur.5")
date.vars <- c("date_bas", "dob", "outcome_date", "dat_visit.1", "dat_visit.5")
stage.filter <- c("GFR stage 3a", "GFR stage 3b", "GFR stage 4")
imputation.data[factor.vars] <- lapply(imputation.data[factor.vars], as.factor)
imputation.data[date.vars] <- lapply(imputation.data[date.vars], as.Date)
```
# Rename columns to match DAG variables and exclude GFR stage 1, 2 and 5
imputation.data <- imputation.data %>% mutate(outcome = ifelse(outcome=='Censored', 0, 1), outcome = as.factor(outcome)) %>% rename(ACEI_ARB_T0=med_rasi.1, ACEI_ARB_T1=med_rasi.5, BP_T0= sbp_dnmp.1, BP_T1=sbp_dnmp.5, ESRD=outcome, Intervention=treatment, PU_T0=uprot.1, PU_T1=uprot.5, SE_T0=urinenatrium.1, SE_T1=urinenatrium.5, eGFR_T0= mrdr.175.1) %>% filter(ckd_stage %in% stage.filter)

# Initialize imputation
set.seed(123567)
i <- mice(imputation.data, maxit=0)
meth <- ini$meth
pred <- ini$pred

# Setup imputation
meth[["whr"]]<- "I(waist/hip)"
incl = c("age", "sex")
excl = c("pid", "waist", "hip", "obesity", "visit_type.5", "age_fu.1", "age_fu.5", "date_base", "time.1", "time.5")
pred <- quickpred(data=imputation.data, mincor=0.1, minpuc=0.25, include=incl, exclude=excl)

# Rows indicate the values to be imputed and columns indicate predictor variables
pred[c("waist","hip"), "whr"] <- 0
pred[c("waist","hip"), "bmi"] <- 1
pred["whr", c("waist","hip")]<- 1

pred[c("ldl.1","hdl.1"), "bmi"] <- 1

pred[grep("((\.|5))\((._1)\)", rownames(pred), invert=T), grep("((\.|5))\[(._1)\]", rownames(pred))]<- 0

pred[grep("((\.|1\|5))\((._0|1)\)"", rownames(pred)), "eGFR_T0"] <- 1
pred[grep("((\.|5))\[(._1)\]", rownames(pred)), "mrdr.175.5"] <- 1
pred[grep("((\.|1\|5))\[(._0|1)\]"", rownames(pred)), "outcome_date"] <- 1

pred["BP_T1", "dbp_dnmp.5"] <- 1
pred["dbp_dnmp.5", "BP_T1"] <- 1
pred["sbp.5", "dbp.5"] <- 1
pred["dbp.5", "sbp.5"] <- 1

# Imputation of 34 datasets
imp <- parlmice(imputation.data, meth=meth, pred=pred, print=FALSE, n.core=2, n.imp.core=17)

# Import DAG

# Setup DAG
# This is the DAG with the blood pressure treatment goal as mediator

dag.masterplan <- dagitty("dag {
  ACEI_ARB_T0 [pos="0.000,2.000"]
  BP_T0 [pos="0.000,1.000"]
  BP_T1 [pos="1.000,1.000"]
  ESRD [outcome,pos="2.000,3.000"]
  Intervention [exposure,pos="0.000,5.000"]
  PU_T0 [pos="0.000,3.000"]
  SET0 [pos="0.000,0.000"]
  eGFR_T0 [pos="0.000,4.000"]
  ACEI_ARB_T0 -> BP_T1
  ACEI_ARB_T0 -> ESRD
  BP_T0 -> BP_T1
  BP_T1 -> ESRD
  Intervention -> ESRD
  Intervention -> BP_T1
  PU_T0 -> BP_T0
  PU_T0 -> BP_T1
  PU_T0 -> ESRD
  SET0 -> BP_T0
  SET0 -> BP_T1
  SET0 -> PU_T0
  SET0 -> ESRD
  eGFR_T0 -> ACEI_ARB_T0
  eGFR_T0 -> BP_T1
  eGFR_T0 -> ESRD
  eGFR_T0 -> PU_T0
  eGFR_T0 -> SET0
}")

plot(dag.masterplan)

# Check DAG implications

# Get conditional independencies from DAG
condindep <- impliedConditionalIndependencies(dag.masterplan)

# There is no automated local testing for imputed data and it does not work if NAs are present
# A workflow for this is implemented manually below

# Create sets of lists to store results

99
result.implications <- vector(mode = "list", length = length(condindep))
implications <- vector(mode = "list", length = length(condindep))

# Obtain all local tests (using linear regression) for all implied conditional independencies
for (j in seq_along(condindep)) {
    # Create a regression formula from the implied conditional independencies
    x <- condindep[[j]]$X
    y <- condindep[[j]]$Y
    z <- condindep[[j]]$Z

    # Normalise the independent variable if it is not binomial
    if (grepl("(ACE|ARB|T(0|1))|(ESRD)|(Intervention)", y)) {
        eq <- paste0(x, " ~ ", y)
    } else {
        eq <- paste0(x, " ~ scale(" , y , ")")
    }

    for (i in seq_along(z)) {
        if (grepl("(ACE|ARB|T(0|1))|(ESRD)|(Intervention)", z[i])) {
            eq <- paste0(eq, " + ", z[i])
        } else {
            eq <- paste0(eq, " + scale(" , z[i] , ")")
        }
    }

    # Perform the local test on the imputed data
    # Use logistic regression for binomial dependent variables and linear regression for continuous dependent variables
    if (grepl("(ACE|ARB|T(0|1))|(ESRD)|(Intervention)", x)) {
        implications[[j]] <- with(imp, glm(as.formula(eq), family = binomial))
    } else {
        implications[[j]] <- with(imp, lm(as.formula(eq)))
    }

    # Description of the local test
    lcltest.descr <- paste0(x, " ~ ", y)
    if (length(z) > 0) {
        lcltest.descr <- paste0(lcltest.descr, " ~ ")
        for (i in seq_along(z)) {
            lcltest.descr <- paste0(lcltest.descr, z[i])
            if (i < length(z)) {
                lcltest.descr <- paste0(lcltest.descr, " + ")
            }
        }
    }

    # Obtain the pooled result

100
result.implicitations[[j]] <- cbind(lcltest.descr, summary(pool(
    implicitations[[j]])))[2, ]

# Combine the results and export to .xlsx file
rownames(result.implicitations) <- NULL
result.implicitations[, c(3:ncol(result.implicitations))] <-
    round(result.implicitations[, c(3:ncol(result.implicitations))], 3)
write.xlsx(x=result.implicitations, path="implicationsBP_uprot.xlsx")
Appendix E

Edge Estimation Script

This appendix contains the script for the edge estimation of our DAGs, specifically the DAG with the blood pressure treatment goal as mediator.

# Load required packages
library(dplyr)
library(lubridate)
library(mice)
library(dagitty)
library(survival)
library(writexl)
library(boot)
library(Amelia)
filter <- dplyr::filter

# Impute missing data
# This CSV file contains the baseline and fifth visit values of 742 patients in long format
imputation.data <- read.csv("Z:/data//imputationdata.csv")

# Transform date and factor columns
"ntx_bas", "obesity", "med_rasi.1", "med_diur.1", "med_rasi.5", "med_diur.5")
date.vars <- c("date_bas", "dob", "outcome_date", "dat_visit.1", "dat_visit.5")
stage.filter <- c("GFR stage 3a", "GFR stage 3b", "GFR stage 4")
imputation.data[factor.vars] <- lapply(imputation.data[factor.vars], as.factor)
imputation.data[date.vars] <- lapply(imputation.data[date.vars], as.Date)

# Rename columns to match DAG variables, prepare for Cox
# proportional hazards regression,
# and exclude GFR stage 1, 2 and 5
imputation.data <- imputation.data %>%
mutate(outcome = ifelse(outcome=="Censored", 1, 2),
        outcome = as.factor(outcome),
        outcome.time = interval(date_bas, outcome_date) / months(1))

rename(ACELARB_T0=med_rasi.1, ACELARB_T1=med_rasi.5, BP_T0=sbp_dnmp.1, BP_T1=sbp_dnmp.5,
        ESRD=outcome, Intervention=treatment, PU_T0=uprot.1, PU_T1=uprot.5,
        SE_T0=urinenatrium.1, SE_T1=urinenatrium.5, eGFR_T0=mdrd.175.1)

filter(ckd_stage %in% stage.filter)

# Test Cox proportional hazards assumption
fit <- coxph(Surv(outcome_time, as.numeric(ESRD)) ~
             Intervention + BP_T1 +
             ACELARB_T0 + PU_T0 + SE_T0 + eGFR_T0,
             data = imputation.data)
ggcoxph(cox.zph(fit))

# Initialize imputation
set.seed(123567)
ini <- mice(imputation.data, maxit=0)
meth <- ini$im
pred <- ini$im$pred

# Setup imputation
meth["whr"] <- "~I(waist/hip)"

incl = c("age", "sex")
excl = c("pid", "waist", "hip", "obesity", "visit_type.5", "age_fu.1", "age_fu.5",
         "date_bas", "time.1", "time.5", "outcome_time")
pred <- quickpred(data=imputation.data, mincor=0.1, minpuc=0.25,
                   include=incl, exclude=excl)

# Rows indicate the values to be imputed and columns indicate
# predictor variables
pred["waist","hip"] <- 0
pred["waist","hip", "bmi"] <- 1
pred["whr", "waist","hip"] <- 1
pred["ldl.1","hdl.1", "bmi"] <- 1
pred[grepl("(\d.5)|(\_T1)",rownames(pred),invert=T), grepl("(\d.5|
         \_T1)",rownames(pred))] <- 0
pred[grep("(\.(1|5))|\.(\.(0|1))", rownames(pred)), "eGFR_T0"] <- 1
pred[grep("(\.)|\.(T0|1)", rownames(pred)), "mdrd.175.5"] <- 1
pred[grep("(\.(1|5))|\.(T0|1))", rownames(pred)), "outcome_date"] <- 1
pred["BP_T1", "dbp_dump.5"] <- 1
pred["dbp_dump.5", "BP_T1"] <- 1
pred["sbp.5", "dbp.5"] <- 1
pred["dbp.5", "sbp.5"] <- 1

# Imputation of 34 datasets
imp <- parlmice( imputation.data , meth=meth, pred=pred, print=FALSE, n.core=2, n.imp.core=17)

# Import DAG
# Setup DAG
# This is the DAG with the blood pressure treatment goal as mediator
dag.masterplan <- dagitty("dag {
  ACEI_ARB_T0 [ pos="0.000 ,2.000\""]
  BP_T0 [ pos="0.000 ,1.000\""]
  BP_T1 [ pos="1.000 ,1.000\""]
  ESRD [ outcome , pos="2.000 ,3.000\""]
  Intervention [ exposure , pos="0.000 ,5.000\""]
  PU_T0 [ pos="0.000 ,3.000\""]
  SE_T0 [ pos="0.000 ,0.000\""]
  eGFR_T0 [ pos="0.000 ,4.000\""]
  ACEI_ARB_T0 -> BP_T1
  ACEI_ARB_T0 -> ESRD
  BP_T0 -> BP_T1
  BP_T1 -> ESRD
  Intervention -> ESRD
  Intervention -> BP_T1
  PU_T0 -> BP_T0
  PU_T0 -> BP_T1
  PU_T0 -> ESRD
  SE_T0 -> BP_T0
  SE_T0 -> BP_T1
  SE_T0 -> PU_T0
  SE_T0 -> ESRD
  eGFR_T0 -> ACEI_ARB_T0
  eGFR_T0 -> BP_T1
  eGFR_T0 -> ESRD
  eGFR_T0 -> PU_T0
  eGFR_T0 -> SE_T0
}")
plot(dag.masterplan)
# Generate regression formulas

# To calculate the coefficients for edges from parents of a
# binomial variable, use logistic regression,
# for edges from parents of a continuous variable, use linear
# regression, and
# for edges from parents of the outcome variable, use Cox
# regression

# Create a list with each variable and its parents
depindvars <- vector(mode = "list", length = length(names(dag.masterplan)))
for (j in seq_along(names(dag.masterplan))) {
  depindvars[[j]]$X <- names(dag.masterplan)[j]
  depindvars[[j]]$Y <- parents(dag.masterplan, names(dag.masterplan)[j])
}

# Create the list of regression formulas
formulas <- list()
formulaindex <- 1
for (i in seq_along(depindvars)) {
  x <- depindvars[[i]]$X
  y <- depindvars[[i]]$Y
  if (length(y) > 0) {
    # For binary variables, use logistic regression
    if (grepl("ACEI ARB T (0|1)", x)) {
      formula <- paste0("glm(" , x , " ~ " )
      for (k in seq_along(y)) {
        # Scale continuous independent variables
        if (grepl("(ACEI ARB T (0|1)| Intervention)", y[k])) {
          formula <- paste0(formula , y[k])
        } else {
          formula <- paste0(formula , "scale(" , y[k] , ")")
        }
        if (k < length(y)) {
          formula <- paste0(formula , " + ")
        } else {
          formula <- paste0(formula , ", family='binomial')")
        }
      }
    } # For the survival outcome, use Cox proportional hazards regression
    else if (grepl("ESRD", x)) {
      formula <- paste0("coxph(Surv(outcome_time, as.numeric(" , x , ")) ~ " )
      for (k in seq_along(y)) {
        # Scale continuous independent variables
        if (grepl("(ACEI ARB T (0|1)| Intervention)", y[k])) {
          formula <- paste0(formula , y[k])
        } else {
          formula <- paste0(formula , "scale(" , y[k] , ")")
        }
      }
    } else {
      formula <- paste0(formula , "scale(" , y[k] , ")")
    }
  } else {
    # For continuous variables, use linear regression
    formula <- paste0("lm(" , x , " ~ " )
    for (k in seq_along(y)) {
      # Scale continuous independent variables
      if (grepl("(ACEI ARB T (0|1)| Intervention)", y[k])) {
        formula <- paste0(formula , y[k])
      } else {
        formula <- paste0(formula , "scale(" , y[k] , ")")
      }
    }
  }
}

# Create the list of regression formulas
formulas <- list()
formulaindex <- 1
for (i in seq_along(depindvars)) {
  x <- depindvars[[i]]$X
  y <- depindvars[[i]]$Y
  if (length(y) > 0) {
    # For binary variables, use logistic regression
    if (grepl("ACEI ARB T (0|1)", x)) {
      formula <- paste0("glm(" , x , " ~ " )
      for (k in seq_along(y)) {
        # Scale continuous independent variables
        if (grepl("(ACEI ARB T (0|1)| Intervention)", y[k])) {
          formula <- paste0(formula , y[k])
        } else {
          formula <- paste0(formula , "scale(" , y[k] , ")")
        }
        if (k < length(y)) {
          formula <- paste0(formula , " + ")
        } else {
          formula <- paste0(formula , ", family='binomial')")
        }
      }
    } # For the survival outcome, use Cox proportional hazards regression
    else if (grepl("ESRD", x)) {
      formula <- paste0("coxph(Surv(outcome_time, as.numeric(" , x , ")) ~ " )
      for (k in seq_along(y)) {
        # Scale continuous independent variables
        if (grepl("(ACEI ARB T (0|1)| Intervention)", y[k])) {
          formula <- paste0(formula , y[k])
        } else {
          formula <- paste0(formula , "scale(" , y[k] , ")")
        }
      }
    } else {
      formula <- paste0(formula , "scale(" , y[k] , ")")
    }
  } else {
    # For continuous variables, use linear regression
    formula <- paste0("lm(" , x , " ~ " )
    for (k in seq_along(y)) {
      # Scale continuous independent variables
      if (grepl("(ACEI ARB T (0|1)| Intervention)", y[k])) {
        formula <- paste0(formula , y[k])
      } else {
        formula <- paste0(formula , "scale(" , y[k] , ")")
      }
    }
  }
}
if (k < length(y)) {
  formula <- paste0(formula, " + ")
} else {
  formula <- paste0(formula, "")
}

# For continuous variables, use linear regression
else {
  formula <- paste0("lm(" , x, " ~ ")
for (k in seq_along(y)) {
  # Scale continuous independent variables
  if (grepl("(ACEI|ARB|T(0|1))\|\{Intervention\}", y[k]) {
    formula <- paste0(formula, y[k])
  } else {
    formula <- paste0(formula, "scale(" , y[k], ", ")")
  }
  if (k < length(y)) {
    formula <- paste0(formula, " + ")
  } else {
    formula <- paste0(formula, "")
  }
}
  formulas[[formulaindex]] <- formula
  formulaindex <- formulaindex + 1
  # Skip variables without parents
} else {
  next
}

# The edge weights are computed for each of the imputed datasets
# Create sets of lists to store results
result.imptest <- vector(mode = "list", length = length(names(dag.masterplan)))
imptest <- vector(mode = "list", length = length(names(dag.masterplan)))

# Obtain coefficients (using regression) for all parents of each variable and combine using Rubin’s rules
for (j in seq_along(formulas)) {
  # Perform the regression on the imputed data
  imptest[[j]] <- with(imp, eval(parse(text=formulas[[j]])))

  # Obtain the pooled result
  result.imptest[[j]] <- cbind(formulas[[j]], summary(pool(imptest[[j]])))
}
# Combine the estimates for all edges
result.estimations <- do.call(rbind, result.imptest)

# Estimate edge weights and significance with bootstrap

# The standard error of the regression is computed for each of the imputed datasets by bootstrapping

# Define number of bootstrapping samples
bootNum <- 5000

# Create list to store bootstrap results of each imputation
botes <- array(NA, dim=c(c(nrow(result.estimations),2), imp$im))

# Define function to perform all regressions on a bootstrap sample
boreg <- function(data, indices, formulalist) {
  
  # Retrieve bootstrap sample
d <- data[indices, ]

  # Create list to store result of regression for each formula
  indivregs <- vector(mode = "list", length = length(formulalist ))

  # Perform each regression on the bootstrap sample and store
  # result for (f in seq_along(formulalist)) {
  # fit <- with(d, eval(parse(text=formulalist[[f]]))))
  # indivregs[[f]] <- coef(fit)
  # }

  # Compress individual lists of regression results into one
  # list
  # containing all regression results for this bootstrap sample
  regresults <- unlist(indivregs)
  return(regresults)
}

# Perform bootstrapping on each of the 34 imputed datasets
for (i in 1:imp$m) {
  
  # Complete original dataset with imputed variables of
  # imputation i
  compldata <- complete(imp, i)

  # Generate bootstrap samples and apply 'regress' function to
  # obtain
  # regression coefficients for each regression of each
  # bootstrap sample

boreult <- boot(data=compldata, statistic=boregress, R=
    bootNum, formula=list=formulas, parallel="snow")
boverview <- cbind(boreult$t0, #coef
    apply(boreult$t, 2, sd)) #se(coef)
btest[, , i] <- boverview

# Create list to store results of pooling
result.bootest <- matrix(NA, nrow=nrow(result.estimations), ncol
    =4)
colnames(result.bootest) <- c("Boot_est", "Boot_SE", "Lower_95",
"Upper_95")

# Pool results of 34 imputed datasets
boTestMeld <- mi.meld(q=btest[,1,], se=btest[,2,], byrow=F)
result.bootest[,1] <- boTestMeld$q.mi
result.bootest[,2] <- boTestMeld$se.mi
result.bootest[,3] <- result.bootest[,1] - (1.96 * result.bootest
    [,2])
result.bootest[,4] <- result.bootest[,1] + (1.96 * result.bootest
    [,2])

# Add bootstrap results to estimations and export to .xlsx
result.estimations <- cbind(result.estimations, result.bootest)
rownames(result.estimations) <- NULL
result.estimations[,c(3:ncol(result.estimations))] <-
    round(result.estimations[,c(3:ncol(result.estimations))],
    digits=6)
write.xlsx(x=result.estimations, path="estimationsBP_uprot.xlsx")
Appendix F

Mediation Analysis Script

This appendix contains the script for the final mediation analysis of our DAGs, specifically for the DAG that contains the blood pressure treatment goal as mediator.

# Load required packages
library(dplyr)
library(lubridate)
library(mice)
library(survival)
library(writexl)
library(boot)
library(Amelia)
filter <- dplyr::filter

# Impute missing data
# This CSV file contains the baseline and fifth visit values of 742 patients in long format
imputation.data <- read.csv("Z:\data\imputationdata.csv")

# Transform date and factor columns
factor.vars <- c("treatment", "sex", "caucasian", "diabetes", "cvd", "chd", "stroke", "smoke", "ntx_bas", "obesity", "med_rasi.1", "med_diur.1", "med_rasi.5", "med_diur.5")
date.vars <- c("date_bas", "dob", "outcome_date", "dat_visit.1", "dat_visit.5")
stage.filter <- c("GFR stage 3a", "GFR stage 3b", "GFR stage 4")
imputation.data[, factor.vars] <- lapply(imputation.data[, factor.vars], as.factor)
```r
imputation.data[date.vars] <- lapply(imputation.data[date.vars], as.Date)

# Rename columns to match DAG variables, prepare for Cox proportional hazards regression,
# and exclude GFR stage 1, 2 and 5
imputation.data <- imputation.data %>%
  mutate(outcome = ifelse(outcome=="Censored", 1, 2),
         outcome = as.factor(outcome),
         outcome.time = interval(date_bas, outcome_date) %>%
         months(1))

rename(ACEIARB[T0=med_rasi.1, ACEIARB_T1=med_rasi.5, BP_T0= sbp_dnmp.1, BP_T1=sbp_dnmp.5, 
        ESRD=outcome, Intervention=treatment, PU_T0=uprot.1, 
        PU_T1=uprot.5, 
        SE_T0=urinenatrium.1, SE_T1=urinenatrium.5, eGFR_T0= mdrd.175.1)

filter(ckd_stage %in% stage.filter)

# Initialize imputation
set.seed(123567)
ini <- mice(imputation.data, maxit=0)
meth <- ini$meth
pred <- ini$pred

# Setup imputation
meth["whr"] <- "%I(waist/hip)"

incl = c("age", "sex")
excl = c("pid", "waist", "hip", "obesity", "visit_type.5", "
        age_fu.1", "age_fu.5",
        "date_bas", "time.1", "time.5", "outcome_time")
pred <- quickpred(data=imputation.data, mincor=0.1, minpuc=0.25,
        include=incl, exclude=excl)

# Rows indicate the values to be imputed and columns indicate
# predictor variables
pred[c("waist","hip"), "whr"] <- 0
pred[c("waist","hip"), "bmi"] <- 1
pred["whr", c("waist","hip")]

pred[c("ldl.1", "hdl.1"), "bmi"]

pred[grepl("\([1-5]\)\ vulner\(T1\)"), rownames(pred), invert=T], grepl("\([1-5]\)\ vulner\(T1\)"), rownames(pred))]

pred[grepl("\([1-5]\)\ vulner\(T0\)"), rownames(pred), "eGFR_T0"]

pred[grepl("\([1-5]\)\ vulner\(T1\)"), rownames(pred), "mdrd.175.5"]

pred[grepl("\([1-5]\)\ vulner\(T0\)"), rownames(pred), "outcome_date"]

pred["BP_T1", "dbp_dnmp.5"]

pred["dbp_dnmp.5", "BP_T1"]

```
pred["sbp.5", "dbp.5"] <- 1
pred["dbp.5", "sbp.5"] <- 1

# Imputation of 34 datasets
imp <- parlmice(imputation.data, meth=meth, pred=pred, print=FALSE, n.core=2, n.imp.core=17)

# Mediation analysis setup

# This code is based on the code written by
# Applied Mediation Analyses: A Review and Tutorial.
# Epidemiology and Health, 39. doi:10.4178/epih.e2017035

# Step 1: Fit a Cox model with the original survival time and status (A), the mediator, and all confounders.
# Step 2: Copy the original data twice and create a new exposure variable (A*) that is equal to
#   the observed exposure in the first copy, and equal to
#   the opposite of the observed exposure in second copy.
# Step 3: Impute the unknown counterfactual survival time in the second copied dataset
#   using the new variable created in step 2 (A*)
#   making sure to limit the maximum value to the last possible observation time.
#   Append the dataset with the imputed values to the first copy of the dataset created in step 2
#   which will result in a dataset twice the length of the original dataset.
# Step 4: Fit a Cox model to the dataset created in step 3
#   using the original status (A), the created status (A*)
#   the survival time and all confounders.
# Step 5: Repeat step 3 and 4 10 times and pool the parameter estimates according to Rubin’s rules.
# Step 6: Repeat step 1–5 1,000 times with a new bootstrap sample each time
#   to obtain the variance estimate.

# This method is slightly altered to
# Perform step 1–3 10 times to obtain 10 datasets with the observed and imputed survival times.
# Perform step 4 1,000 times on each of these 10 datasets with a different bootstrap sample each time.
# Pool the resulting parameter and variance estimate accordingly.

# Finally, these steps are executed on each of our 34 imputed datasets and the result is pooled again.

# Setup maximum survival time, number of imputations and number of bootstrap samples
maxSurvivalTime <- max(imputation.data$outcome.time)
nlmp <- 10
nBoot <- 10^3

# Create a matrix with correct dimensions as base to store Cox regression results
dimParVarEst <- matrix(NA, nrow=4, ncol=2)
dimEst <- matrix(NA, nrow=7, ncol=2)

# Function that creates a dataset with the observed and imputed survival time and outcome
impCounterf <- function(workData, maxSurvTime) {
  # Step 1
  workData$InterventionTEMP <- workData$Intervention
  fitCox <-
  # This formula is specific for the blood pressure treatment goal mediator because of 'BP_T1'
  survreg(Surv(outcome_time, as.numeric(ESRD)) ~
    InterventionTEMP + BP_T1 +
    SE_T0 + BP_T0 + ACEIARB_T0 + PU_T0 + eGFR_T0,
    data=workData)

  # Step 2 and 3
  tempData1 <- workData
  tempData1$InterventionSTAR <- tempData1$Intervention
  tempData2 <- workData
  tempData2$InterventionSTAR <- as.factor(2-as.numeric(
    tempData2$Intervention))
  tempData2$InterventionTEMP <- tempData2$InterventionSTAR
  linPredTemp <- predict(fitCox, newdata=tempData2, type="linear")
  simSurvTimes <- rweibull(nrow(tempData2),
    shape=1/fitCox$scale,
    exp(linPredTemp))
  tempData2$ESRD <- as.factor(1 + 1*(simSurvTimes < maxSurvTime))
  tempData2$outcome_time <-
    round(simSurvTimes*(simSurvTimes < maxSurvTime) +
    maxSurvTime*(simSurvTimes >= maxSurvTime), digits=0)
  expData <- rbind(tempData1, tempData2)
  return(expData)
}

# Function that fits a Cox regression with the observed and counterfactual data and
# the observed and counterfactual intervention and all confounders
# with bootstrap samples, given as indices
parEst <- function(mediationData, indices) {
  # Retrieve bootstrap sample
d <- mediationData[indices, ]

  # Step 4
  fitMed <-
with (d, coxph(Surv(outcome_time, as.numeric(ESRD)) ~ Intervention + InterventionSTAR + SE_T0 + BP_T0 + ACEI_ARB_T0 + PU_T0 + eGFR_T0))

return(coef(fitMed))
}

# Mediation analysis execution

# Create a list to store results of parameter and variance estimates for each of the 34 imputed datasets
impEstimates <- array(NA, dim=c(dim(dimParVarEst), imp$m))

for (i in 1:imp$m) {
  complData <- complete(imp, i)
  # Create a list to store results of parameter and variance estimates for each of the 10 imputed datasets
  bootEstimatesTemp <- array(NA, dim=c(dim(dimEst), nImp))

  # Make 10 datasets with the created counterfactual intervention and imputed counterfactual outcome,
  # as well as the original observed intervention and outcome
  # and perform Cox regression on 1,000 bootstrap samples of each of these datasets
  for (j in 1:nImp) {
    counterfData <- impCounterf(complData, maxSurvivalTime)
    # Perform bootstrapping to obtain parameter and variance estimate for this imputed dataset
    bootstrapTemp <- boot(data=counterfData, statistic=parEst, R =nBoot, parallel="snow")
    bootstrapResult <- cbind(bootstrapTemp$t0, #coef
      apply(bootstrapTemp$t, 2, sd)) #se(coef)
    bootEstimatesTemp[, , j] <- bootstrapResult
  }

  # Compute indirect, direct, and total effect and mediated proportion
  bootEstimates <- array(NA, dim=c(dim(dimParVarEst), nImp))
  for (j in 1:nImp) {
    # Compute indirect effect
    bootEstimates[1, , j] <- bootEstimatesTemp[1, , j]
    # Compute direct effect
    # Compute total effect
    # Compute mediated proportion
  }
}
Pool results of the 10 imputed datasets

```r
bootEstMeld <- mi.meld(q=bootEstimates[,1,], se=bootEstimates[,2,], byrow=F)
parVarEst <- matrix(NA, nrow=4, ncol=2)
parVarEst[,1] <- bootEstMeld$q.mi
parVarEst[,2] <- bootEstMeld$se.mi

impEstimates[,i] <- parVarEst
```

Pool bootstrapped imputations results

Create a list to store results of pooling of the 34 imputed datasets

```r
result.mediation <- matrix(NA, nrow=4, ncol=5)
rownames(result.mediation) <- c("IE", "DE", "TE", "Q")
colnames(result.mediation) <- c("coef", "exp(coef)", "se(coef)", "lower95", "upper95")

# Pool results
impEstMeld <- mi.meld(q=impEstimates[,1,], se=impEstimates[,2,], byrow=F)
result.mediation[,1] <- impEstMeld$q.mi
result.mediation[,2] <- exp(result.mediation[,1])
result.mediation[,3] <- impEstMeld$se.mi
result.mediation[,4] <- result.mediation[,1] - (1.96 * result.mediation[,3])
result.mediation[,5] <- result.mediation[,1] + (1.96 * result.mediation[,3])

result.mediation <- round(result.mediation, 6)
result.mediation <- cbind(effect = c("IE","DE","TE","Q"), as.data.frame(result.mediation))
write.xlsx(x=as.data.frame(result.mediation), path="mediationBP_uprot.xlsx")
```