

*Digital Comprehensive Summaries of Uppsala Dissertations  
from the Faculty of Social Sciences 211*

# Design and analysis with observational data

*Protocols and modeling with the aim of causal  
inference*

PAULINA JONÉUS



ACTA UNIVERSITATIS  
UPSALIENSIS  
2023

ISSN 1652-9030  
ISBN 978-91-513-1836-3  
urn:nbn:se:uu:diva-504846



UPPSALA  
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, Kyrkogårdsgatan 10, Uppsala, Friday, 15 September 2023 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Docent Jenny Häggström (Umeå University).

### **Abstract**

Jonéus, P. 2023. Design and analysis with observational data. Protocols and modeling with the aim of causal inference. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 211. 29 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1836-3.

This thesis consists of six papers that study the design and analysis with observational data.

There is a growing interest in using real-world evidence (RWE) for regulatory purposes. The belief is that observational data can make drug development more efficient and speed up patient access to new drugs. Paper I presents a study protocol for a comparative effectiveness evaluation of two recently reimbursed hormonal treatments (NHTs) given to patients with advanced prostate cancer. The study protocol aims to present the study design, which is done without access to outcome data. Paper II presents the results from the same comparative effectiveness evaluation in clinical practice. The study shows the strength of using a matched sample and IV strategies simultaneously, even though a lack of precision using the IV analysis can be noticed.

Paper III presents a study protocol from one of a few comparative effectiveness evaluations of the NHTs against Standard of Care (SoC). Almost no patients were prescribed any of the two drugs before June 2015, as the drugs were yet to be reimbursed, creating a possibility of using historical controls. Paper IV presents the results from the comparative effectiveness evaluation. We cannot rule out that the difference in mortality maybe due to confounding. However, using a bounding strategy of the effect, we do not have sufficient evidence to show that NHT reduces mortality compared to SoC.

In Paper V, we investigate how high-dimensional data on healthcare consumption can be used when adjusting for imbalances between groups in an observational study. Our method employs a two-level neural attention model, where it is possible to include high-dimensional daily health data.

Paper VI introduces the smooth transition duration model. This model allows for analysis of policy changes when the outcome of interest is the duration until some event and when the policy change introduces different regimes, i.e., before and after the change and in the proposed model, we allow for the change between regimes to be gradual.

*Keywords:* observational studies, protocols & guidelines, epidemiology, public health, prostate disease

*Paulina Jonéus, Department of Statistics, Uppsala University, SE-75120 Uppsala, Sweden.*

© Paulina Jonéus 2023

ISSN 1652-9030

ISBN 978-91-513-1836-3

URN urn:nbn:se:uu:diva-504846 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-504846>)

*Dedicated to Majken*



# List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I Johansson P., Jonéus P. and S. Langenskiöld (2021) Study protocol for a comparative effectiveness evaluation of abiraterone acetate against enzalutamide: a longitudinal study based on Swedish administrative registers. *BMJ Open*, 11:e052610.
- II Johansson P., Jonéus P. and S. Langenskiöld (2023) Causal Inferences and real-world evidence: A comparative effectiveness evaluation of abiraterone acetate against enzalutamide.
- III Jonéus P., Johansson P. and S. Langenskiöld (2021) A study protocol for a comparative effectiveness evaluation of antiandrogenic medications against Standard of Care.
- IV Jonéus P., Johansson P. and S. Langenskiöld (2023) Novel hormonal therapy versus Standard of Care – a registry-based comparative effectiveness evaluation for mCRPC-patients.
- V Jonéus P. (2023) Estimating treatment effects with high-dimensional EHR data and historical controls.
- VI Jonéus P. and J. Lyhagen (2023) A smooth transition duration model.

Reprints were made with permission from the publishers.



# Contents

1	Introduction .....	9
2	Background .....	11
2.1	The Potential Outcomes Framework .....	11
2.1.1	The unconfoundedness assumption .....	13
2.2	High-dimensionality .....	15
2.3	Censoring .....	16
3	Scientific methodology in practice .....	18
4	Summary of papers .....	20
4.1	An effectiveness evaluation of Novel hormonal therapy for mCRPC-patients .....	20
4.2	Paper I: Study protocol for a comparative effectiveness evaluation of abiraterone acetate against enzalutamide: a longitudinal study based on Swedish administrative registers ....	21
4.3	Paper II: Causal Inferences and Real-World Evidence: A comparative effectiveness evaluation of abiraterone acetate against enzalutamide .....	21
4.4	Paper III: A study protocol for a comparative effectiveness evaluation of antiandrogenic medications against Standard of Care. ....	22
4.5	Paper IV: Novel hormonal therapy versus Standard of Care – a registry-based comparative effectiveness evaluation for mCRPC-patients. ....	23
4.6	Paper V: Estimating treatment effects with high-dimensional EHR data and historical controls .....	23
4.7	Paper VI: A smooth transition duration model .....	24
5	Acknowledgements .....	25
	References .....	27



# 1. Introduction

Humans tend to reason about their everyday actions in terms of causal effects. For example, one may believe that if doing A, then B will happen. If, instead, C is chosen, then D will be the outcome. The will to understand the world starts at an early age. Each time a child attempts a new task, for example, they can observe the consequences of actions and, eventually, learn from them. For instance, a two-year-old who gets hold of a remote control may start experimenting with it by pressing it repeatedly in different ways. Suddenly, music starts playing. The child soon realises that it is the act of pushing the play button that causes the music to start playing. At this age, already, the child has begun to reason about cause (pressing the play button) and effect (the music starts playing), see, e.g., Gopnik et al. (2001).

Before proceeding with methodologies and tools for estimating causal effects, we must specify a *cause*. Holland (1986) emphasises the need to separate the two questions of the ‘cause of an effect’ (why can I hear music now?) and the ‘effect of a cause’ (what happens if I push the play button?). There is an essential difference between understanding the causal mechanisms and the effect of a specific intervention. In this thesis, the latter question is in focus; wanting to estimate the effects of causes. One should also remember that the effects of causes are always relative to other causes. In other words, if interest lies in the effect of a treatment, no treatment (or other treatment) must be an alternative.

In papers I to V, causality is defined in terms of potential outcomes, as introduced by Neyman in the early 20th century (Neyman, 1923, 1990; Rubin, 2005) and extended by Rubin (1974). Causal effects refer to the difference in outcomes that can be attributed to a particular cause, as opposed to the outcomes that would have been observed without that cause. In the life of a toddler, again, if you did find that button (which I believe you did), you will never know what would have happened without pushing it at a specific time. You could guess but not know. This framework will be presented closely in Section 2.1. In paper VI, another type of intervention is examined with a different approach. This paper introduces the smooth transition duration model. This censoring model is designed to model the duration dependence on external variables, allowing the duration time to vary with smooth transitions over different regimes. Here we do not pre-define the time of an intervention in the model. Instead, we allow the dependency to vary over time and can thereby examine if the data supports the claim that the duration time is affected by some known, exogenous intervention. This approach will be presented more closely in Section 2.3.

In contrast to the toddlers' way of exploring cause and effect, being somewhat cumbersome and sometimes, honestly speaking, random, researchers should aim to rely on *scientific methodology*, with the key difference that the scientific process is more explicit in following steps and standards.

In general, because of the range of questions and topics covered in science, no single 'scientific method' is used in all circumstances and agreed upon among all scientists. The view taken in this thesis is that there are some important principles, at least for the type of questions studied in this thesis, that sound scientific inquiry should strive to adhere to. As discussed in Kosso (2011), a few essential aspects of scientific methods are highlighted here. First, a scientific approach should provide a systematic and organised research method. It involves formulating hypotheses and determining the design of experiments or studies before collecting and analysing data. Such a *systematic approach* ensures that research is undertaken structured and rigorously, reducing risks of potential bias and errors. Empirical testing is an important component of the scientific method. An *evidence-based approach* ensures that conclusions and claims are grounded in actual observations and measurements rather than opinions or speculation. Finally, scientific studies should be conducted to allow other researchers to replicate the analysis independently. The scientific methodology promotes the concept of *repeatability*, implying a precision and explicit presentation of experimental conditions; this allows for self-correction and refinement of scientific knowledge over time.

In parts of this thesis, these approaches follow the registration of a pre-analysis plan or study protocol. While pre-analysis plans are common when conducting experiments, such as clinical trials, they have not been widely used in observational studies. The pre-analysis plans in this thesis belong to this second category. Hopefully, they can serve as a template for researchers on how to write pre-analysis plans when working with observational data, thereby aiding transparency and reproducibility in the scientific process. The pre-publication of study plans is discussed more thoroughly in Section 3.

## 2. Background

At its best, scientific research helps us understand the world, develop new technologies, solve problems, and make informed decisions about important issues that affect society and our daily lives. Here, statistical methodologies play a crucial role. One thing often emphasised while studying statistics is that correlation does not imply causation. In other words, just because two things are correlated does not mean that one causes the other. Sometimes the correlation is spurious. It may seem that there is a clear cause- and effect, but we are only *ignoring a common cause*, which may lead to wrong conclusions. One such example can be found in Quinn et al. (1999). They find that young children who sleep with the lights on seem much more likely to develop myopia (near-sightedness) later in life and suggest that the absence of a daily period of darkness during early childhood is a potential factor in myopia development. This result was later questioned by the finding of an association between myopic parents and nursery lighting (Zadnik et al., 2000). Myopia among children does not seem to be an effect of nursery lighting, as first claimed, but a result of myopic parents seeming to prefer leaving the light on, and myopia being hereditary. It can be worth noting that in some applications, we could have a *bidirectional causality*. One example is the found association between periodontal pathogens and systemic disease. There is sufficient evidence to suggest that periodontal disease can cause adverse systemic conditions and that certain systemic diseases cause periodontal disease, e.g., Bui et al. (2019).

In the following, we aim to disentangle the effect of a cause, where the cause is some intervention, in situations where we can rule out a bidirectional causality.

### 2.1 The Potential Outcomes Framework

The history of the concept of causal effects can be traced back to the works of philosophers such as Aristotle, Hume and Kant, who explored the idea of causality and its implications (Losee, 2017). However, the formalisation of causal inference as a statistical framework emerged in the mid-20th century. This thesis builds on the Neyman potential outcomes framework and the Rubin Causal Model. This framework is precise about the definition of a causal effect, which refers to the difference in the outcome for an individual if she was treated, as opposed to if she would not have been treated, which demonstrates

the need for a contrafactual, in addition to the observed, outcome (Holland, 1986).

We begin with  $n$  units and let  $Y$  be the outcome of interest, assuming two different treatments. The potential outcomes framework assumes that each individual  $i$ ,  $i = 1, \dots, n$  has two *potential outcomes*, one for the treatment  $Y_i(1)$  and one for the control condition  $Y_i(0)$ . The causal effect is defined as the difference between these potential outcomes for individual  $i$ . Each unit is exposed to either treatment or control, where  $W_i = 1$  if unit  $i$  is given the active treatment and  $W_i = 0$  if unit  $i$  is given the control treatment.

We assume no interference between units and no hidden versions of the treatment, known as the stable unit treatment value assumption (SUTVA). The observed outcome  $Y_i$  is then a function of the treatment assignment  $W_i$ , and the potential outcomes:

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0), \quad (2.1)$$

and the effect the treatment would have on individual  $i$  is thus a comparison between  $Y_i(1)$  and  $Y_i(0)$ , for example,  $Y_i(1) - Y_i(0)$ . As seen from Equation 2.1, only one potential outcome can be observed for individual  $i$  at a given time, known as the *fundamental problem of causal inference* (Holland, 1986).

A key component is the *assignment mechanism*, which determines who receives which treatments and, thereby, which potential outcomes are realised and which are missing. Imbens and Rubin (2015) consider two basic restrictions on assignment mechanisms, which will be maintained (and discussed further) in this thesis. First, we require the assignment mechanism to imply a non-zero probability for each treatment value for every individual. Second, we assume the assignment is unconfounded. This assumption disallows the dependence of the assignment mechanism on the potential outcomes.

Randomised experiments are generally considered the gold standard for estimating the causal effects of treatments. In an ideally conducted randomised experiment, the assignment mechanism can satisfy all three restrictions on the assignment process and allows for identification of the average causal effect:

$$ATE = E(Y_i(1) - Y_i(0)) = E(Y_i | W_i = 1) - E(Y_i | W_i = 0). \quad (2.2)$$

The difference between the sample means in the two groups yields an unbiased estimator of the average causal effect. Nevertheless, randomised experiments can sometimes be infeasible to conduct. There can be ethical considerations not allowing for randomisation of treatment, or it is impossible to conduct in practice. Consequently, we often rely on non-randomised, observational studies. In contrast to classical randomised experiments, the assignment mechanism may not be controlled or known by the researcher (Imbens and Rubin, 2015). Observational studies, or real-world data (RWD), can, in addition, be a complement to RCTs and, for example, accelerate drug development.

Each unit has a vector of characteristics, or covariates, denoted  $X_i$ . Now, in addition to SUTVA, we assume *unconfoundedness*

$$(Y_i(0), Y_i(1)) \perp W_i | X_i, \quad (2.3)$$

and *overlap*

$$0 < e(X_i) < 1, \quad (2.4)$$

for all units, where  $e(X_i) = P(W_i = 1 | X_i)$  is the propensity score. The propensity score plays a central role (Rosenbaum and Rubin, 1983). If treatment assignment is unconfounded given  $X_i$ , then treatment assignment is unconfounded given the propensity score:

$$(Y_i(0), Y_i(1)) \perp W_i | e(X_i), \quad (2.5)$$

that is, instead of adjusting for all pre-treatment variables, it is sufficient to adjust for the propensity score, allowing for identification of average causal treatment effects.

### 2.1.1 The unconfoundedness assumption

The term *confounding* comes from the Latin word *confundere* and can be explained as ‘to mix-up’. For the English philosopher Mill, confounding meant ‘intermixture of causes’, which he defined as two or more causes, ‘modifying the effects of one another’ (Morabia, 2011). Suppose, for example, we want to assess the impact of a new drug on patient mortality. To do so, we compare the mean mortality between two groups: patients who received the new drug (treatment group) and those who did not (control group). Suppose further that this drug is not suitable for older patients. If we then fail to account for age, where the treated patients are systematically younger than the controls, the observed difference in sample means is subject to bias in estimating the average causal effect. However, by the unconfoundedness assumption, the causal effect is identified if we know and can observe the entire range of confounding variables.

In a randomised experiment, the treated and control groups are guaranteed to be identical in terms of background covariates in expectation and different methods aim to replicate this in observational studies, such as by matching. By choosing well-matched samples of the original treated and control groups, the researcher reduces bias due to differences in covariate distribution among treated and controls. Matching has an intuitive appeal; the only difference if we compare individuals identical in all covariates is the treatment status (Stuart, 2010). The problem is choosing treated and controlled subjects to achieve balanced groups. In one-to-one exact matching, one treated individual

is matched with one control with the same covariate values, resulting in well-matched samples. However, this method has the disadvantage of throwing away data, as it may only be possible to match exactly on a few variables. Numerous matching algorithms are available to achieve balanced groups. The main differences between these algorithms lie in how the differences between two individuals are measured and what algorithm is chosen to minimise it (Rosenbaum, 2020).

Balance on all covariates can be achieved by matching or weighting on the propensity score alone. The matter of iteratively checking the specification of the propensity score model is not controversial in the theoretical literature on matching. Instead, one of the major benefits of matching is that because outcome data are not used in propensity score estimation, one may consider various models of treatment assignment (Rubin, 2008). The genetic matching algorithm provides an alternative to the iterative process of checking and improving overall covariate balance, which can guarantee asymptotic convergence to the optimally matched sample. By construction, the algorithm will improve covariate balance, if possible, as measured by the particular loss function chosen to measure balance (Diamond and Sekhon, 2013). A covariate balance method that circumvents modeling the propensity score is denoted entropy balancing (Hainmueller, 2012). This approach uses a re-weighting scheme that assigns a scalar weight to each sample unit. The re-weighted groups then satisfy balance constraints imposed on the covariate distributions' sample moments. The constraints ensure that the re-weighted groups are balanced on the specified moments.

Unconfoundedness requires adjusting for differences in values of observed pre-treatment variables, removing systematic biases from comparisons between treated and controls. Unfortunately, this assumption is not testable. When conducting an observational study, the researcher should always be able to answer why two individuals, who are identical on measured covariates, receive different treatments. The answer can be that there is some natural randomness; for example, treatment can be given as a result of a lottery or two individuals, identical on measured covariates, can be born in different cohorts or at different sides of a municipality border, resulting in different treatments. However, if the potential outcomes are influenced by unobservable covariates, which affect the potential outcomes, then treated and untreated are not directly comparable, even after adjusting for observable covariates.

One way to assess this concern is to design the comparison group so that the unobservables are *likely* to be balanced. By carefully balancing on a large set of important and observed covariates, the influence of the unobservable covariates can, in some situations, be negligible. There are, in addition, different ways to assess the unconfoundedness assumption by, for example, pseudo-outcomes and robustness of estimates (Imbens and Rubin, 2015).

There are alternative approaches when the assumption of no unmeasured confounders is not met. For example, an Instrumental Variables (IV) approach

can be used, where the researcher identifies an instrumental variable correlated with the treatment but not directly associated with the outcome, except through its influence on the treatment (Angrist and Krueger, 2001). For example, McClellan et al. (1994) are interested in the causal effect of more intensive treatments on mortality. Patients who receive different treatments are assumed to differ in unobservable health characteristics. An IV approach is chosen, as the authors find that the distance to the hospital strongly predicts how intensively a patient will be treated. If the distance to the hospital is uncorrelated with the potential outcomes but affects treatment, then the distance is a valid instrument.

There is an increased interest in combining the two methods of matching samples and instrumental variables estimation. However, it is crucial to be aware of the impact on the results of the analytical technique chosen and its appropriate use in observational studies, as underlying assumptions of the technique used can be violated. For example, unmeasured confounders may not be balanced in case of propensity score use, or instruments may be related to the outcome directly or to unmeasured confounders, or is not strong enough, in case of instrumental variables estimation use (Laborde-Castérot et al., 2015). Therefore, discussing assumptions, weaknesses, and strengths is essential when combining different methods.

## 2.2 High-dimensionality

The assumption of unconfoundedness is often only plausible after conditioning on many confounders. When the dimension of the covariates grows, it quickly becomes difficult to find matching groups with overlap in all dimensions. Therefore, the researcher often needs some dimension reduction of potential confounders. As mentioned in Section 2.1, an important advance was made by Rosenbaum and Rubin (1983), introducing the propensity score.

All matching methods assume no unobserved differences between the treatment and control groups, conditional on the observed covariates. Therefore, including all variables related to the treatment assignment and the outcome in the matching procedure is essential. The selection process should be done without access to the observed outcomes to avoid variable selection based on estimated effects. It should be based on previous research and scientific understanding (Stuart, 2010).

In large part of this thesis, we have a high-dimensional dataset comprising all healthcare consumption and socioeconomic background variables for the sample of interest. Therefore, dimension reduction is needed, of which three methods are highlighted here. First, a qualitative study, including expert views and previous research, aims to pinpoint the important variables to be included in the analyses. Second, we employ an exploratory factor analysis in combination with a Least Absolute Shrinkage and Selection Operator (LASSO)

regression to estimate the propensity score. The estimated factors and propensity score can be included in the matching and weighting approaches to obtain balanced groups. These two methods thus require pre-selection and construction of variables from the high-dimensional data on health. Therefore, we also explore the possibility of using the raw health data in a neural network model to predict the probability of treatment.

## 2.3 Censoring

In many studies within medical research, the primary outcome is the time until an event of interest. If the event occurred for all individuals, many analysis methods would be applicable. However, when individuals have yet to experience the event at the end of the study, known as censoring, survival analysis methods might be necessary (Yamaguchi, 1991). Modeling survival is thus not restricted to medical research, and since the early 1980s, the econometric applications include, for example, duration of unemployment or duration in labour market programs (Van den Berg, 2001). Duration data then usually measure the time an individual spends in a specific state, denoted duration time, and we want to model time until leaving that state (Kiefer, 1988).

Survival data are generally modelled in terms of survival and hazard. Assuming  $T$  is continuous time until some event, the survival function is defined by

$$S(t) = P(T > t), 0 < t < \infty. \quad (2.6)$$

The hazard function  $h(t)$  represents the instantaneous rate at which events occur if an event has not yet occurred at time  $t$  (Kalbfleisch and Prentice, 2002). Two important features characterise duration data. The first important feature is, as mentioned above, that the data may be censored. A second characteristic is that the values of many covariates in these models may change over time.

There are two major groups of methods for analysing hazard rates: fully or partially parametric methods and non-parametric methods. In the parametric case, the hazard is a function of covariates  $x_i$ , and the functional form is commonly specified as

$$h(t, x) = \lambda_0(t)e^{\beta x} \quad (2.7)$$

where  $x$  is a vector of independent observable variables. Different assumptions about the baseline hazard,  $\lambda_0(t)$ , can be made; it is unspecified in a Cox relative risk model but can be parametrised so that time is assumed to follow some known distribution, commonly, an exponential with  $\lambda_0(t) = 1$ , a Weibull with  $\lambda_0(t) = \alpha t^{\alpha-1}$  or a log-logistic with  $\lambda_0(t) = \frac{\alpha t^{\alpha-1}}{1+t^\alpha}$  specification is used.

Often we are interested in the effects of policy changes, that is, treatments that affect all individuals in a given population. If the outcome under study

is the duration in a specific state, then, in general, simple evaluations, e.g. a before and after evaluation, or a comparison of a difference in an affected population to that of a difference in an unaffected population, is generally not possible. The reason is that the duration examined could exist in both policies.

Consider, for example, an analysis of stricter monitoring of the unemployed with unemployment insurance on the duration of unemployment. The unemployment duration is then measured under both policies in the stock of unemployed at the time of the policy change. A further concern with the analysis is that the policy may take time to implement but will be gradual from one regime to another. This situation will likely be the case with stricter monitoring, as caseworkers take time to implement the new monitoring rules. Paper VI introduces a model that allows for analyses of gradual effects of policy changes on a duration outcome, when we have a policy change that is best described by different regimes.

### 3. Scientific methodology in practice

*Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passions, they cannot alter the state of facts and evidence.*

John Adams, 1770

In recent decades, concerns have been raised regarding empirical science. The replication crisis in social sciences refers to the failure to replicate a large fraction of published experiments. Ioannidis (2005) argues in the paper 'Why Most Published Research Findings Are False' that published research findings suffer from, for example, publication bias, multiple hypothesis testing and low statistical power. *Publication bias* occurs when the study results affect the decision to publish or distribute the research. The consensus is that statistically significant results are easier to publish. In addition, non-replicable publications are also found to be cited more than replicable ones. Multiple hypothesis testing, and *p-hacking*, occurs when the researcher searches for statistically significant results. See, e.g., Brodeur et al. (2020); Serra-Garcia and Gneezy (2021). Scientific methodologies should allow us to accept unexpected results. This section discusses different proposed tools to reduce the risks of the mentioned types of biases.

As many subjective decisions are part of the research process and can affect the outcomes, the best defence against subjectivity is to expose it. Transparency in data, methods, and process allows the rest of the community to see the decisions made by the researcher, question them, offer alternatives, and test them in further research (Silberzahn et al., 2018). Here, following Rubin (2007), we define the *design stage* as all contemplating, collecting, organising, and analyses of data that takes place before seeing any outcome data. Carefully implementing the design is essential for drawing objective inferences for causal effects in practice. Paper II and IV in this thesis build on pre-analysis plans published before access to outcome data. The pre-analysis plans, or study protocols, are presented in Paper I, Paper III and in Johansson et al. (2021).

Pre-registration of studies highlights the importance of the design of observational studies, and publication of statistical analyses plans for observational studies has been intensely debated in the last decade. A well-written study plan can serve as a template for researchers. It also helps ensure transparency and reproducibility in research by providing a detailed plan that others can review and follow. In addition, writing a pre-analysis plan may have the benefit of forcing the researchers to think through their hypotheses beforehand, improving the quality of the research design and data collection approach.

Not all are expressly positive. For example, some authors fear that pre-registration of the analysis plans could create false security that data are of high quality, discourage publication of important accidental findings, and delay publications due to bureaucratic procedures (Hiemstra et al., 2019). However, pre-registration of studies serves at least three aims to improve research findings' credibility and reproducibility. It highlights which analyses were planned a priori, ensures that methods can be replicated and findings confirmed and reduce selective outcome reporting and publication bias. The latter addresses the issue of p-hacking, as the researcher must specify in advance how data will be analysed. See, e.g., Imbens (2021); Casey et al. (2012); Banerjee et al. (2020); Olken (2015) for a discussion of the use of pre-analyses plans.

## 4. Summary of papers

### 4.1 An effectiveness evaluation of Novel hormonal therapy for mCRPC-patients

Papers I to V in this thesis address the same empirical problem. The effectiveness of two recently reimbursed hormonal treatments given to patients with advanced prostate cancer is evaluated, comparing them to each other and Standard of Care (SoC). Prostate cancer (PC) is reported to be the most commonly diagnosed form of cancer in Sweden and is also the fifth leading cause of death in men worldwide. Almost all mortalities arise when the patients have progressed to the advanced stage of metastatic castrate-resistant prostate cancer (mCRPC). Various treatment alternatives are available for patients with mCRPC. However, in the last two decades, chemotherapy and novel hormonal therapy (NHT) have revolutionised the treatment in mCRPC patients (Heidenreich et al. (2013); Rawla (2019); Socialstyrelsen (2018)).

There is a growing interest in using real-world evidence (RWE) for regulatory purposes. The belief is that observational data can make drug development more efficient and speed up patient access to new drugs. The evaluations in this thesis concern the use of enzalutamide (ENZ) and abiraterone acetate (AA) in clinical practice from June 2015, corresponding to when these drugs were reimbursed for mCRPC patients in Sweden. Data are collected from population registers administrated by the National Board of Health and Welfare (NBHW), Statistics Sweden (SCB), and the National Prostate Cancer Register (NPCR).

The population is restricted to all men in the NBHW register with a prostate cancer diagnosis before 2017. All in and out-patient care visits, including length of stay and registered ICD-10 codes, are provided in the available data. Similarly, all collected prescriptions (ATC codes) are included. Additionally, we combine the health measures with a vast set of socioeconomic variables that might affect treatment assignment. What is not registered in the data is the mCRPC diagnosis. The substantial dataset on the included patients' health and socioeconomic factors combined with a qualitative study enables the groups to be balanced on what is deemed to be important covariates. The outcome data are added after the publication of pre-analysis plans. The main outcome of interest is mortality.

There are multiple methodological challenges to address when evaluating the NHTs, but also advantages in defining a valid design. Papers I to V aim to offer a solution to some of these challenges.

## 4.2 Paper I: Study protocol for a comparative effectiveness evaluation of abiraterone acetate against enzalutamide: a longitudinal study based on Swedish administrative registers

Paper I presents a study protocol for a comparative effectiveness evaluation of AA against ENZ in clinical practice. The study protocol aims to present the study design, which is done without access to outcome data. Here, a matching approach is used. Therefore, the design builds on the assumption that observed covariates capture the patients' health and socioeconomic status. In other words, the observed data contain everything that jointly determines treatment and potential outcomes. Furthermore, all covariates included are measured before the first prescription of AA or ENZ and can not be affected by treatment.

To obtain balanced groups, we use the genetic matching algorithm. We first need to summarise the information in the high-dimensional data when specifying the variables to be included in the algorithm. The dimension reduction is done in three different ways. First, age, waiting time and other covariates found in the qualitative work and all important discrete variables are included directly. Second, exploratory factor analysis is introduced to reduce the dimension of our large subset of continuous variables. We end up with nine factors included in the matching scheme. Finally, the genetic matching algorithm can include the estimated propensity score as a covariate. To enable a flexible specification of the propensity score, we estimate a logit model using LASSO regression.

The results of the matched samples are evaluated. None of the covariates shows a large difference before matching. Still, it can be noted that both the prevalence of diabetes and secondary malignant neoplasms, including age interactions, and acute myocardial infarction seem to differ, with a standardised difference of above 10 per cent, between the two treatment groups. The matching is successful in removing these differences as well as the difference in the estimated propensity score.

The pre-analysis plan also introduces the planned analyses, including sensitivity analyses to be evaluated in Paper II.

## 4.3 Paper II: Causal Inferences and Real-World Evidence: A comparative effectiveness evaluation of abiraterone acetate against enzalutamide

Paper II presents the results from the comparative effectiveness evaluation of AA against ENZ in clinical practice. The main analysis is based on a matching strategy, and a sensitivity analysis uses differences in prescription practices

across counties in an IV setting. This paper aims to illustrate how an observational study can be based on a pre-published protocol when outcome data are added after the publication of the pre-analysis plan. Paper II builds on two pre-analysis plans, where the matching strategy is presented in Paper I and the sensitivity analysis is presented in Johansson et al. (2021).

The results from the two analyses show an increased mortality risk from prescribing AA compared to ENZ. In addition, the matched sampling analysis also suggests an increased risk of skeleton-related events. Further, the study shows the strength of using a matched sample and IV strategies simultaneously, even though a lack of precision using the IV analysis can be noticed.

When assessing the identifying assumptions in the designs and analyses, we find no statistically significant effects on covariates from the NPCR, using the same regression analysis as in the main analysis. Therefore, we have no reason to believe that available data from the population registers are insufficient to control for confounding bias.

#### 4.4 Paper III: A study protocol for a comparative effectiveness evaluation of antiandrogenic medications against Standard of Care.

Paper III presents a study protocol from one of a few comparative effectiveness evaluations of the NHTs against SoC. Almost no patients were prescribed any of the two drugs before June 2015, as the drugs were yet to be reimbursed, creating a possibility of using historical controls. This approach solves the problem when patients who are not treated when the drugs are available are unsuitable for forming a control group. The study protocol aims to present the study's design; this is done without access to outcome data. An entropy balancing scheme is chosen to obtain balanced groups. The rich registry data and the lack of the NHTs as an option for the comparisons provide support for the validity of the design. The defined comparison group should be as similar as possible to the treatment group, and the two should have been provided with the same quality of health care. If the comparison group is sampled close in time to the NHT group, these two restrictions may hold.

For each month after the PC diagnosis, we create covariates describing the monthly health of the patients. For the treatment group, this is done up to the date of prescription of an NHT. For the comparisons, on the other hand, as we do not have information on the mCRPC diagnosis, health data are created for all periods up until 36 months after the diagnosis. Covariates measuring the health progression and pre-diagnosis variables are then included in the entropy balance scheme. Finally, weights are estimated directly from imposed balance constraints on the covariates. After re-weighting, the total weight of

the comparison group exactly matches that of the treatment group for a given treatment month.

While the design yields balanced observed covariates, this is an observational study with the usual limitations. In particular, one must recognise the possibility that unobserved confounders are not balanced. For this reason, placebo regressions are introduced.

#### 4.5 Paper IV: Novel hormonal therapy versus Standard of Care – a registry-based comparative effectiveness evaluation for mCRPC-patients.

Paper IV presents the results from the comparative effectiveness evaluation of NHT against SoC in clinical practice. The evaluation is limited to patients treated within three years after diagnosis, i.e., with a very quickly progressed prostate cancer. For this group, we find a substantial increase in mortality for NHT patients if prescribed an NHT rather than being given SoC.

We rely on population registers and experts when pre-specifying the covariates used in the entropy balancing. Based on the results from one of the two placebo regressions, we cannot rule out that the difference in mortality may be due to confounding. Instead, using a bounding strategy of the effect, we do not have sufficient evidence to show that NHT reduces mortality compared to SoC.

#### 4.6 Paper V: Estimating treatment effects with high-dimensional EHR data and historical controls

In Paper V, we investigate how high-dimensional data on healthcare consumption can be used when adjusting for imbalances between groups in an observational study. In certain studies, treatment is prescribed almost deterministically based on the patient’s health, making it challenging to find a comparable control group within the current cohort. In contrast to the approach presented in Paper III, where an entropy balance scheme is used to achieve balance between the treated in the current cohort and a group of historical controls, we present how information on the not yet treated in the current cohort can help.

In health evaluations, a high degree of background information on health is commonly incorporated to encompass all relevant confounding variables. Our method employs a two-level neural attention model denoted RETAIN, introduced by Choi et al. (2016), when predicting the patients’ treatment probabilities in the historical control group. This model can include high-dimensional daily health data without constructing and pre-selecting among variables.

We apply our approach to the evaluation presented in Paper III. We find that the RETAIN model can pick up important covariates affecting prescription of

the NHTs, as judged by experts. The estimated effects on mortality align with earlier findings.

#### 4.7 Paper VI: A smooth transition duration model

Paper VI introduces the smooth transition duration model. This model allows for analysis of policy changes when the outcome of interest is the duration until some event and when the policy change introduces different regimes, i.e., before and after the change. We illustrate the method using duration data from the Queensland electricity market. In 2007, the deregulation of the Australian electricity market began when the Queensland government started to implement a move towards full retail competition. Energy prices in South East Queensland were deregulated in 2016. A highly relevant question is if deregulation led to increased incentives of strategic behaviour by market participants and if this affected electricity prices.

In the proposed model, we allow for the change between regimes to be gradual. By modeling this duration time, we investigate if there is a change over time and if the behaviour of market participants is affected by regulatory changes. The results show that there seems to be more than one transition in the study period. A transition from one regime to another is found to occur at the beginning of 2016 and in the first half of 2007, indicating an increase in the rate at which price events occur. As the Australian electricity market deregulation began in 2007, with energy prices in South East Queensland deregulated in 2016, the findings align with what would be expected.

## 5. Acknowledgements

So many people have a part in this work. I want to express my deepest gratitude to the following individuals whose support, guidance, and encouragement have contributed to the completion of this thesis.

My main supervisor, Mattias Nordin, deserves special recognition for his boundless optimism (I am not exaggerating here) and engaging mentorship. Your expertise and vast knowledge are truly inspiring. I am immensely thankful for your support throughout these years; our friendship and your belief in my abilities have been invaluable. I also want to extend my sincere gratitude to Per Johansson, my second supervisor, whose impact on my academic and personal growth cannot be overstated. Unlike myself, who tends to be more silent in my curiosity, Per's inquisitive nature has inspired and challenged me in so many ways. I am so glad I have got the opportunity to work with you.

I want to thank all my PhD colleagues and friends, only mentioning a few here. Ingrid, you have been an irreplaceable presence in my life during these years. I cannot imagine a better-shared office or a more understanding friend. Your support during my most challenging moments has kept me going. Lukas and Johan, thank you for being there for me, providing assistance and laughter. Our friendship has been a constant source of support through my (and yours?) ups and downs. Sebastian, your dedication has profoundly impacted my work. Our collaborative efforts and the papers we have written together were so beneficial in writing this thesis. Alexander and Mårten, I've appreciated your company since we started, sharing the smallest office in the department.

I want to thank Inger Persson and Ingeborg Waernbaum as directors of studies at the PhD level and dear colleagues for their structure, encouragement, and guidance. Your commitment to fostering a nurturing academic environment has been invaluable. To Lisbeth Hansson and Johan Lyhagen, thank you for encouraging me to embark on this journey. I am grateful to Thommy Perlinger for his dedication to teaching and for giving me responsibilities as a student. To all at the Department of Statistics, thank you for the courses, feedback, and encouragement that have enhanced my academic growth and development. I want to thank Sophie Langenskiöld, taking me on the project that forms a substantial part of this thesis.

Outside the academic sphere, there are so many people that I would like to thank.

Katina, your availability and support have been a lifeline throughout this process (and my life). Your belief in me has been invaluable, especially during my extreme self-doubt moments. I am forever grateful for your presence and

the laughter you bring when I need it most, and for providing me with self-reflection when required.

To Anja, Elin, Linda, Jessica and Rebecca (among many others!), thank you for showing me the wonders of Uppsala from the day we started pol kand together, in what feels like a million years ago. Without your friendship and support, I would not have been able to endure the challenges and joys of this city, and I am fortunate to have you all in my life. Amanda, thank you for always being a lifeline, bringing me food, wine, and good advice. Anders, I want to express my gratitude (to you and your family) for your never failing support.

Isabell, thank you for always being there, you are a rock. Elin, Sofia, Klara, and Evelina, thank you for reminding me of the world beyond the academic sphere; you have provided much-needed balance and perspective. I cannot emphasise enough how much I have relied on our friendship over the years.

I want to thank my parents, your unwavering confidence and belief in me have always been a driving force. Thank you, Carl and Ulrika, and a special thanks to Bettan. Also, although they are no longer with us, I want to express my appreciation for mormor and morfar; their memory continues to inspire me. *Tack for allt jag fick medskickat.*

Erik, you have been able to lift my spirit and ground me back to reality when I've doubted and hesitated (and sometimes complained). Just before you entered my life, I thought I'd never finish this PhD project but you helped me find my way back. Last but not least, Majken, you are my sunshine and a never-ending source of motivation. I love you.

# References

- Angrist, J. D. and Krueger, A. B. (2001), 'Instrumental variables and the search for identification: From supply and demand to natural experiments', *Journal of Economic Perspectives* **15**(4), 69–85.
- Banerjee, A., Duflo, E., Finkelstein, A., Katz, L. F., Olken, B. A. and Sautmann, A. (2020), In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics, Working Paper 26993, National Bureau of Economic Research.
- Brodeur, A., Cook, N. and Heyes, A. (2020), 'Methods matter: P-hacking and Publication Bias in Causal Analysis in Economics', *American Economic Review* **110**(11), 3634–3660.
- Bui, F. Q., Almeida-da Silva, C. L. C., Huynh, B., Trinh, A., Liu, J., Woodward, J., Asadi, H. and Ojcius, D. M. (2019), 'Association between periodontal pathogens and systemic disease', *Biomedical Journal* **42**(1), 27–35.
- Casey, K., Glennerster, R. and Miguel, E. (2012), 'Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan', *The Quarterly Journal of Economics* **127**(4), 1755–1812.
- Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A. and Stewart, W. (2016), RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism, in D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 29, Curran Associates, Inc.
- Diamond, A. and Sekhon, J. S. (2013), 'Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies', *Review of Economics and Statistics* **95**(3), 932–945.
- Gopnik, A., Sobel, D. M., Schulz, L. E. and Glymour, C. (2001), 'Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation.', *Developmental Psychology* **37**(5), 620.
- Hainmueller, J. (2012), 'Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies', *Political Analysis* **20**(1), 25–46.
- Heidenreich, A., Pfister, D., Merseburger, A., Bartsch, G. et al. (2013), 'Castration-resistant prostate cancer: where we stand in 2013 and what urologists should know.', *European Urology* **64**(2), 260–265.
- Hiemstra, B., Keus, F., Wetterslev, J., Gluud, C. and van der Horst, I. C. (2019), 'Debate-statistical analysis plans for observational studies', *BMC Medical Research Methodology* **19**, article no: 233.
- Holland, P. W. (1986), 'Statistics and causal inference', *Journal of the American Statistical Association* **81**(396), 945–960.
- Imbens, G. W. (2021), 'Statistical significance, p-values, and the reporting of uncertainty', *The Journal of Economic Perspectives* **35**(3), 157–174.

- Imbens, G. W. and Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.
- Ioannidis, J. P. (2005), 'Why most published research findings are false', *PLoS medicine* **2**(8), e124.
- Johansson, P., Jonéus, P. and Langenskiöld, S. (2021), 'A study protocol for an instrumental variables analysis of the comparative effectiveness of two prostate cancer drugs', *arXiv:2110.04164*.
- Kalbfleisch, J. and Prentice, R. (2002), *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics, Wiley.
- Kiefer, N. M. (1988), 'Economic duration data and hazard functions', *Journal of Economic Literature* **26**(2), 646–679.
- Kosso, P. (2011), *A summary of scientific method*, Springer Science & Business Media.
- Laborde-Castérot, H., Agrinier, N. and Thilly, N. (2015), 'Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review', *Journal of Clinical Epidemiology* **68**(10), 1232–1240.
- Losee, J. (2017), *Theories of causality: from antiquity to the present*, Routledge.
- McClellan, M., McNeil, B. J. and Newhouse, J. P. (1994), 'Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?: Analysis Using Instrumental Variables', *JAMA* **272**(11), 859–866.
- Morabia, A. (2011), 'History of the modern epidemiological concept of confounding', *Journal of Epidemiology & Community Health* **65**(4), 297–300.
- Neyman, J. (1923, 1990), 'On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9', *Statistical Science* **5**(4), 465 – 472, 1990.
- Olken, B. A. (2015), 'Promises and perils of pre-analysis plans', *Journal of Economic Perspectives* **29**(3), 61–80.
- Quinn, G. E., Shin, C. H., Maguire, M. G. and Stone, R. A. (1999), 'Myopia and ambient lighting at night', *Nature* **399**(6732), 113–114.
- Rawla, P. (2019), 'Epidemiology of prostate cancer', *World Journal of Oncology* **10**(2), 63–89.
- Rosenbaum, P. R. (2020), 'Modern algorithms for matching in observational studies', *Annual Review of Statistics and Its Application* **7**, 143–176.
- Rosenbaum, P. R. and Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.
- Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**(5), 688.
- Rubin, D. B. (2005), 'Causal inference using potential outcomes: Design, modeling, decisions', *Journal of the American Statistical Association* **100**(469), 322–331.
- Rubin, D. B. (2007), 'The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials', *Statistics in Medicine* **26**(1), 20–36.
- Rubin, D. B. (2008), 'For objective causal inference, design trumps analysis', *The Annals of Applied Statistics* **2**(3), 808 – 840.
- Serra-Garcia, M. and Gneezy, U. (2021), 'Nonreplicable publications are cited more than replicable ones', *Science Advances* **7**(21), eabd1705.

- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E. et al. (2018), 'Many analysts, one data set: Making transparent how variations in analytic choices affect results', *Advances in Methods and Practices in Psychological Science* **1**(3), 337–356.
- Socialstyrelsen (2018), 'Cancer i siffror 2018 populärvetenskapliga fakta om cancer, Cancerfonden och Socialstyrelsen i samarbete', <https://www.cancerfonden.se/om-cancer/statistik/cancer-i-siffror>. 2018-6-10.
- Stuart, E. A. (2010), 'Matching methods for causal inference: A review and a look forward', *Statistical Science* **25**(1), 1–21.
- Van den Berg, G. J. (2001), Chapter 55 - Duration Models: Specification, Identification and Multiple Durations, Vol. 5 of *Handbook of Econometrics*, Elsevier, pp. 3381–3460.
- Yamaguchi, K. (1991), *Event history analysis*, Sage Publications.
- Zadnik, K., Jones, L. A., Irvin, B. C., Kleinstein, R. N., Manny, R. E., Shin, J. A. and Mutti, D. O. (2000), 'Myopia and ambient night-time lighting', *Nature* **404**(6774), 143–144.

# Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 211*

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences”.)

Distribution: [publications.uu.se](http://publications.uu.se)  
urn:nbn:se:uu:diva-504846



ACTA UNIVERSITATIS  
UPSALIENSIS  
2023