Regression Discontinuity Design for AI-assisted histopathology data

Master of Science in Technology
Thesis
University of Turku
Department of Computing, Faculty of Technology
Digital Health
Master's Degree Programme in Information and Communication Technology
2023
Jouni Kettunen

Supervisors:
Leo Lahti
Guilhem Sommeria-Klein

UNIVERSITY OF TURKU
Department of Computing, Faculty of Technology

JOUNI KETTUNEN: Regression Discontinuity Design for AI-assisted histopathology data

Master of Science in Technology Thesis, 83 p., 16 app. p.
Digital Health
Master's Degree Programme in Information and Communication Technology
March 2023

The aim of the thesis was to review use of quasi-experimental regression discontinuity design method in histopathological data-analysis and to assess the technical possibility of using the method with algorithmic data. Regression discontinuity design is a rarity in medical research and little used in histopathological data-analysis. The data produced in histopathological analysis relies on cut-offs when reviewing eligibility to diagnose/treatment and fulfils the main assumption of continuity. Also, data around the divisive cut-off can be proven to be randomly divided. Proof-of-concept for regression discontinuity design was done using digitalized breast tissue slides (n = 129) analyzed in Aiforia Cloud. The observations were divided by using a proliferation protein Ki-67's positivity ratio as a cut-off; $< 14\%$ were in moderate group (pathologist n = 37, algorithm n = 47) and $\geq 14\%$ positives (pathologist n = 92, algorithm n = 82). Outcome was chosen to be severity of the decease (CFR%), derived from five-year survival. Thesis used a pre-made API to fetch algorithmic results to R, where the data handling and analysis were done. For realistic study one needs thousands of observations, thus used dataset with n = 129 subjects was too small, also there were only nine observations in outcome. For the reasons, thesis can't give any clinically relevant judgments. Technically it was possible to make a pipeline from Aiforia Cloud to R and proceed to regression discontinuity design. In the future one might consider repeating this study with enough observations and more robust outcome.

Keywords: regression discontinuity design, quasi-experiment, Ki-67, causal effect, breast cancer

TURUN YLIOPISTO
Tietotekniikan laitos, Teknillinen tiedekunta

Jouni Kettunen: Regression Discontinuity Design for AI-assisted histopathology data

Diplomityö, 83 s., 16 liites.
Terveysteknologia
Tietotekniikka, diplomi-insinööri
Maaliskuu 2023

Diplomityön tarkoituksena oli selvittää kirjallisuuskatsauksen ja konseptitodistelun avulla kvasikokeisiin kuuluvan regressioepäjatkuvuusasetelman käyttöä histopatologisen aineiston analysoinnissa. Kirjallisuuskatsauksen perusteella regressioepäjatkuvuusasetelma on vähän käytetty menetelmä lääketieteellisessä tutkimuksessa, eikä sitä ole juurikaan sovellettu histopatologian alalla, toisin kuin ekonomiassa ja sosiologiassa. Edellä mainittujen alojen tutkimus kuitenkin osoittaa regressioepäjatkuvuusasetelman olevan monipuolinen menetelmä useisiin tutkimuskysymyksiin. Histopatologiassa tuotettava syöpädiagnostiikkaa tukeva tieto soveltuu kirjallisuuden perusteella käytettäväksi regressioepäjatkuvuusasetelmassa, sillä tärkeimmät ennakko-oletukset jatkuvuuden ja jakavan kynnysarvon satunnaisuuden suhteen täyttyvät. Regressioepäjatkuvuusasetelma-analyysi toteutettiin käyttäen digitalisoiduista rintakudosleikkeistä Aiforia Cloud pilvipalvelussa tuotettua tietoa. Havainnot (n = 129) jaettiin kahteen ryhmään käyttäen prosenttiosuutta proliferaatioproteiini Ki-67 värjäytyneistä soluista jakavana tekijänä siten, että $<$14% olivat kohtalaisia (patologidatassa n = 37, algoritmilla n = 47) ja $\geq$14% positiivisia (patologidatassa n = 92, algoritmilla n = 82). Lopputulemaksi valittiin taudin vakavuutta kuvaava CFR% luku, joka johdettiin potilaiden viiden vuoden selviytymisseurannasta. Työssä käytetty ohjelmointirajapinta haki algoritmilla analysoidut tulokset pilvipalvelusta ja tietoja käsiteltiin R-ohjelmointikielellä. Regressioepäjatkuvuusasetelma-analyysiin luotettava toteutus vaatii tuhansien havaintojen joukkoja, joten diplomityössä täytetty havaintojoukko oli liian pieni. Myös käytetty lopputulema oli osittain soveltumaton johtuen rintasyöpäpotilaiden hyvästä eloonjäännistä viiden vuoden aikana, näistä syistä työn kliinistä merkittävyyttä ei voida luotettavasti arvioida. Teknisesti toteutus oli onnistunut osoittaen Aiforia Cloud pilvipalvelusta voitavan noutaa analyysituloksia, joita pystytään käyttämään tehtäessä regressioepäjatkuvuusasetelma-analyysia. Jatkotutkimuksena suositellaan kokeen toistamista käyttäen suurempaa havaintojoukkoa sekä soveltuvampaa lopputulemaa.

Asiasanat: regressioepäjatkuvuusasetelma, kausaalivaikutus, kvasikoe, Ki-67, rintasyöpä

Contents

Acronyms

AI Artificial intelligence.

ASCO American Society of Clinical Oncology.

ATE Average Treatment Effect.

BW Bandwidth.

CAD Computer Aided Diagnostics.

CAP College of American Pathologists.

CE-IVD Conformité européenne for in vitro diagnostic devices.

CFR Case Fatality Rate.

CI Confidence Interval.

ER Estrogen receptor.

ESMO European Society for Medical Oncology.

GG Grading Group.

GS Gleason score.

H&E Hematoxylin-Eosinophil.

HER2 Receptor tyrosine-protein kinase erbB-2.

HUS Helsinki University Hospital.

IHC Immunohistochemical.

IKWG International Ki67 in Breast Cancer Working Group.

ISUP International Society of Urological Pathology.

Ki-67 Proliferation marker protein Ki-67.

NDA Non-disclosure agreement.

PD-L1 Programmed death-ligand 1.

PR Progesterone receptor.

PSA Prostate Specific Antigen.

RCT Randomized Controlled Trial.

RDD Regression Discontinuity Design.

SRSR Suomen Rintasyöpäyhdistys Ry.

TPS Tumor Positive Score.

WHO World Health Organization.

WSI Whole Slide Image.

1 Introduction

1.1 Background

The core of modern science is acquiring empirical evidence and reviewing the causality of some phenomena X to cause Y to happen. Randomized experiment, where subjects are divided to "treatment" and "control" groups, is considered to be the best method for a researcher to gather reliable evidence to estimate the average change in Y due to treatment; the average causal effect or more formally the ATE. [1]

The field of medicine relies heavily to the instance of the randomized experiment, Randomized Controlled Trial (RCT),in gathering the evidence for clinical questions [2]. By using randomization to achieve unconfoundedness and blinding the given treatment treatment – so that patients (single blinded) or patients and researchers (double blinded) are unaware of the given treatment –, the RCT aims to minimize systematic error [2]. The most rigorous RCTs will eventually be part of the revision(s) leading to formation of clinical guidelines [3], [4].

Cancer treatment and diagnostics guidelines – formatted as described above – have cutoffs and treatment largely depends on how patient falls into those categories. The guidelines and diagnostics are done by microscopic examination of the tissue composition, i.e. histopathology, which is cornerstone of cancer diagnostics [5]–[7]. Traditionally the sample review and diagnose has been done by a pathologist

using a light microscope, though the era of digitization has brought the use of slide scanners to send image to be reviewed on computer as Whole Slide Image (WSI) [8]. Moreover, the evolution of computers and especially the machine learning algorithms have made possible to develop computerized methods – such as convolutional neural network – for analysing histology WSI [8].

Though the RCT aim to eliminate the systematic error, the human factor makes plausible that the study results have some discrepancies. Histopathological analysis for example, is a task where concentration is needed for a long period of time and repetitive decision are made. This might lead to a phenomenon called "Decision Fatigue" where a quality of the subject's decision worsens due to long lasting session of decision making [9].

The use of computerized methods as an aid for decision making, CAD, might help to reduce decision fatigue and improve the diagnostic quality. For example, in a study where 24 pathologists used CAD in breast cancer diagnostics, pathologists accuracy and sensitivity improved, regardless of the expertise level of a pathologist [10]. Also, the time used to make a diagnose reduced with a median of 10 second, largest reduce being for those with a fellow status [10]. In some cases, the use of algorithms for WSI analyzes might offer a possibility to produce data that is in reproducible, higher in agreement, more sensitive and specific than humans alone [11].

By using CAD to analyse WSI the new RCTs might produce more precise data and given the reduced time consumption and increased cost efficiency, it might even become feasible to review the previous RCTs and compare the human alone and CAD results. With more precise results, it might become possibly to identify only the subjects in the borderline and review the treatment effect in those specific cases. However, if the aim of the revision is to effect to guidelines, the revision needs to be done with randomized experiment.

Using a RCT might not be an option for reviewing the existing guidelines, as the RCT is not only a relatively costly and heavy study setting, but also as there might be some unethical issues preventing the whole study. An alternative for RCT comes within quasi-experiments – a broad family of research designs – that makes possible to estimate the causal effect without random assignment. The estimate is valid internally as the quasi-experiments can control the confounding factors for observed and for unobserved, external validity comes as the subject's environment – natural context – isn't intervened. Thus, the quasi experiments share features from experiments and non-experiments. [12]

1.1.1 Choosing the topic

This thesis is done to Aiforia Technologies Oyj (hereon Aiforia), a software as a service company, developing convolutional neural network algorithms (AI-models) for analysing mainly histopathological WSIs with an aim to make the histopathological process more accurate, efficient and reproducible. Aiforia utilizes cloud services provided by the big tech vendors allowing Aiforia to provide the software as a cloud-based service (Aiforia Cloud), with flexibility and good cost-efficiency. [13], [14]

The process of using Aiforia Cloud for CAD includes first to scan the tissue slide to WSI, then use the CE-IVD marked Aiforia® Clinical Suite Viewer to analyze the image with pre-defined algorithm identifying the staining based biomarker(s). The resulting report holds the quantities of the bio-marker(s) and morphological features such as contours for tissue and tumor, which the pathologists then reviews and makes possible corrections before giving a pathology report about the sample.[13], [14]

In addition to CE-IVD marked Clinical Suite Viewer, Aiforia holds five CE-IVD marked AI-models certified for EU-usage: prostate cancer (Gleason GG), breast cancer (PR, ER and Ki-67) and in lung cancer PD-L1 [13], [14].

In a discussion with Aiforia's Chief Technology Officer Tuomas Ropponen (MSc

Tech.), rose a question if there would be some case in the literature where quasi-experiment and more precisely Regression Discontinuity Design (RDD) would have been used for histopathological data. As an extension to literature review, a small-scale proof of concept study was considered to be needed to research the technical possibility of conducting a RDD at the end of previously described "Aiforia pipeline".

1.2 Aims of the thesis

In this thesis a relatively new – in medical and histopathological perspective – quasi-experimental approach for estimating causal effect – the Regression Discontinuity Design (RDD) – is reviewed. The RDD leverages effectively the cut-off-based nature of many studies and utilizes the randomized variation resulting from subjects' lack to fully control the assignment variable near the known cutoff [15]. The RDD is proven to resemble random experiments [15] which might make RDD a good candidate to be used also in medical research [12].

Aiforia can provide tools for the research community to get cost-effectively robust, quantifiable, and consistent data, and with more robust data it might be possible to efficiently and specifically review the discontinuity area – area determining most of the cutoffs. It is in Aiforia's interest to show the existence of RDD and to investigate could – in general – the RDD be used in medical research and moreover to investigate could RDD be technically done at the end of "Aiforia pipeline".

The research questions will be answered in two stages with following aims:

1. A literature review: familiarize reader to the concept of RDD, find out is RDD used in medical literature (especially in histopathology), find out is treatment/diagnosis cut-off based for Aiforia's CE-IVD marked cancer AI-models.

2. Technical implementation for RDD "pipeline" is done so that: WSI analysis are done in Aiforia Cloud from where data is extracted to local computer using

Aiforia API and the extracted data will be used when RDD is implemented using R-programming language.

In addition, this thesis aims to support the research around RDD by gathering and presenting the most recent knowledge of the topic.

1.3 Structure

The rest of the thesis is organized as follows. The chapter 2 outline the key concepts that are needed when discussing about experiments, the chapter determines what is meant with an experiment and when can experiment be randomized, while the second chapter (2.1) explains what is meant when it's said experiments to find out causal effect of X to Y. The last two chapters (2.2&2.3) will briefly explain what is the effect that experiments, try to discover how experiments are done (chapter 2.2) and introduce the concept of quasi-experiment as a subsection of experiments and provide a short introduction to RDD.

Regression discontinuity design section from chapter 3 will move on to familiarize the reader to the general theory behind RDD discussing the key assumptions needed for RDD to be valid and the importance of choosing bandwidths, as well as the downside of requiring large number of samples. The chapter 3.1.2 will cover the two approaches and their sub types, by which RDD can be done: the local polynomial and local randomization. The chapter 3.1.3 will cover the sub types of RDD in cases where the division to treatment-control groups isn't perfect: fuzzy and kink RDD. Last chapter (3.1.4) of the RDD theoretical section will cover how the validity of the RDD results should be proven.

The literature review: section 3.2 and its subsections (3.2.1&3.2.2) will present in general: is RDD used in medical research 3.2, is it used in histopathology 3.2.1

and are the cancer bio-markers used to make cutoffs and do the cutoffs guide treatment/diagnosis. The section 3.2 will end the theoretical part.

The materials & methods section 4 starts the technical part of the thesis by presenting how the technical implementation was done (what was done, what was used, why) as well as explaining the study material and the acquisition of it.

The result section starts from chapter 5 onward and consists of multiple parts. At first the overall results for the study material are presented in chapter 5.1 and then as a part of broader introduction to results the 5.2 presents the survival analysis – which was the key to determine if there was some difference in mortality. The final part of general results is the chapter 5.3 presenting the calculated Case Fatality Rate (CFR) that were needed as an outcome variable for RDD.

RDD results are presented separately for pathologist (chapter 5.4) and for algorithmic (5.5) data. Both chapters include the whole RDD pipeline presented in the theoretical section (3-3.1.4).

The final two chapters are about discussing the findings from literature review, assessing was the technical implementation succeeded, suggesting future research subjects (chapter 6) and finally making a brief summary about the findings/suggestions (chapter 7).

2 Random experiments

The aim of the theoretical section of the thesis is to familiarize reader to the main concepts of random experiments and treatment effect before covering the Regression Discontinuity Design that implements heavily on the aforementioned. The chapter is started by determining what is meant with an experiment and what is a randomized experiment.

Shadish et. al. [16] define an *experiment* to be a study where effects are observed by intervention and experiment is *randomized* when some random process defines whether subject receives a treatment or an alternative condition. The random experiment – originating from proposals of Sir Ronald Fisher – is considered as the De facto method in medicine for treatment outcome research largely by its benefits; if randomisation is applied properly, it yields two or more groups with study subjects probabilistically similar to each other and the similarity results to that any observed difference should be a result of the treatment [1].

Also, assuming the conditions to be met, the random experiment provides a way to evaluate the results importance in real life. The difference between groups can be translated to of

The resulting difference between the groups is regarded as the treatment effect (τ), of which effect size can be calculated. Also, when the τ is statistically relevant and is inside a defined confidence interval, the biological relevance might be evaluated [1].

2.1 Causality

As the aim of an experiment is to find the reason why some factor X (cause) made Y to happen (effect), i.e. the causal inference. This brings in the concept of causality which is an inseparable part of the experiments in various concepts of it (both in real world as in quasi-experiments). The theory of causality has its roots in philosophy and for a long period causality had a critical reception amongst statisticians. [1], [16]

An excellent example of the causalities philosophical core comes the book Shadish et. al [16] who cite the philosopher John Stuart Mill's explanations that causality only exists when the cause: preceded the effect, was related to the effect and no other alternative explanation for the effect can be found.

Traditionally the explanation for the causality has been as singular: Y would've not happened without X. Currently though more emphasize has been given to the multiplexed nature of the causality by combining the explanations for causality to probability theory, raising the concept of probabilistic causation. Judea Pearl illustrates the theory of probabilistic causation by dividing the causality into probability of necessity, probability of sufficiency and to probability of necessary and sufficient causation. [1]

Probability of necessity (PS) is the standard explanation for the causation, emphasising the singular nature and the necessity of the cause to produce the effect when alternative processes are absent; the effect (Y) would've not happened without the cause (X). Probability of necessity is a commonly used type of cause in both legal and epidemiological settings, though particularly in epidemiology use of necessity needs to be done with caution as the probability of necessity can't be directly observed from the frequency distributions. Limitations exists as there's no control of the possible third (exogenous) factor that affects to exposed and unex-

posed subjects results (*Confounding*) and there's a possibility for the cause to be affected by a precursor state (*Sensitivity to the generative process*). [1]

Given an observational study type with an epidemiological focus, the probability of necessity can be thought to assess how probable it would be the disease not to occur without the exposure, when exposure (x) did raise the disease (y). [1]

In this settings excess risk ratio, evaluating the excess rate of occurrence of a particular health effect associated with exposure to some agent, can be used to measure probability of necessity only when the limitations are covered. When the independent variables X aren't dependent on the variable Y, i.e., there's no confounding, the exogeneity assumption holds and when no individual in the population can be helped by the exposure to the risk factor, i.e., no prevention, the monotonicity holds. If these assumptions don't hold, one needs to notice the boundaries or make corrections to the data or use combined data (experimental and observational). It's also plausible to use non-monotonic models but using them is out of scope of this thesis. [1]

Probability of sufficiency (PN) in short is about "*general tendency of certain event types to produce other event types*" with the "*probability that a healthy unexposed individual would have contracted the disease [y] had he or she been exposed [y]*" [1]. In concordance Shadish et. al. [16] wrote that cause is something that promotes the events needed for the effect to take place, cause won't act by itself, rather than together with multiple factors. As the definition $PS \triangleq P(y_x|y', x')$ emphasizes the absence of x and y, it gives probability to how sufficient the cause was to produce the effect. The sufficiency is most often used in epidemiological settings where the aim is to e.g., make some assessments for about how sensitive the population is to some risk factor. [1].

Probability of Necessity and Sufficiency (PNS) is an extension to probability of necessity and sufficiency with an assumption of exposure actually causing a disease and measures the exposures probability of being a necessary and sufficient cause for a disease. The result is the likelihood of a subject to be affected by PN and PS. [1]

The PNS, PS and PN can be written out using Pearl's example of leukaemia deaths and radiation. The PNS would give probability for a randomly chosen child –if not exposed to radiation – would both die of leukaemia if exposed and survive. PN would give survival probability of an exposed and deceased (to leukaemia) child without exposure. Finally, the PS would give the probability of death to leukaemia under exposure for an unexposed surviving child. [1]

2.2 Treatment effect

As mentioned, experiments seek to find out causality between treatment and effect, and especially in random experiments the treatment is binary – subjects are either treated or not treated. As the effect of the treatment varies between subjects, it's feasible to measure the average difference of the pair of potential outcomes, averaged over the entire populations $\tau^P = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)$. The formula is more precisely the formula of *sample* of the treatment effect, i.e. ATE for sub population. The formula is explained in so that summation occurs over all $\mathbb{N}$ individuals in the population, where $Y_i(1)$ are treated and $Y_i(0)$ are untreated (realized outcomes). The sub population, can be (and often is) the treated and thus the ATE becomes ATE for treated (ATT): a summation over all of treated units $\sum_{i:W_1=1}$. [1], [17]

2.3 Quasi-experiment

The Quasi-Experiment: An experiment in which units are not assigned randomly [16]

Several methods to conduct randomized experiments exists and have been studied for years. Common to many methods is the missing researcher intervention and that the methods don't *per se* do additional randomization. These non-intervention methods are generally referred to *Quasi-Experimental Studies* and they utilize existing data as a study material. The assignment to intervention and control group is done already in the source data. The use of source data groups is considered to be perfect or near-perfect option ("as good as random") to make opposite groups. Division is considered to have enough precision to produce unbiased data for estimates of causal intervention effects. [12]

The Regression Discontinuity Design (RDD) is a method belonging to the broad family of quasi-experiments and a method that many researchers have found to be a powerful tool to perform Randomized Controlled Trial (RCT) mimicking experiments. The method was developed by Thistlethwaite & Campbell in 1960 to test causal hypotheses in circumstances where random assignment to experimental and control groups is not possible [18]. From the 60's to mid-90's, the RDD was a niche method with very little use, but since the use of RDD has boomed especially in economics and the method has proven to be a versatile tool to be used in various discontinuity based study settings [15], [19]. The RDD has a stronger relationship to RCT than any other quasi-experiment, Lee & Lemieux [15] state that the RDD should not be viewed as a method, rather than a *"description of a particular data generating process"*. The RDD is particularly intriguing as it shares with random assignment the ability to produce unbiased estimates of the Average Treatment Effect (ATE) [16].

3 Regression discontinuity design

3.1 Theory of RDD

The RDD setting requires some continuous assignment (running) variable X with a pre-determined cut-off point (c) that determines the eligibility to treatment (Y) – the outcome variable. The cutoff determined division formulates two groups acting as a control and a treatment groups, generating randomized variation which – according to Lee & Lemieux (2010) – is an effect of the subjects' inability to accurately control the assignment variable X near the pre-determined cutoff. More formally; when a dummy variable $D \in \{0, 1\}$ marks getting treatment, it becomes $D = 1$ if $X \leq c$ and $D = 0$ if $X \geq c$. This suggests that the discontinuous jump in Y at c is the causal effect resulting from X, thus assuming the relationship between Y and X to be otherwise linear, the treatment effect – τ – can be assessed with linear regression: $Y = \alpha + D\tau + X\beta + \epsilon$. [15]

The figure 3.1 illustrates how the τ can be viewed as an estimate of the causal effect of X to Y. Given a subject with X exactly c, the Y needs to be determined for both groups. Excluding Y, we assume that X has no impact on factors and therefore B' can be assumed for Y for subject having X precisely c and belonging to the treated. And aforementioned still holding true, A" would be for the subject belonging to the control group and therefore the causal effect would be B' - A". To truly estimate the τ at $X = c$ it's a must to use data not just in c but in area around

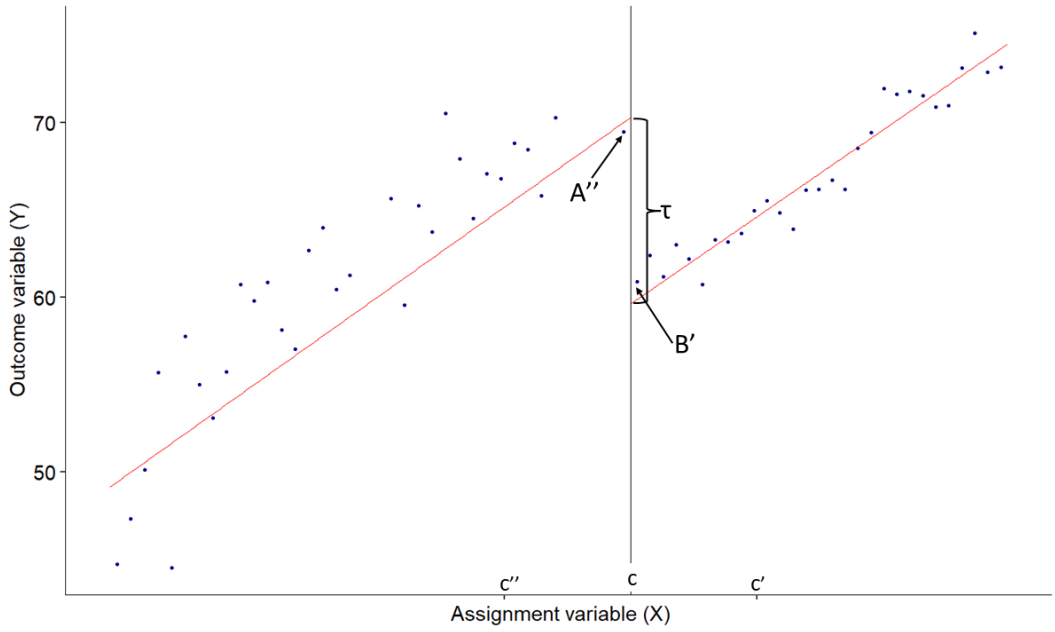

Figure 3.1: **Sharp RDD presented by linear model**. In an imaginary Regression Discontinuity Design (RDD) study the assignment variable X is blood glycohemoglobin level, outcome Y is the risk to diabetic retinopathy and cutoff (c) for insulin treatment is >6.4%. The sudden jump indicates that all exceeding 6.4% got treatment and all under c were left untreated. For the insulin treated (B') the treatment effect τ shows negative effect on retinopathy compared to untreated (A').

the c (points c", c'). However, it needs to be noted that care must be used in the choice of the area around c: if the area is narrow, the less data there are and wider area might lead to biased estimations. [15]

3.1.1 Key assumptions

RDD to be valid, key assumptions need to hold; the necessity of the continuity and random selection on the observables. The continuity assumption requires that all other factors are continuous with respect to X, so that the difference of $Y_i(1) - Y_i(0)$ can be used as a causal effect of the treatment. When subject can't *completely*

control the assignment variable X, it results the conditional expectation of the outcome variable to be continuous in the running variable. The inability also leads to a stochastic error and variation in the treatment, which leads the area around the discontinuity threshold to be considered "as good as" randomized.[15], [20]

Formally, the random selection becomes fulfilled in RDD as in treated $X \geq c$ and thus $D = 1$ is always true, in the control it's expressed that $X < c$ and $D = 0$. The requirements in X leads the variation to disappear from D and hence – according to Lee & Lemieux (2010) – D can't be correlated with any other factor. When the conditions are met, resulting from the continuity between average outcomes – $E[Y_i(0)|X]$ and $E[Y_i(0)|X]$ – and X, the RDD has the potential to estimate the ATE of X at the c (equation for sharp regression discontinuity (τ_{SRD}) 3.1). [15], [20]

$$\tau_{SRD} = \mathbb{E}[Y_i(1) - Y_i(0)|X = c] = \mathbb{E}[Y_i(1)|X = c] - \mathbb{E}[Y_i(0)|X = c]. \tag{3.1}$$

Here, it's important to stress that the ATE holds – as well as the whole design – only locally for individuals around the cutoff point. This is something that Lee & Lemieux (2010) describe not to be a drawback, but to better understand as a *weighted* ATE across all individuals (instead of *local* ATE). Lee & Lemieux (2010) present the mathematical evidence that only the ratio $f(c|W = w, U = u)/f(c)$ separates the RDD's *weighted* ATE from actual ATE. With the ratio, the weights directly present the ex-ante probability distribution of subjects X being close to the c. The comparison between – or determining closeness – overall ATE and RDD's *weighted* ATE isn't possible to know, as each subject's weight is unobserved. [15]

Also, when choosing RDD as a study design, researcher is rarely interested about neither end of the line. If the researcher is interested about the treatment effect of some factor X to Y, it's unfeasible to look the effect from those who won't get the treatment under any circumstance or from those who will get treatment regardless of what [15]. Consider a RCT to investigate effectiveness of surgery to minor bone

fractures: the results won't be informative for those having only bruise, they won't be operated, and one certainly will operate those with bone completely cut, again the results aren't informative. Instead, one is interested of the minor fractures to investigate would a cast be as good as surgery, this is the area between a bruise and a cut.

RDD's need for larger sample sizes is also a noteworthy topic, when applying the study type. Goldberger revealed that with a nonclustered design, the RDD needs 2.75 times more cases than the RCT to be comparable in statistical precision (efficiency and power) [21]–[23]. The efficiency of the RCT relative to the RD design was calculated by noting the reciprocal of one minus the square of the correlation between treatment and baseline variables. Here the reciprocal is the ratio of variances of the regression coefficients for treatment and indicates the efficiency of the RCT relative to the RD design [21], [22].

In another words, the RDD is a "cutoff-based RCT" and data in it is less randomized compared to actual RCT and thus RDD needs more observations to generate randomization and to achieve statistical significance. The need for larger sample size – with different effect size – is illustrated by Cappelleri & al. (1994) in the comparative table of sample sizes between RCT and RDD [22]. When a regularly used statistical power of 0.8 is assumed, the RDD needs a minimum 1.85 times more data for a large effect (0.8) and a maximum 2.73 times more for a small (0.2) effect. For example, at the 0.025 significance level and a power of 0.8 a small effect is observable in RDD with 1078 cases compared to RCT's 395 [22]. The need for larger sample size shouldn't pose a problem in the time of various registries, as the data can be extracted with decent effort and cost.

Bandwith selection is also a key issue in RDD as already covered in previous chapters; too large will create bias and too small will fail to capture the variation

[15]. In the RDD bandwidth h determines the proportion of the observations used to estimate the local linear regression. Data is usually binned and the c and regression line fitted according to the bins. In the bandwidth selection, it isn't advisable to trust solely on heuristic estimation on proper size, but to instead use some test-based judging and leave the heuristics for falsification, where the recommended bin width can be eg. halved [15], [24]. McCrary (2008) states that to avoid the biased decisions, a researcher should rely on subjective choice done by some automatic method [24]. This condition is fulfilled eg. in the RD packages, which is a family of programs used to run RDD's in SATA and R [25].

3.1.2 Local Polynomial and Randomization approaches to RDD

The two types of approaches to RDD existing are the Local Polynomial and Local Randomization methods [15]. The two differ to some degree on their underlying assumptions, yet both have the same idea on the how the results should be interpreted.

The Local Polynomial method has been discussed in the preceding chapters when formulas and theory has been introduced. In general, the polynomial approach is further divided into **parametric** and **non-parametric** approach, where the parametric denotes a linear regression with an assumption of linearity in regression in the assignment variable X: $Y = \alpha + D\tau + X\beta + \epsilon$. From the assumption follows that if the functional form is not linear and linear regression is applied, specification error might get enlarged at the cutoff and thus generate bias in RD estimates of the treatment effect (τ). [15]

A possibility to relax estimates of the regression function comes by using the non-parametric approach, where two main approaches are to include higher order

polynomials and to use kernel regression [15]. When polynomial functions of X are included to the regression model, the line will try to curve around the points creating a smoothing effect to the model [15]. Fitting higher order polynomial is important as robustness check but as a drawback, using higher orders might lead to nonsensical results when the model uses data further from the cutoff to predict the value of Y [15], [26]. Thus, it's highly recommended to use use estimators based on local linear or quadratic polynomials. [26]

Kernel functions act more as a local method and are thus suited for estimating the local effect at a particular point, such as the cutoff [15]. However, as the kernels gives more weights to data close to specific data point instead of looking more broadly, kernel approach might lead to biased results as well [15]. E.g., consider two bins A and B, the kernel calculates local means of Y (A' and B') for values of X in the bins on upward sloping line. If the A' and B' happen to be slightly lower (A') and upper (B') than the actual points (A and B) at the cutoff, the τ B'- A' is larger than the true τ, the difference of B - A. [15]

Local Randomization is based on an assumption that when subjects don't have a exact control to X, there exists a small window around the cutoff where the study type is "as good as" randomly assigned. More precisely, when the window is defined as $W_0 = [\bar{x} - w0, \bar{x} + w0]$, subjects scores above or below c in that window are predicted to be *as if random assignment.* As the regression functions are constant functions of the score in the possible window around $\bar{x}$, in the $\bar{x}$ itself regression functions will also be continuous functions of the score. The preceding can't be true in reverse, however. These conditions lead the assumptions in Local Randomization to be stronger than in Local Polynomial approach. [15], [27]

The local randomization doesn't require that much extrapolation use of and thus the effect of smoothing is reduced. Also, randomization has independence of the baseline covariate, thus consistent estimate of τ doesn't require including the

covariates. These benefits are specifically useful when identifying the treatment effect in cases where the observations around the c come in scarce or when the running variable is discrete. Overall, the local randomization approach is useful and advisable for testing the validity of any RD Design and could be used as solely in a design where the running variable is categorical. [15], [27]

3.1.3 Fuzzy and Kink RD

Fuzzy RD

differs from the previously introduced deterministic *sharp* RDD by the compliance. In sharp RD the probability of receiving the treatment changes sharply from 0 to 1 at the c and the compliance is perfect [15]. The *fuzzy* RD is a **probabilistic** design where the compliance is partial and crossing the c determines the treatment partially [15]. Fuzzy settings are usual e.g., in healthcare, where the treatment might not be solely dependent on crossing the c in X, but still the subjects aren't in total control of the X.

Fuzzy RD is visualized in the figure 3.2, where the two slopes won't change drastically from 0 to 1 but instead are more leaned to each other and connected at the c. When monotonicity and excludability are met, and it can't be that one subject crossing the c would make some to a complier and other to non-complier and also that the running variable X can't impact to the outcome Y otherwise than impacting to treatment type. Based on the assumptions, the formula for Fuzzy RD treatment effect (τ_F) is [15]:

$$\tau_F = E[Y(1) - Y(0) \mid unit\, is\, complier,\ X = c]$$

The *compliers* are those being treated only when cutoff rule ($X_i \leq c$) is met. The fuzzy design otherwise relies to exact same assumption – imprecise control over running variable X –, requires same checks as previously presented and can be thought as a randomized experiment where only the *intent to treat* is randomized

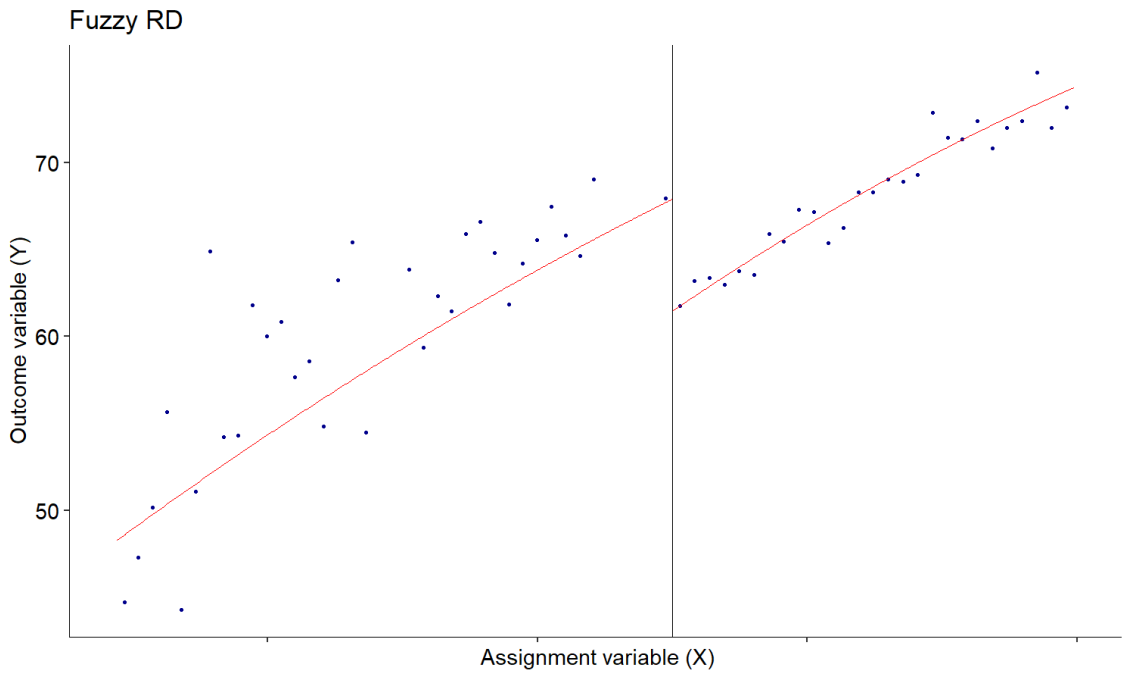

Figure 3.2: **Fuzzy RD, 2nd order polynomial**. Following the example of imaginary Regression Discontinuity Design (RDD) study presented in sharp regression discontinuity, now some eligible to insulin won't start the treatment and some right below the cutoff will start insulin by doctors' prescription. The setting where the jump isn't obvious is called *fuzzy* regression discontinuity.

and the compliance is imperfect. [15]

Kink

The Kink RDD can be either sharp or fuzzy but the common to them is that instead discontinuity, a change – a kink – at the cutoff in the relationship between the running variable and the treatment variable, allows implementation of the Regression Discontinuity Design (RDD). The kink design is illustrated in the 3.3, where in the black regression line (observed) a small kink is seen at the cutoff (indicated by the red line), the green dashed line indicates the expected results. Relying to

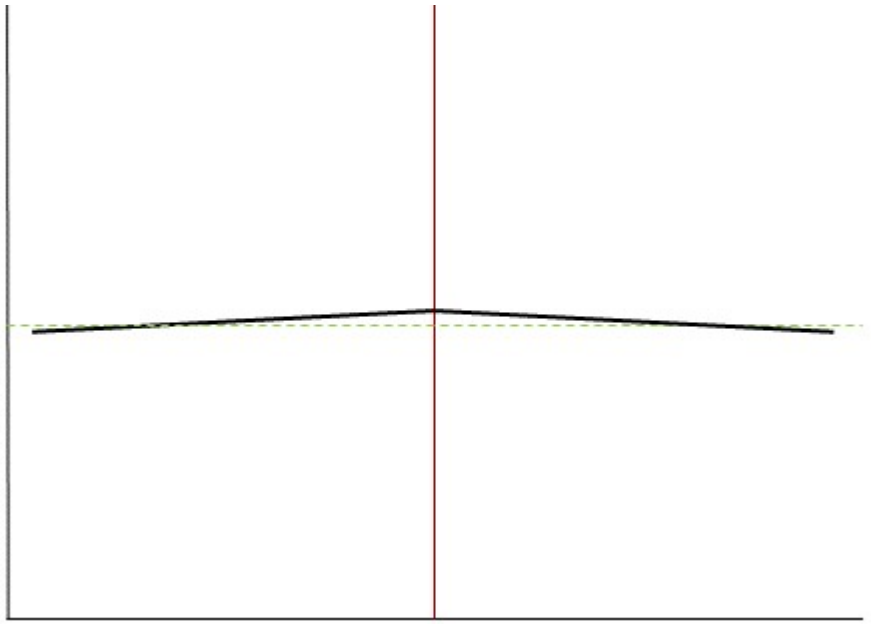

Figure 3.3: **Kink RD, illustration**. In the illustration, the faint green line represents baseline that would be expected without any treatment. Black line is the current situation displaying some bending around the cutoff marked by red line. Regression Discontinuity Design (RDD) can be done for the kink area as the treatment alters the observed status.

the similarity (excluding the treatment) between two groups, kink RD uses local polynomial regressions in estimating ATE at the c. [28]

The type of the kink design is determined with similar logic as previously described. In **sharp kink** has, the treatment variable is a deterministic function of the the assignment variable at the cutoff and it's required that all but the first derivative progresses smoothly across the threshold. In **fuzzy kink** RD, the relationship of the running variable and the treatment variable doesn't completely follow the rule *everything progresses, but the first derivative, progresses smoothly across the threshold*, a deviation that might arise from noncompliance. [28]

3.1.4 Checking the validity of RDD

When it comes to presenting proofs behind study assumptions, the RDD is no different from other statistical methods: the validity and falsification checks are a crucial step when conducting the study design. In RDD presenting the results (validation, falsification and analysis itself) in graphical format is an important part of the study and should be considered as a golden standard; RD graph should show data as discrete binned local averages and the regression line from a low-order polynomial [15].

Testing the manipulation is a vital check for the key assumption in RDD of the subject being unable to precisely manipulate the value of the running variable. To check the assumption, one can use histogram with bin widths as small as possible to visualize and to get a rough idea of how the groups are divided. When sorting isn't presence, the observations should be evenly distributed across the cutoff and no bunching should occur. However, the histogram shouldn't be solely used for manipulation, but instead it should be statistically tested by using a density test such as McCrary's (2008) or Cattaneo et. al (2020). The density test relies on the continuity-based null hypothesis implying that at the cutoff there is no difference between the densities of the running variable of treated and control observations. [15], [29]

Testing the similarity of the groups as regards to observable characteristics will provide proof for the subjects inability to manipulate the running variable. This testing can be done by comparing the variables determined *before* the assignment to treatment – predetermined covariates – and placebo outcomes, variables determined *after* the assignment that – relying on the knowledge of treatment's causal mechanism – aren't affected by the treatment. The predetermined covariates and placebo outcome should be tested against the running variable similarly as the actual out-

come variable. This will provide robustness check, because if the RD design is valid, the treatment isn't observable in placebo and predetermined covariates, thus the null hypothesis of no treatment effect remains. [15], [29]

Placebo cutoffs should be used to test the validity of the actual cutoff. As the expectation in RDD is that the cutoff determines the treatment, design shouldn't show any statistically significant treatment effect at the false cutoff. Cattaneo et. al (2020) propose a continuity-based analysis for alternative cutoffs, where the false cutoff is set ± few steps away from the actual c. [29]

Sensitivity of the results should be checked by using a range of bandwidths, and a range polynomial orders. However, as Cattaneo et. al. [29] state, the false bandwidth selection should be done with small ranges around the actual BW, for otherwise the results will have either too much variance (too narrow) or too much bias (too large) and result to unreliable falsification test. Eg. Cattaneo et. al. has used two types of automatic BW selection — coverage error probability & mean squared error — doubled the optimal values and plotted all variations; the different BWs should be in concordance with the optimal point estimate [29]. RDDs using different polynomial orders should be used together with the different BWs and to be presented as a table, which can as Lee & Lemieux present: help to "rule out overly restrictive specifications." [15]. Also using weighted kernels is advised for a robustness check of the results to differing bandwidths; the weighted kernels shouldn't make much difference to the treatment effect when there isn't sensitivity to the choice of BWs [30].

Sensitivity to observations near the cutoff is checked to observe the possibly manipulated observations. The idea of the "donut-hole approach" is that if subject could've manipulated the running variable, it would have happened near the cutoff

that determines the treatment. By excluding observations around the cutoff in a continuity-based manner with a pre-defined range from the cutoff – i.e. making holes – and repeating the RDD with the subsetted data, confirmation for the observations of the original data can be obtained. In summary, the point estimates and CIs should follow similar tendency regardless of the exclusion. [29]

3.2 RDD in medical literature

The cut-off feature by self makes the RDD interesting to use in medical data-analysis, where measured data is de-facto diagnosis method and cut-offs in the data form the sole base to treatment decisions. Additional fact that advocates the use of RDD to analyse medical data is how RDD uses subject's inability to precisely control the assignment variable near the cut-off as a sanity check and doesn't have "as good as randomized" assumption, rather than stating that the variation is a direct consequence of the inability [15], [31]. The inability of the subject to control the running variable can be statistically proven and especially in medical data the inability can be logically argued to be true [31].

The transparency and perfect knowledge of the selection process give the RDD a strong background compared to other quasi-experiments. Though the fact is that to succeed, RDD needs a data set 3-4 times larger than the one in RCT [15], the quantity of the data might not be a problem, as the digital archives for medical records provide a rich source of the data. This holds true also in cancer diagnostics where many countries have their own dedicated cancer registries that can be used for research purposes.

Despite the relatively easy access to data and the suitability of the RDD to medical data, increasing popularity of the RDD hasn't reached medical researchers for reasons that are somewhat speculative [31], [32]. Two systematic reviews — Moscoe et. al. ([33]), Hilton et.al ([32]) — from 6 years span show some growth in

the appliance of the RDD in medical research field and shed some light to the the method.

In 2015 Moscoe et. al. found with a crude key word search 193 studies [33], an in 2021 Hilton Boon et. al found 325 studies that met the review criteria and were from medical field. Most of the studies applied RDD to some socio-economic related issue, such as insurance policy, and the studies that applied RDD in clinical measures were a clear minority [32], [33]. The usability and rareness of the RDD in medical data lead to interest about the experiment's usability in medical data derived from cancer diagnostics.

Review based search following an additional keyword-based search – made by the author of the thesis – revealed that the use of RDD in cancer related topics was a rare sight, with under 10 studies to use RDD in some cancer related diagnostics and none of the studies had used algorithmic data in them.

3.2.1 RDD in histopathology

The literature review was enhanced so that it was to be discovered if RDD was used in some/any medical studies related to histopathology and also to review the biomarkers of the cancers that are in Aiforia's scope (see 3.2.2), if the biomarkers would have scale to be used in treatment-based division for RDD.

It can be said that based on the reviews, the use of clinical measure as a running variable X is somewhat rare; Hilton Boon et. al found 31 studies using clinical outcome [32] and author of the thesis found three (3) studies related to clinical pathology.

Smith et. al used RDD to review effectiveness of human papilloma virus vaccine in prevention of cervical dysplasia and anogenital warts so that the running variable is birth date used to indicate if the subject received the vaccine or not and as an

outcome incident cases of detected cervical dysplasia and anogenital warts [34]. As the cervical dysplasia can only be confirmed by histopathological sample, this study met the criteria.

Gild et. al used RDD to compare the effectiveness of early salvage radiotherapy to late one after radical prostatectomy for prostate cancer so that the running variable is Prostate Specific Antigen (PSA) used to determine if the subject prostatectomy group was early or not and as an outcome metastasis-free survival [35]. As the histopathological Gleason score had been diagnosed from the study subjects, this study met the criteria.

Shoag et. al used RDD to find out the effectiveness of PSA triggered biopsy screening to prostate cancer–specific or overall mortality so that the running variable is PSA used to determine if a biopsy was taken or not and as an outcome prostate cancer–specific and overall mortality [36]. As the histopathological Gleason score had been diagnosed from the study subjects, this study met the criteria.

3.2.2 RDD for analyzing biomarkers in cancer diagnostics

In the histopathological cancer diagnostics, the IHC stains are highly important as the prevalence of the antigens that the stains color give insight about the cancer type and prognosis, e.g., in breast cancer diagnostic guidelines the IHC stains are directly advised to be used [37].

This said, the antigens that IHC helps to detect are biomarkers for the cancer and the algorithms used in CAD detect these biomarkers indirectly, as human would. In general agreement exists about the usability of the IHC stains and the biomarkers, but the clinical cut-offs have some variation due to population differences and laboratory variations [5]–[7]. The thesis has reviewed the cut-off and treatment recommendations from College of American Pathologists (CAP) & American Society

of Clinical Oncology (ASCO), European Society for Medical Oncology (ESMO), International Ki67 in Breast Cancer Working Group (IKWG) and Suomen Rintasyöpäyhdistys Ry (SRSR). The biomarkers are referred to only by their established abbreviations and reader is advised to view the full terminology from the glossary. Selected cancers/biomarkers are selected by the interest of Aiforia.

Breast cancer diagnostic uses biomarkers to divide cancer to four different subtype – luminal A, luminal B, triple negative/basal like, HER2-positive – the use of adjuvant endocrine and anti-HER2 therapies are based on the hormone sensitiveness (ER/PR), presence of growth factor (HER2) and proliferation (Ki-67) [6], [7], [37], [39]. The table 3.1 is used to illustrate the cutoffs for different sub-types, the numbers represent the percentage of positive cells, indicated by colorization by the biomarker, out of total cells; in general HER/PR positivity is observed when the proportion is ≥ 1, Ki-67 positivity is somewhere ≥ 5 to ≥ 14 and HER2 ≥ 10 [6], [7], [37]–[41].

Lung cancer diagnostic relies on differentiating small cell carcinoma from non-small cell carcinoma's – adenocarcinoma and squamous-cell carcinoma. The carcinoma is categorized from H&E and/or IHC stained histopathological sample using World Health Organization (WHO) 2015 guidelines *Classification of Lung Tumors* [42], [43]. The full features of lung cancer diagnostic are out of thesis scope, and only the biomarker PD-L1 and Tumor Positive Score (TPS) method are covered. The PD-L1 staining is in Aiforia's interests and the TPS is used for guiding treatment.

The PD-L1 expression is determined from tissue sample using some IHC stain specific for PD-L1 antigen – e.g. 22C3 pharmDx assay for pembrolizumab –, expression of the PD-L1 is related to better treatment response for anti-PD-1 drugs. The TPS score is a derivative of the PD-L1: division from the number of positively stained cells divided by the number of all viable tumour cells is multiplied with 100. If patient has a non-small cell carcinoma and the TPS is ≥ 50, patient might be

eligible for treatment anti-PD-1 drug pembrolizumab. [43], [44]

Prostate cancer diagnostic is mainly based on the 2014 International Society of Urological Pathology (ISUP) guidelines which have been accepted as global guidelines by the WHO in 2016 [45]. The diagnosis relies on morphological findings from H&E-stained tissue samples and uses the Grading Group (GG) which have been derived from Gleason score (GS). Traditional GS uses a "ground grade pattern" 1-5 which generates the histological range – 2-10 – by summing the score of primary pattern to secondary pattern. GG combines the scores so that GS$\leq$ 6 = GG1, GS 3+4 = GG2, GS 4+3 = GG2, GS8 = GG4, GS>8 = GG5, respectively. [45]–[47]

Based on GS/GG and PSA patients are divided to three groups with low (GS: $\leq$ 6/GG: 1), moderate (GS: 3+4 /GG: 2) and high risk (GS: 4+3 or higher /GG: $\geq$ 3) cancer. The treatment for the local prostate cancer varies such that with low-risk cancer is only followed, moderate risk treatment is radical prostatectomy and radiation, high risk treatment is radical prostatectomy and radiation with a possible combination to other treatment forms. [47], [48]

Table 3.1: **Cutoffs for the biomarkers used for breast cancer diagnostics.**

	ER%	PR%	HER2%	Ki-67%
Luminal A	Positive: $\leq$1 CAP: low pos 1-10	Positive: $\geq$1 or high >20 IKWG/SRSR	Negative <10	Low ESMO/SRSR [1]<10 IKWG very low $\leq$5
Luminal B	Positive: $\geq$1 CAP: low pos 1-10	Negative (<1) or low ($\leq$ 20)	Negative <10	High SRSR >20 IKWG >30
Luminal B (HER2 pos)	Negative <1	Any value	Positive >10	Any value
HER2 pos	Negative <1	Negative <1	Positive >10	
Triple neg	Negative <1	Negative <1	Negative <10	prognostic marker[2] (high $\geq$ 10 / low <10)

In the table, the top row presents abbreviations of the biomarkers that are used for diagnosing and classifying the type (first column) of breast cancer. [1] Ki-67% and Luminal A type, see SRSR 2019 diagnostic guides [38], [2] prognostic marker used only by Keam et.al [39].

4 Materials & Methods

4.1 Ethical consent

In 10.2.2021 the Helsinki University Hospital (HUS) Ethics Committee II has given a positive ethical statement (HUS/415/2021 §24) for Aiforia Technologies Oyj for a study *"Image analysis based on artificial intelligence as a clinical tool for cancer diagnostics. Aiforia- AI for Image analysis"*. The ethical permission (in Finnish) is attached to appendix E. Aiforia has the rights to use WSI for the purposes mentioned in the ethical permission (see E). In addition, with the images, Aiforia uses patient diagnostic data as mentioned in the ethical permission.

The thesis aim is to study the usability of RDD to algorithm based and biologically relevant data, more precisely to cancer diagnostics. Aim of the thesis is in concordance with the Aiforia's larger project aiming to utilize the AI based image analysis data in clinical analysis. Thus, the thesis is covered with the Aiforia's ethical permission and doesn't require a separate ethical permission. (Aiforia Director of Clinical Products J. Juhila, personal communication, 19.8.2022)

4.2 Data acquisition

The thesis uses digitized breast tissue slides (WSIs) from the Helsinki Biobank pathology archives; images had been uploaded to Aiforia's cloud-based platform. The WSIs are from patients who have given a consent to keep samples in the biobank

and have been treated in the Helsinki Biobank area. The samples are completely pseudonymized upon arriving to the biobank and neither Aiforia nor the author of the thesis can personalize the WSI cases used in the study. Furthermore, the once pseudonymized WSI cases were pseudonymized again for the purposes of this thesis.

The thesis used 129 digitalized slides (n = 129) that had been collected from women with a breast cancer diagnosis. The samples had been IHC stained revealing the Ki-67 protein; pathologist's calculated percentage of positively stained cells was used to categorize the cases as high ($> 14\%$) or moderate ($\leq 14\%$) Ki-67 class. Thesis used the original pathology reviewed Ki-67 data and additional clinical information: age, five-year survival, age at death.

The HUS Ki-67 is calculated as an average from three areas that a pathologist has determined as most representative areas for the whole tissue (J. Juhila, personal communication, 19.8.2022). The "hot spot analysis" is done so that pre-defined number of areas are analyzed so that from each hot spot the total number of cells and total number of positive cells are calculated and the nuclear positivity is calculated as a relative percentage. The Ki-67 positivity is then calculated as an average of the calculated areas.

The WSI analysis were done in Aiforia CloudAiDev, which is one of the Aiforia's cloud-based platforms for image analysis. The analysis were done by using a pre-trained algorithm (AI-model) *OptimV5_EPI75VC_Cell11_25000* that identifies the tissue area, then proceeds to classify tissue to epithelium and from epithelial area counts positive and negative cells. Thesis used the reported nuclear positivity values as a grouping variable. The algorithmic results are, by contrast, a result from the whole slide analysis, meaning that the positivity percentage is the relative percentage from all cells that the algorithm has found from the whole slide divided by the number of all positive cells from the whole area. Due to the discrepancy between algorithmic and human results, the comparison of the Ki-67 is challenging.

4.3 Data analysis

The data analysis for the thesis were done using the R programming language (versions 4.2.1 upwards) run in RStudio development environment (version 2022.07.2). RStudio enables efficiently editing/running the code in a graphical user interface which also efficient utilisation of R packages and using proofreader for code. The attachment D lists the R packages used in the thesis. The whole project in a R format together with requirements file listing all used packages is found at: `https://gitlab.utu.fi/jojuket/thesiskettunen`, the git folder does not include the clinical data used creating the survival and RDD analysis as the data is protected with Non-disclosure agreement (NDA).

Accessing data with API

The data from the analysis was held in the CloudAiDev server and for analysis it needed to be retrieved into R. To access Aiforia data trough API developer the *aiforiaAPIutils* package, by Aiforia's developer Ville Koponen, was used. At first necessary credentials, subscription name and the unique identifier for the image analysis were filled after which data from the analysis was retrieved with a R script.

Data tidying

The image analysis produces a huge amount of data stored in a data frame of nested lists. The API call was written so that it leaves out all results from spatial metrics and also discards the tissue and epithelium area results. Limiting the API call was done by setting the get_coordinates false and using a filter based on values that cells gain from the colouring (positive or negative). The benefit of limiting the retrieved data was that the retrieved data would fit into memory at once and accessing the cell values was markedly simpler than it would been if other levels would've been used. The data downloaded from the Aiforia cloud server was saved in .rds format

for further processing.

The previously saved .rds file was opened in R as a gzip format to reduce the memory usage. As mentioned, the structure of the .rds object was a data frame consisting of lists, so utilizing R's list apply (*lapply*) together with *dplyr* package's piping operator made possible to loop function across the file. The script made a selection inside the nested list to get the individual cases (labelled by slide id) and categorized the cases based on cell value (positive or negative) and bind the column rows to a data frame. As the acquired data consisted only of absolute values, it was necessary to count the percentages for the grouping. The percentage calculation was done while looping through the values.

For the final analyzes, there was a need to get the HUS pathologist values, patient ages and time to death values. These values were in two excel files that needed to be downloaded locally and uploaded to R. In accordance with Aiforia's IT Manager Jan Valkonen and Director of Clinical Products Juuso Juhila, a decision was made that the author could use the files for data extraction, but local copies were agreed to be deleted right after extraction.

As the three files (data frame in R and two excels) needed to be merged, the re-pseudonymization wasn't done at the step one. The three files had a common slide id column which was used to combine the files into one data frame in R. The merging needed to be done in two pieces as the algorithm data didn't have a column for case ids. First the data frames holding survival times and the clinical data were merged together and the formatted new data frame was used as a template to which the data algorithm data was merged.

Data grouping

After the merge was done, a re-pseudonymization was done to the final data frame. The data was also tidied by replacing 'null' with 'NA', and re-coding the Ki-67 and

death columns. Ki-67 groups were formed by HUS diagnostic value (Ki-67 14%) so that if value is above 0.14 (14%) group is positive (1), else moderate (0). The death column was coded based on the value in time to live after diagnosis column. If column had a value, patient was dead during follow up (1), else patient had lived during the follow up (0). To use survival time as a variable in further analysis, a new column was introduced where survival time was counted as months. If death didn't occur during the five year follow up, values was 60 (5∗12), else the value was 12 times the time lived after diagnosis.

4.3.1 Survival calculations

Despite having a small dataset, there were some number of deaths (n = 8) during the five year follow up, thus enabling the calculation of survival time and deriving Case Fatality Rate (CFR) from the data. Prior to analyzes, the data was grouped as mentioned in previous chapters. As there were only 8 deaths during the follow up, the data was heavily right censored, meaning that most of the patients didn't experience the event (didn't die) until completion of the follow up (value 0) and true unobserved event is to the right of the follow up (censoring time) [49]. The survival function takes the censoring into account when calculating the survival rates and thus censoring won't be an issue [49].

The Kaplan-Meier estimator for survival times was calculated and visualized using R packages *survival, ggsurvfit* and *ggplot2*. The packages were chosen as they were earlier used in Aiforia's projects and for being somewhat a standard choice for survival time calculation. The Kaplan-Meier estimator is a semiparametric method to analyze unadjusted probability of surviving beyond a certain time point [49]. The time taken for development is a continuous outcome and the point of interest - death in the case of thesis - is a binary outcome [49].

Once data was grouped and tidied accordingly, *survival* package's *Surv()* function

was used to create a R data structure, nominally survival object. The next step was to create survival curves with *survfit* function for the entire dataset and to visualized with *ggsurvfit* package's function *survfit2*. The curves and survival times were calculated for the whole dataset (n = 129) and separately for positive (n = 92) and moderate (n = 37) Ki-67 groups.

4.3.2 Case Fatality Rate

As described in the section 3, the RDD needs some running variable (X) that has impact on the outcome data (Y). For purposes of the thesis, the Ki-67 was used as a running variable with a cutoff of 14%. For successfully calculating the RDD, the outcome variable needs multiple data points, thus utilizing the results from survival analyzes in Regression Discontinuity Design wasn't feasible as the survival analysis calculates rates in group wise manner, not so much for individual bases, resulting lesser data points. To get a better outcome variable, a decision was made to derive Case Fatality Rate from the time to live value.

The CFR expresses the percentage of cases of a specified condition which are fatal within a specified time. The CFR equation 4.1 determines the risk to be calculated as a number of deaths from a disease (in a given period) divided by the number of diagnosed cases in the same period of time [50].

$$Case\ Fatality\ Rate = \frac{Deaths\ from\ a\ disease\ (n)}{Diagnosed\ cases\ (n)} \times 100 \tag{4.1}$$

For the purposes of the thesis CFR calculation, the data was binned to groups based on the Ki-67%.. Two rounds on binning were done: binning by the value that the HUS clinician had used, producing 21 groups and binning the algorithm produced values by pre-decided cutoffs (n = 15). Once data was binned, the CFR was calculated with the equation 4.1 so that the numerator was the number of cases on the binned group and denominator was the ones exceeding cut-off 14% and being

assigned as Ki-67 positive (n = 92). As the Ki-67 grouping was done based on the values of the HUS clinician, denominator remained same when the calculations were done with the algorithm produced values.

As the CFR value was calculated in bins, each individual in the Ki-67 bin got the same CFR value; the HUS reference Ki-67 bins could be merged directly with the CFR values, but the heterogeneous algorithm results in the reference table were binned before merging the CFR values and the binned result groups were used as a guiding column for the merge.

4.3.3 RDD calculations

As mentioned in the chapter 1.2, the main focus of the thesis was to analyze Proliferation marker protein Ki-67 with Regression Discontinuity Design. The R supports a number of different RDD packages and the selection of which package to use was done by emphasizing active scientific work related to the package and the support provided to the package. The chosen RD Packages collection is actively maintained and is based on the research done by Sebastian Calonico & al. [25].

The RDD runs were done with applying best practices (see chapter 3) of the RDD analyses and it was evaluated: is the process of assigning treatment rule-based, is the design fuzzy or sharp, observe discontinuity in running variable around cutoff, observe discontinuity in outcome across running variable, determining the size of the cutoff and finally with falsification/validation methods give evidence for the reliability of the analyzes [32], [51], [52]. The evaluations were done by utilizing the R's built in linear models, RD-packages of Cattaneo & al. and by visualizing with *ggplot*.

For RDD analyses, the examples from Titiunik & Cattaneo in the 2021 Summer Institute of the National Bureau of Economic Research [52] and the lecture of Dr. Andrew Heiss' lecture in course PMAP 8521 [51] were used as reference when

compiling the design analyses.

Inspecting RDD data

First steps included visually inspecting the data to determine overall structure and to exclude the possibility of manipulation. R's *ggplot* package was used to graph the whole data and to inspect the type of the RDD – fuzzy vs. sharp. The graph's X-axis held value for Ki-67% and the Y-axis for CFR, from the graph it was visually inspected if all positives were above the cutoff and negatives below the cutoff, type was further confirmed with data tabulation.

Manipulation of clinical data or specifically cancer grade is basically impossible for the patient and for a doctor it would be not only illicit but also completely unethical. Still, the data can and should be inspected for manipulation even when dealing with clinical data. Inspection was done with *ggplot* by drawing a histogram of the two groups and inspecting the visual appearance. A confirmatory check was done with RD Packages *rddensity* that performs a local polynomial density estimation to check whether units are sorting [53].

The density test puts the running variable (X) into certain number of bins and calculates the frequencies that are used as a dependent variable in a local linear regression; the density test uses t-test with a null hypothesis of continuity around the cutoff [24], [53], [54]. The histogram of the density test is plotted, making easier to interpret the density curve and density changes around the cutoff.

In the thesis two packages – *rrdensity* and *rdplotdensity* – were used to first calculate test statistics and p-values at the a prespecified cutoff and to construct density plots from the test statistics. The t-test's result for statistical significance of the difference in the two points on either side of the cutoff is outputted under "Robust" section and in the plot, confidence intervals were drawn to visualize the statistical difference of the discontinuity of the difference. With overlapping confi-

dence intervals and $p > 0.05$, null hypothesis of continuity fails to reject indicating no evidence for manipulation or bunching [51].

Discontinuity in outcome across running variable

To observe the discontinuity in the outcome, at first the analyses were done parametrically by using linear regression. Both Cattaneo & Titiunik and Heiss recommend to use centered data when doing linear regressions to reduce risk of mix-ups and to make interpretation easier [51], [52]. Thus, a new variable 'R' was created to illustrate how many points above or below 14 the CFR was, this was done by by subtracting the cutoff from the running variable.

Next to visualize the differences – the discontinuity – between groups, a linear regression mode was fitted and plotted to the data by using *ggplot* such that the running variable was on the x-axis, the outcome variable on the y-axis, and points were colored by the group, respectively. The plotting was done separately to both centralized running variable (c = 0) and to "raw" running variable (c = 14).

To measure the actual effect of being assigned to either of the two Ki-67 groups, R's builtin *lm* was used to make linear regression models with a 1st order polynomial regression. The linear models' coefficients and respective p-values reveal the direction and the statistical significance of the difference. The use of 1st order polynomial is based on the suggestions in literature that using higher polynomial regression could lead to the change of overfitting the data [26], [52]. The first linear regression model *lm* was calculated to the entire data, hence the name model_simple. It's noteworthy that if the *entire data* would be used in final analyses, it would leave to false specifications, as in RDD the interest is in the area around the cutoff [51].

To achieve a correct values for the linear regression, the data need to be fitted by using different Bandwidth (BW)s. The BW selection was described in the chapter 3, but in short: in RDD there needs to be some estimation of the usable BW.

This can be done by estimating proper BW from the regression plot (area around the cutoff) or by using some empirical knowledge. In the thesis case, the dataset sets clear limitations to the used BWs; any smaller BWs wouldn't have worked as the size of the dataset is limited, especially in the observations around the cutoff. For these reasons, BWs was set to 10 and 20, and compared to each other.

The result from the ordinary least square analyse with arbitrary chosen BWs was compared to the results from the *rdrobust* package of the RD Packages series. The *rdrobust* package provides the means for "data-driven graphical and analytical statistical inference in RD designs" [25]. There are many advantages of using the tools provided by the RD packages; the tools provide easier and more robust way to build the local-polynomial point estimators and count confidence intervals for average treatment effect. Also, the tools enable a data-driven method to analyze the BW and provide a set of parameters to fine tune the RDD models and to plot the results for visual and explanatory analysis. [25]

When making RDD with *rdrobust*, the recommendations from literature were again followed and 1st order polynomial regression was used to avoid overfitting (local linear regression default value, p = 1). The kernel function used to construct the local-polynomial estimator (*kernel*) and the procedure used to compute the variance-covariance matrix estimator (*vce*) were adjusted to make comparison of the results, other settings were left as such.

The BW selection to RDD can be done separate to the analysis with *rdbwselect* but as the *rdrobust* uses the exact same function for bw selection and outputs the bw along with the effect size, running the *rdbwselect* separately wasn't seen necessarily.

The kernel function is out of the scope of the thesis and hence will be covered only briefly. The kernel choice affects to the kernel density estimation, estimation of probability density function ($f(x)$). The $f(x)$ best describes how the whole 100%

probability mass is distributed over the x-axis. Kernel estimate ($\hat{f}(x)$) is a non-parametric method for empirical representation of the probability density function ($f(x)$). In the kernel estimation, a kernel (K) is a standardized weighting function and the kernel determines the shape of the weighting function. The properties of $\hat{f}(x)$ are determined by the K and the BW h that effects to smoothness. [55], [56]

To summarize, the shape of K influences how much weight is given to observations around the cutoff and the figure 4.1 illustrates the shapes of the K's used in *rdrobust*. From the figure it can be observed that the blue uniform K is unweighted, giving all observations the same weight regardless of the distance from the cutoff ($c = 0$). Epanechnikov (red) – parabolic – gives weights following a curve and the weight diminishes the further the point is from the c. Triangular (green) determines the weight linearly and is used as a default K by the *rdrobust*.

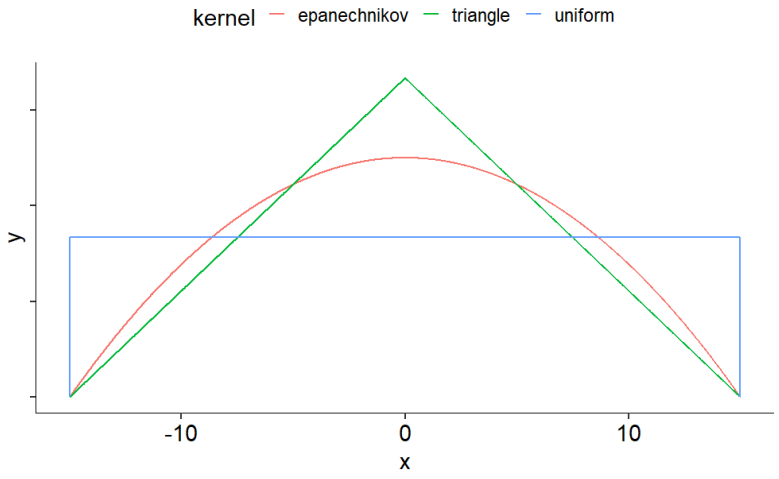

Figure 4.1: **Comparison of the kernels in a common coordinate system.** The image illustrates how the kernels use the x-y coordination to choose and weight observations. X0 indicates the cutoff point and used kernels differ in respect of inclusion of the more distant observations (eg. bandwidth of -10 to 10).

Variance-covariance matrix estimator *vce* is also out of the scope of the thesis and will be only briefly covered. Variance is defined as a measure of the spread around a center of random variable mean (variability), covariance measures how two random variables associate in their joint variability. In the variance-covariance matrix, the main diagonal contains the variances and rest of the matrix contain the co-variances. The matrix is used for determining the statistical significance of regression coefficients and for determining the confidence intervals for each coefficient. [57], [58]

When determining a variance-covariance matrix, is that the variance of the errors is constant and does not depend on x. If the condition is met, error is determined as a homoscedastic and if not, error is determined as a heteroskedastic (error is not constant). The presence of heteroskedasticity can make produce biased and inconsistent estimator of the covariance matrix of the parameter estimates, thus affecting to the statistical inference. [57], [59]

To overcome the effect of heteroskedasticity, *rdrobust* has plug-ins for heteroskedasticity consistent standard error (*hc*) and for nearest neighbor variance estimator (*nn*). The options for setting *hc* in *rdrobust* are without weights (*hc0)* and with weights from 1 to 3 [25]. The hc estimators have been compared to some extent; with small sample sizes hc0 might produce biased SE leading to overly liberal inferences in regression models. Out of the *hc* estimators 1-4, the hc3 might have the best small sample properties, though the rest too compare better than hc0.[59]

Local randomization RDD for an additional robustness check was carried out after the polynomial RDD. The analyzes were done using the *rdlocrand* – a subset of the RD Packages [25]. At first, optimal window was searched using *rdwinselect* to both centralized and non-centralized data. The analyses were adjusted such that starting from window sized of seven, window length was increased by one to achieve 10 windows' with 5000 repetitions (wmin = 7, wstep = 1, reps=5000).

5 Results

5.1 Overall characteristics

The data consisted of 129 of breast cancer cases, collected from women with diagnosed breast cancer. Cancer was diagnosed based on the histopathological samples reviewed by a HUS pathologist (HUS-pos%). Two cancer classes were present in the data, Carcinoma ductale and Carcinoma lobulare, these classes weren't used in characterising the data, respectively. A summary of the whole dataset is seen in table 5.1.

In the table, survival time is presented in months and is derived from the original data as explained in the chapter 4.3. The HUS-pos% value is from the original data and the AI-pos% and AI-pos% represent the values calculated by the algorithm from the WSI's. Positive AI percentage represents the percentage of the of positive cells out of the total cells, negative percentage illustrates the percentage of negative cells. The positivity values between algorithm and pathologist do differ to some extent, but the comparison is unfeasible as explained in the chapter 4.2.

As the thesis used grouping by HUS Ki-67%, a more feasible presentation of the characteristics is by dividing the data into groups related to the value. The table 5.2 summarizes this division.

From the table 5.2 it can be seen, that a positive Ki-67 was ∼40% common than a moderate diagnosis and the mean diagnostic Ki-67 value (HUS-pos%) was

Table 5.1: **Descriptive statistics of the study dataset, ungrouped.**

	n	Mean	SD	Median	Min	Max
Age (years)	129	60.55	11.27	60.54	32.78	86.11
Survival (months)	129	58.69	5.68	60.00	20.44	60.00
HUS-pos%	129	28.72	23.05	20.00	1.00	95.00
AI-pos%	129	23.03	17.66	18.20	0.70	74.30
AI-neg%	129	76.97	17.66	81.80	25.70	99.30

Table presents descriptive statistics for the whole dataset using variables age, survival time (i.e. time to death) in months, percentage of Ki-67 positive cells analyzed by HUS human pathologist and algorithm (AI), for algorithm % of Ki-67 negative cells also presented. In the table n = number of observations, mean = statistical mean, SD = standard deviation, median = statistical median, min = smallest observed value, max = largest observed value.

far greater in the positive group than the moderate group. The difference remains even when viewed the Ki-67 results produced by the algorithm, though it can be seen from the min and max values that some of the cases would've changed groups if the grouping would've been made by algorithm. The only independent variables were age and survival time. The boxplot of median ages (figure 5.1) shows the two groups (moderate, positive) boxes overlap and median lines lie within the overlap; a strong indication of similarity and statistically insignificant difference between ages.

In the table 5.3 the aforementioned values are displayed by groups arranged by the algorithm calculated Ki-67%. The values are just for comparison, grouping by algorithm value was not used in any of the analysis, as the actual life-expectancy was dependent on the HUS-calculated value. Incidentally the deceased persons were

in the same Ki-67 group in both categorization methods, but there's a true change that this wouldn't necessarily been the case. The table shows a slight shift towards moderate diagnosis, as the number of positives have diminished by five. The mean and median positivity % of Ki-67 are near to ones reviewed by pathologist, as seen from the table 5.2. Also the age when diagnosed is nearly identical between 5.3 and 5.2: 61 (±3y).

As seen from the boxplot in the figure 5.2 two groups (moderate and positive) had a remarkable difference, as expected, in the Ki-67% and the difference was similar regardless of grouping by a human pathologist (**B**) or an algorithm (**A**). This can be seen as a confirmation for the selected group division.

Table 5.2: **Descriptive statistics of the study dataset, grouped by Ki-67%**

Ki-67 Positive (n = 92)

	Mean	Median	SD	Min	Max
Age (years)	59.14	59.99	11.87	32.78	86.11
Survival (months)	58.17	60.00	6.66	20.44	60.00
HUS-pos%	37.61	30.00	21.55	15.00	95.00
AI-pos%	28.82	42.50	17.34	2.20	74.30
AI-neg%	71.18	46.50	17.34	25.70	97.80

Ki-67 Moderate (n = 37)

	Mean	Median	SD	Min	Max
Age	64.07	64.18	8.83	41.34	78.18
Survival (months)	60.00	60.00	0.00	60.00	60.00
HUS-pos%	6.62	9.00	3.23	1.00	10.00
AI-pos%	8.62	14.00	7.06	0.70	31.60
AI-neg%	91.38	19.00	7.06	68.40	99.30

The table present the descriptive statistics (for explanations, see table 5.1) for the study groups grouped by pathologist evaluated positivity percentage into positive (Ki-67 $\geq$ 14%) and moderate (Ki-67 $<$ 14%) groups.

Table 5.3: **Ki-67 division by algorithm results.**

Ki-67% positivity.

Ki-67 Positive (n = 82)

	Mean	Median	SD	Min	Max
Age (years)	58.73	59.34	11.78	32.78	86.11
Survival (months)	57.95	60.00	7.03	20.44	60.00
AI-pos%	32.16	26.45	15.98	14.70	74.30
AI-neg%	67.84	73.55	15.98	25.70	85.30

Ki-67 Moderate (n = 47)

	Mean	Median	SD	Min	Max
Age	63.74	63.64	9.63	41.34	82.98
Survival (months)	60.00	60.00	0.00	60.00	60.00
AI-pos%	7.10	7.20	3.27	0.70	13.90
AI-neg%	92.90	92.80	3.27	86.10	99.30

The tables below present the descriptive statistics (for explanations, see table 5.1) for the study groups grouped by algorithm calculated positivity percentage into positive (Ki-67 $\geq$ 14%) and moderate (Ki-67 $<$ 14%) groups.

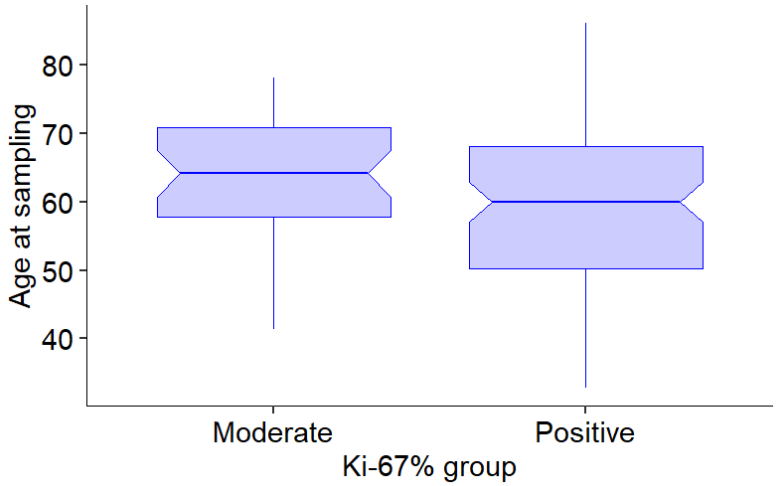

Figure 5.1: **Boxplot for mean ages of Ki-67 groups.** The boxplots describes the the distribution of age in both Ki-67 groups. Boxplots display minimum, first quartile, median, third quartile and maximum values.

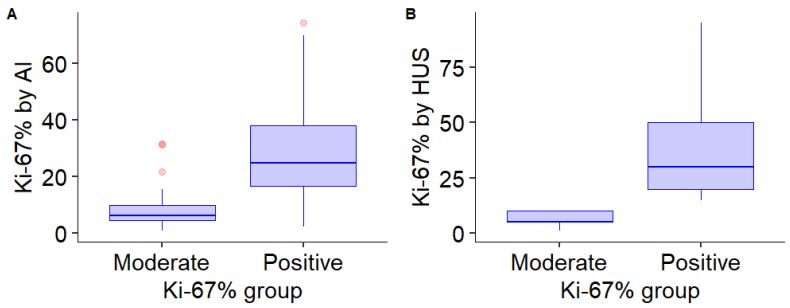

Figure 5.2: **Comparison of Ki-67 percentages inside the groups.** The boxplots describe the distribution of Ki-67 percentages inside the groups. In the image: algorithmic results in **A** and HUS human pathologist **B**.

5.2 Survival analysis

The survival analysis was first done to the whole data, to inspect the overall probability of survival. During the follow up period of five years (6 months) a total of eight people experienced the end event, i.e. were deceased. In the thesis dataset, the overall probability of being alive after five years with a diagnosed breast cancer was 94%. The table 5.4 demonstrates a slight decline in the survival probability during the followup.

Table 5.4: **Overall survival probability in the ungrouped data.**

	12 Month	24 Month	48 Month	60 Month
Probability	100% (100%, 100%)	99% (98%, 100%)	95% (92%, 99%)	94% (90%, 98%)

Table presents the overall survival probability for the ungrouped data; percentages indicate the probability of one being alive at the given time (the top row), brackets below are confidence intervals, receptively.

The Kaplan-Meier plotted survival curve (figure A.1) demonstrates the same decline, where in the x axis is the time and y holds the risk and each step in the curve is an event in the data. The grey Confidence Interval (CI) band – constructed by joining the CI at given time points – is the uncertainty estimate (similar to seen in table 5.4). The uncertainty increases already after first event and gets enlarged at the end, indicating strong uncertainty in the analysis.

To better inspect the survival, the analysis was done with a group division. As mentioned, the division was conducted by using the HUS-reference value, reflecting the situation in the original data. Also, changing the grouping to algorithm based wouldn't have changed the situation as the all deceased (n = 8) had Ki-67% above 14% regardless of the analysis method and thus were positive in both ways, as the

table 5.5 illustrates. Figure A.2 displays the Kaplan-Meier plotted survival curve for the groups. The moderate group is seen with a straight line (orange) as there weren't any events in the group. The positive group's survival curve (cyan) follows the same line as in the overall survival graph, only to have a slightly smaller survival probability (91%). Uncertainty is increasing towards the end of the follow up as visualized by increasing CI in the plot and confirmed by the table 5.6.

Table 5.5: **Ki-67% of the deceased.**

	YI1	YM6	W3O	N6Z	5CZ	F9SS	4T0	01GT
Ki-67 HUS%	80	30	50	50	40	70	40	45
Ki-67 AI%	39	71	69	55	64	31	61	65

Table presents the Ki-67% positivity of the deceased subjects by HUS human pathologist and AI algorithm, the top row indicates the re-blinded study subject id. Id's are blinded so that no one, except the author of the thesis, in Aiforia Technologies Oyj can disentangle the re-blinding and no-one in Aiforia Technologies Oyj, including the author of the thesis, can disentangle the original blinding.

Table 5.6: **Ki-67% grouped survival probability with confidence intervals**.

Ki-67 Group	12 Month	24 Month	48 Month	60 Month
Moderate	100% (100%, 100%)	100% (100%, 100%)	100% (100%, 100%)	100% (100%, 100%)
Positive	100% (100%, 100%)	99% (97%, 100%)	93% (89%, 99%)	91% (86%, 97%)

Table presents the subjects probability of being alive at the given time (the top row), data grouped by Ki-67% value based on HUS human pathologists diagnosis.

5.3 Case Fatality Rate

The CFR's were calculated for both categorization types using the formulae 4.1 described in the chapter 4.3.2.

The diagnosticians had used 21 unique classes in the HUS reference percentage, the bins were done by using the values as such without combining. The algorithm produced results had 116 unique values (observations n = 129) and were collapsed to 16 bins with a step of 5. The results from the CFR calculations are displayed in the Appendix B, where the table B.1 displays the used bins for both groups, table B.2 displays the calculated CFR's for results based on HUS pathologist diagnose and table B.3 displays the calculated CFR's for groups based on algorithmic WSI analysis.

5.4 RDD for pathologist analysed data

The RDD analysis for pathologist analysed data and for algorithm data (5.5) are divided to local polynomial and local randomization methods. Runs were conducted to binned data, for which the CFR's were calculated. The bins are displayed in the Appendix B.

5.4.1 Study type and checks for manipulation

The type of the RDD (sharp, fuzzy, kink) was determined by a plot (fig 5.3A) and a table to confirm the observations in the graph. In the graph (5.3A) the x-axis represents the Ki-67% analysed by a HUS pathologist. The y-axis (True/False) indicates, if the observation belonged to the positive (true) or moderate (false) Ki-67% group. As observed, there are no moderates in positive group and vice versa and the division before and after is sharp. A confirmation for the plot is seen in table 5.7: the table would display if some observation wouldn't belong to the group,

i.e. moderate group would have some observations with value TRUE and positive group would have some observations with value FALSE. Now all the observations are divided sharply, no moderates in positive group or vice versa.

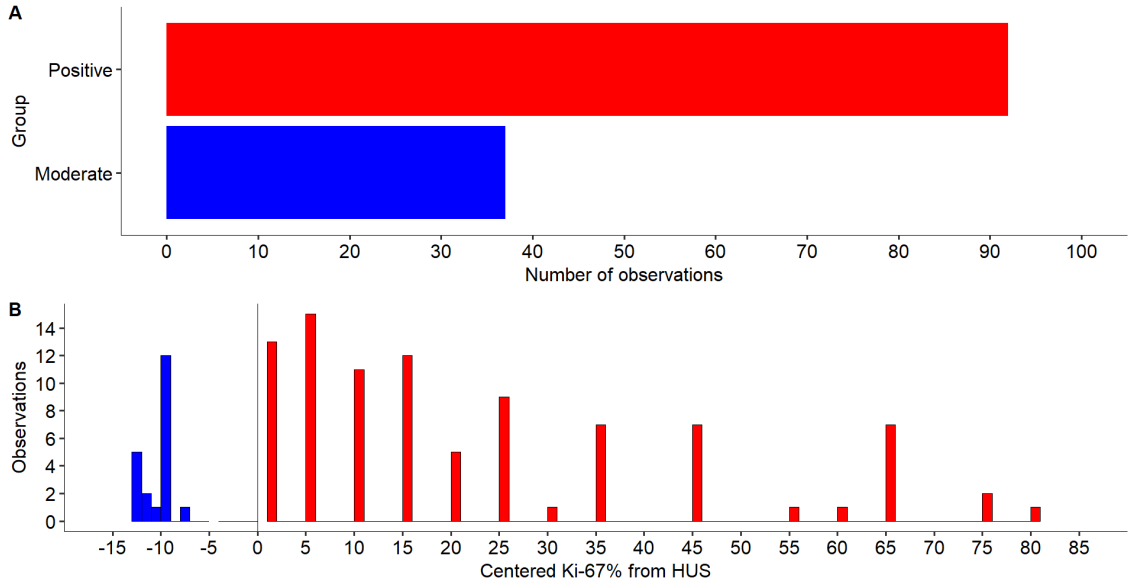

Figure 5.3: **Group overview for pathologist analysed data.** A) distribution of the observations in the groups and B) the number of observations in each centred Ki-67 bin.

To inspect the possibility of manipulation, first a histogram of the two groups (fig 5.3 was drawn to display possible punching around the cutoff. In the histogram, the x-axis represents the Ki-67% analysed by a HUS pathologist and y-axis represents the number of observations of given value in x-axis. The data used in the histogram is centralised to 0, i.e. cutoff (C) has been subtracted from the running variable X (X - C). The histogram (5.3) shows a slight difference in the values right before and right after the cutoff, otherwise the data is really heterogeneous due to using sheer raw values for the histogram. A confirmatory check was done with RD Packages *rddensity* that performs a density test to check whether units are sorting. The density results were subsequently plotted (fig 5.4) and tabulated (5.8).

Table 5.7: **RDD type in HUS data.**

Ki-67 group	Ki-67 ≥ 14	n
Moderate	FALSE	37
Positive	TRUE	92

Table confirms the sharp Regression Discontinuity Design (RDD) type in the HUS data. Each in the moderate group had Ki-67% under 14 (Ki-67 ≥ 14 = FALSE) and each in the positive group had Ki-67% over or equal to 14 (TRUE). I.e. none in moderate group with TRUE and none in positive group with FALSE.

The RD density (5.4) plot made with *rddensity* in the default options, displays the histograms of the Ki-67 groups with faint green at the bottom of the plot and the plotted average density (y-axis) estimator *f(x)* with CI above the histograms as a function of the running variable (x-axis). The color of the plotted average depends of two Ki-67 group, in left (black) is the moderate group and in right (magenta) the positives. The plotted CI's are overlapping though the CI of the moderate

The table 5.8 contains the summary from the *rddensity* command. The density test was done with estimated BW calculation using jackknife consistent estimator (VCE) and a triangular kernel. The estimated bandwidth used was 81 for both groups (BW est.), yielding the test to use all observations (Number of obs.) as

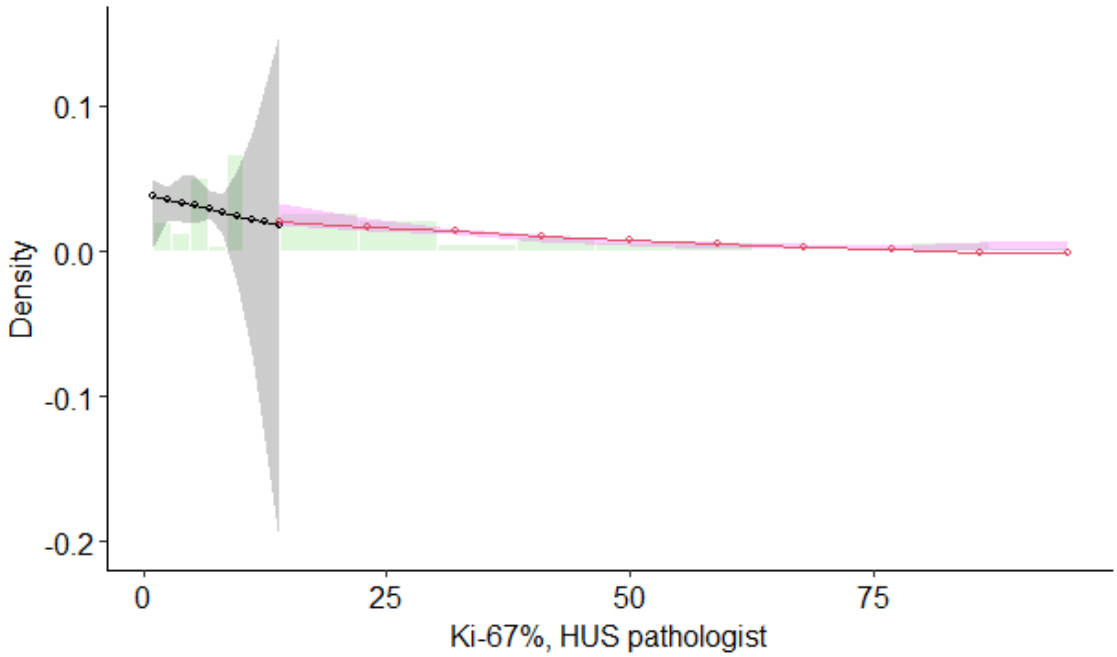

Figure 5.4: **Density plot for HUS grouped data**. The figure is the graphical representation of density test to check whether units are sorting. Left (black) the moderate group and right (magenta) the positives. Line represents the density (y-axis) and shaded area is the confidence interval.

effective sample sizes (Eff. Number of obs.). The rest of the section reports the manipulation test results, where: q indicates polynomial order under which test statistic was constructed (q = 3), p indicates polynomial order under which unrestricted model was constructed (p = 2) and *Robust* indicates final manipulation test results. The final manipulation test result (T) is 0.5912 with a p-value (P >|T|) of 0.5544. Overlapping CIs and p >0.05 indicate a fail to reject the null hypothesis, i.e., no statistical evidence of systematic manipulation of the running variable.

Table 5.8: **Manipulation testing using local polynomial density estimation (HUS data).**

Number of obs = 129
Model = unrestricted
Kernel = triangular
BW method = estimated
VCE method = jackknife

c = 14	Left of c	Right of c
Number of obs	37	92
Eff. Number of obs	37	92
Order est. (p)	2	2
Order bias (q)	3	3
BW est. (h)	81	81

Method	T	P >|T|
Robust	0.5912	0.5544

Table displays the summary of the density test, BW estimated using jackknife consistent estimator (VCE) and a triangular kernel. Final manipulation test result (T) with a *p*-value (P >|T|) of 0.5544 is >0.05 indicating a fail to reject the null hypothesis.

5.4.2 Parametric RDD for HUS data

First test for discontinuity in outcome across running variable and to measure the size of the effect, RDD was done parametric using local polynomial approach with a linear model in R (the *lm* command).

Three types of linear models using 1st order polynomial (p = 1) were calculated to centralized data, one with whole data, two others with BW of 15 and 20, respectively. The table in table (5.9) summarizes the results, from the table it can be instantly seen the the effect of having scarce observations around the cutoff – lm is doable only to a large BW. The row *is_positiveTRUE* indicates a change in CFR if having Ki-67 >14%. The *ki67_centred* is the average change in the response variable (CFR) for a one unit increase in x and the *intercept* mean value of the response variable when x = 0. Below the values, the standard error of the estimate is in brackets (e.g. (0.115)). The coefficient of determination (R2) is around 21% with whole data used and around 55% with BW of 20, a good example why the whole data isn't appropriate for RDD.

The type of the change is observable in the linear plot (fig 5.5). The 1st order polynomial plot is built on centralized data and shows the whole data model with a solid line and the BW of 20 in dashed line. The plot displays the regression line with CI for the moderate group (F) on left and the positives (T) on right. The cutoff (c = 0) splits the group sharply and displays a jump in the data.

Based on the observations in table (5.9) and in the plot (5.5) it can be said that when not restricting the BW there's around 26% increase in CFR is one belongs to the positive group. The negative change in the model restricted to 20 is unlikely and probably caused by misfit data. And it's important to notice that the using whole data is not appropriate either, when the interest is the change around cutoff.

Table 5.9: **Size of the treatment effect as measured by linear model with differing bandwidths (HUS).**

	Full data	Bandwidth = 20	Bandwidth = 15
(Intercept)	0.087	0.390	0.000
	(0.115)	(0.059)	(0.000)
ki67_centered	0.012	0.053	0.000
	(0.003)	(0.006)	(0.000)
is_positiveTRUE	0.262	-0.565	0.000
	(0.167)	(0.105)	(0.000)
R2	0.218	0.556	-
Num.Obs.	129	88	76

Table illustrates the treatment effect measured with using different bandwidths when the parametric RDD was done with 1st order linear regression. The table is done for centered values and in the table; intercept is the average case fatality risk (see 4.1) at the 14 percentage threshold, ki67_centered is the increase in risk for every point above 14 that subjects Ki-67 value is, is_positiveTRUE is shift in intercept (difference between scores at the threshold) when subject belongs to positive group (has Ki-67 ≥ 14) is true. R2 is a measure of the goodness of fit of a model and the Num.Obs. describe the number of observations the model used. Values below in brackets represent standard error, respectively.

5.4.3 Non-parametric RDD using *rdrobust* for HUS data

The second approach for RDD was to the *rdrobust* with different kernels and heteroskedasticity robust (heteroskedasticity consistent k class or HCk class) methods. Although the plots were done using 1st order polynomial, the approach is non-parametric due using kernel regressions.

The results from figure 5.6 display the RD plots done with standard settings but using three different kernels (A: uniform, B: Epanechnikov, C: triangular). The plots were done to centralized data (c = 0); x-axis holds the centralized running variable (Ki-67& by pathologist) and y-axis is the CFR, respectively.

The plots display a slight upward jump and a discontinuation around the cutoff, though the size of the jump and the scale of the y-axis varies depending of the selected K shape (see figure 4.1). The table C.1 in the Appendix B displays the size of the jump (coef), conventionally and robust calculated statistical significance with CI, and the BW used in calculations together with effective number of observations (n) separately on both sides of the cutoff. Three different kernels were analysed using HCk classes between 0 and 3, the nearest neighbour (nn) VCE method could only be used with uniform kernel. The effective number of observations (n) used in BW estimation varies with the type of selected kernel, as expected.

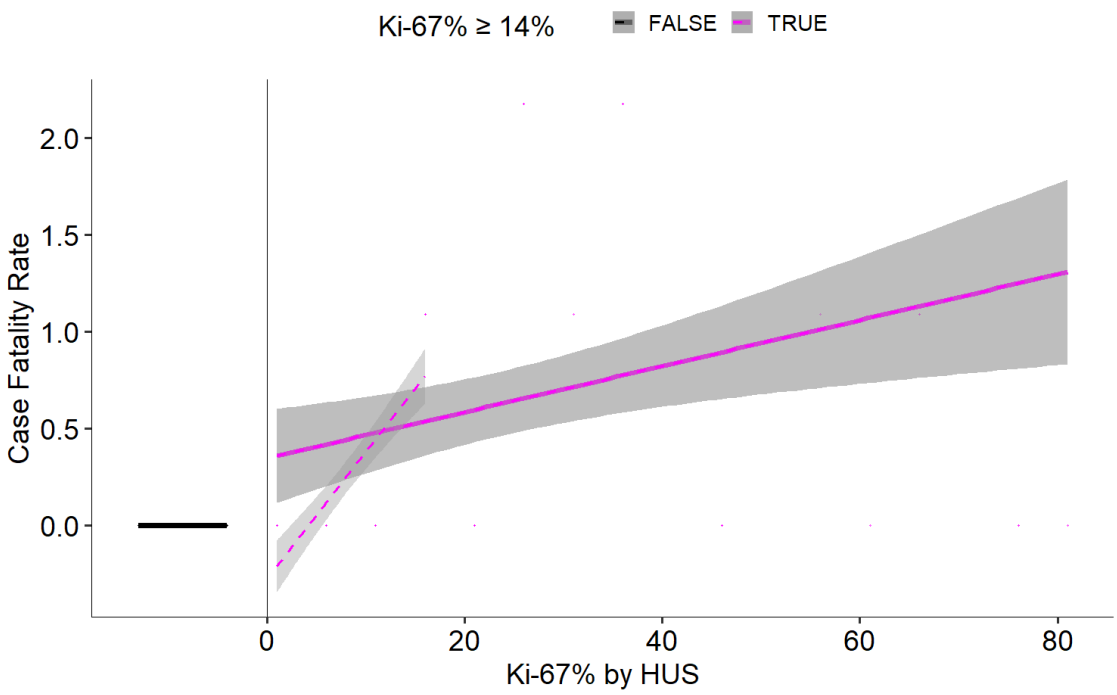

Figure 5.5: **Linear RD model for pathologist analysed data**. Figure shows the sharp Regression Discontinuity Design (RDD) between Ki-67 groups for whole data (solid line) and for bandwidth of 20 (dashed), the graph suggests an increase in Case Fatality Rate (see equation 4.1) when Ki-67% increases.

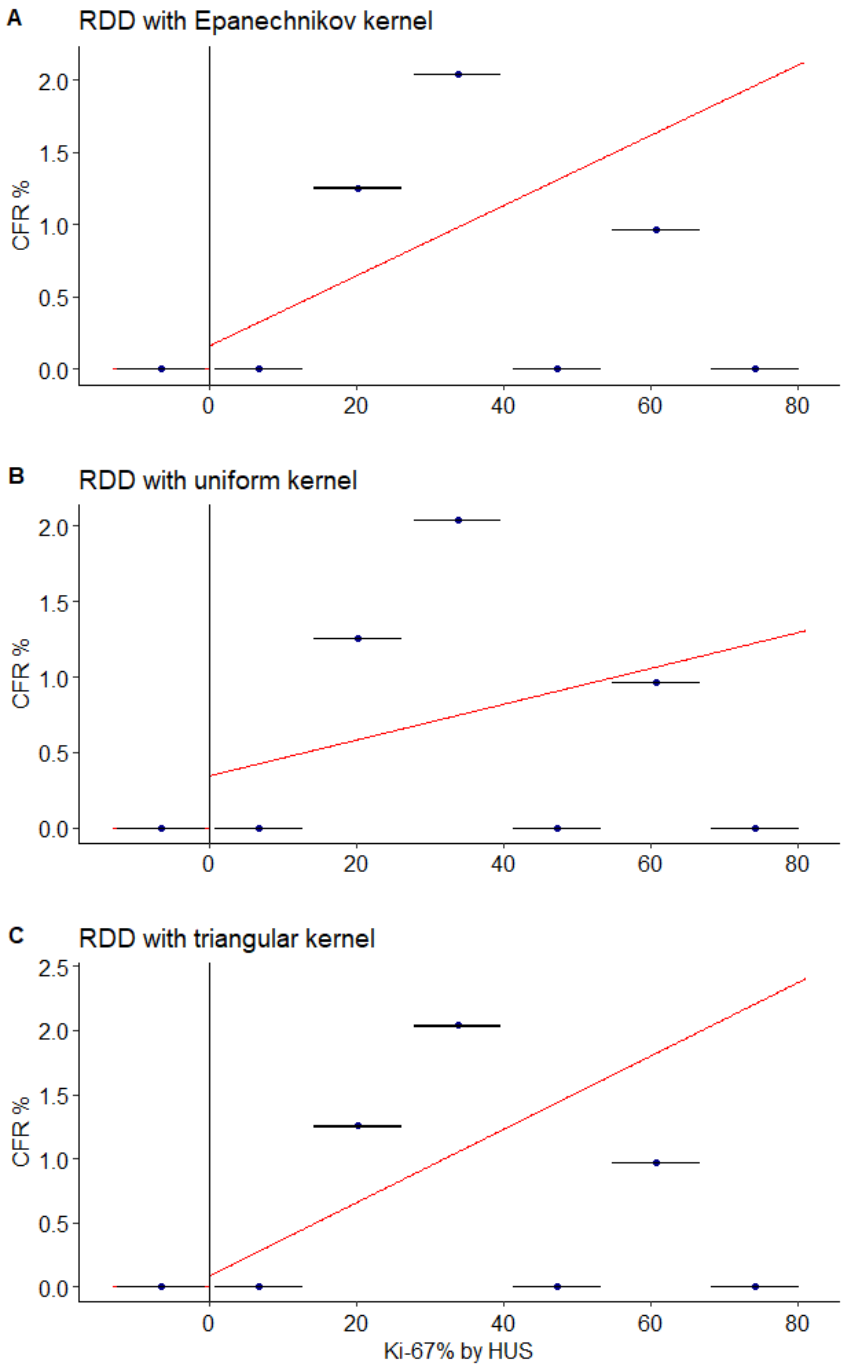

Figure 5.6: **Three non-parametric Regression Discontinuity Design (RDD) for pathologist reviewed Ki-67%.** Figure compiles the results from non-parametric RD analyses from the three different kernels used. In the table CFR = Case Fatality Rate and WSI = Whole Slide Image.

Negative numbers in the coeff column result are a calculative presentation, rather than the direction of the difference. As can be seen from the plots 5.6, all discontinuations lead to an increase in CFR. The BW is rather constant when using Epanechnikov (A) and triangular (C) with different variance estimators; with Epanechnikov BW is ∼16 and with triangular ∼15 in both sides of the cutoff. Out of the two, only the Epanechnikov kernel manages to capture the size of the jump giving it range between ∼0.03-0.05. The conventionally calculated statistical significance gives $p < 0.05$ to all coefs, however the significance is lost with robust statistics.

The table C.1 in the Appendix C shows uniform kernel having far greater variation in BW selection, smallest being 15.7 and largest 26.2, also the sample size varies at the right side of the cutoff. The uniform kernel shows the jump being tenfold larger than Epanechnikov's and gives statistical significance with both methods to all the coefs. The uniform kernel can make a model with nn variance estimator, however leading the coef to disappear.

5.4.4 Local Randomization RDD

The *rdwinselect* outputs the window recommendation and a plot for uniform kernel (see fig 5.7), from where the recommended window is the one just before p changes to be statistically significant. For non-centralized data (A) recommended window was [-15;15] with 76 observations (37 below, 39 above). For the centralized data (B) recommended window was [-5;19] with 50 observations (20 below, 30 above). The results from the *rdrandinf* analyses are listed in the table 5.10, where c is the cutoff (0 for centralized, 14 for non-), n the effective sample size between left and right of the cutoff, win the selected window size, k is the selected kernel, stat the statistic used in the balance tests, T the difference and p the statistical significance of the difference. The local randomization RDD confirms the findings from the local polynomial runs; with the given data only, measurable effects appear when to whole

scale of the data is used. When whole data was used (win length not limited) both uniform and Epanechnikov found ∼0.6% difference-in-means between the moderate group and the positive group. Difference was statistically significant and identical in centralized and non-centralized data. However, as the table 5.10 points out, when the window length was adjusted to that recommended by *rdwinselect* (more realistic), the difference disappeared and was not measurable by any statistic method used in the balance tests.

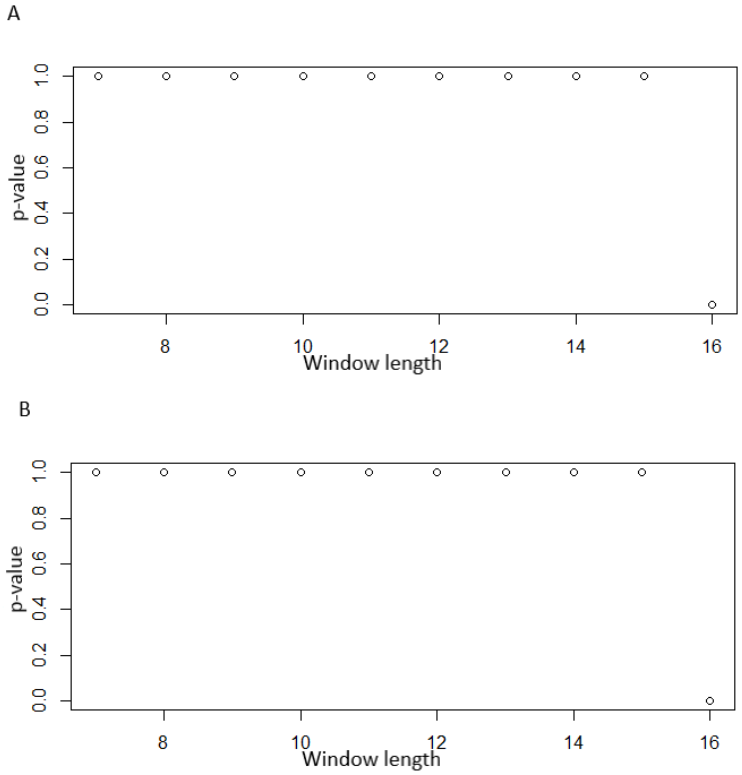

Figure 5.7: **Plots for recommended window size using *rdwinselect***. Results from selecting the most optimal window (one before $p < 0.05$) with finite-sample methods near the cutoff where the assumption of randomized treatment assignment is most plausible. A) windows for non-centralized data, B) for centralized data.

Table 5.10: **Results from local randomization RDD using HUS data.**

c	n	k	stat	win	T	P >|T|
14	37/92	uni	d.m	[1;95]	0.626	0.000
14	37/92	epan	d.m	[1;95]	0.596	0.000
14	37/39	uni	d.m	[-1;29]	0.000	1.000
0	37/92	uni	d.m	[-13;81]	0.626	0.000
0	37/39	uni	d.m	[-15;15]	0.000	1.000
0	37/39	uni	K-S	[-15;15]	0.000	1.000
0	37/39	uni	r.s.z	[-15;15]	0.000	1.000

The table presents a summary of the local randomization Regression Discontinuity Design (RDD) results when using HUS data. In the table c = cutoff, n = number of observations, stat = statistical method used, d.m = difference in means, K-S = Kolmogorov-Smirnov, r.s.z = Rank sum z-stat, win = windows size, T = treatment effect, P >|T| = statistical significance of the treatment effect.

5.4.5 Falsification test results

As discussed in the chapter 3.1.4, to prove/confirm the RDD results, it's crucial to make falsification tests/sanity checks for the data. Continuity and manipulation have been checked in chapter 5.4.1 and different bandwidths have been used in the chapters 5.4.2 & 5.4.4. Other falsification tests were done by plotting RDD for placebo outcomes and cutoffs. Placebo cutoffs were chosen to be double (28) and triple (42) compared to the original (14) and the placebo outcome was chosen to be age, as it can't be affected by any treatment. In realistic dataset the placebo cutoff and outcome shouldn't show any effect, but figure 5.8 illustrates, this not applying for the used data. At placebo cutoffs 28 (5.8A) and 42 (5.8B) there are jumps appearing at the cutoff, also the plot shows an interesting kink and a clear difference when age is used as an outcome variable (5.8C). The plots clearly illustrate the problems of using a small and unevenly distributed dataset.

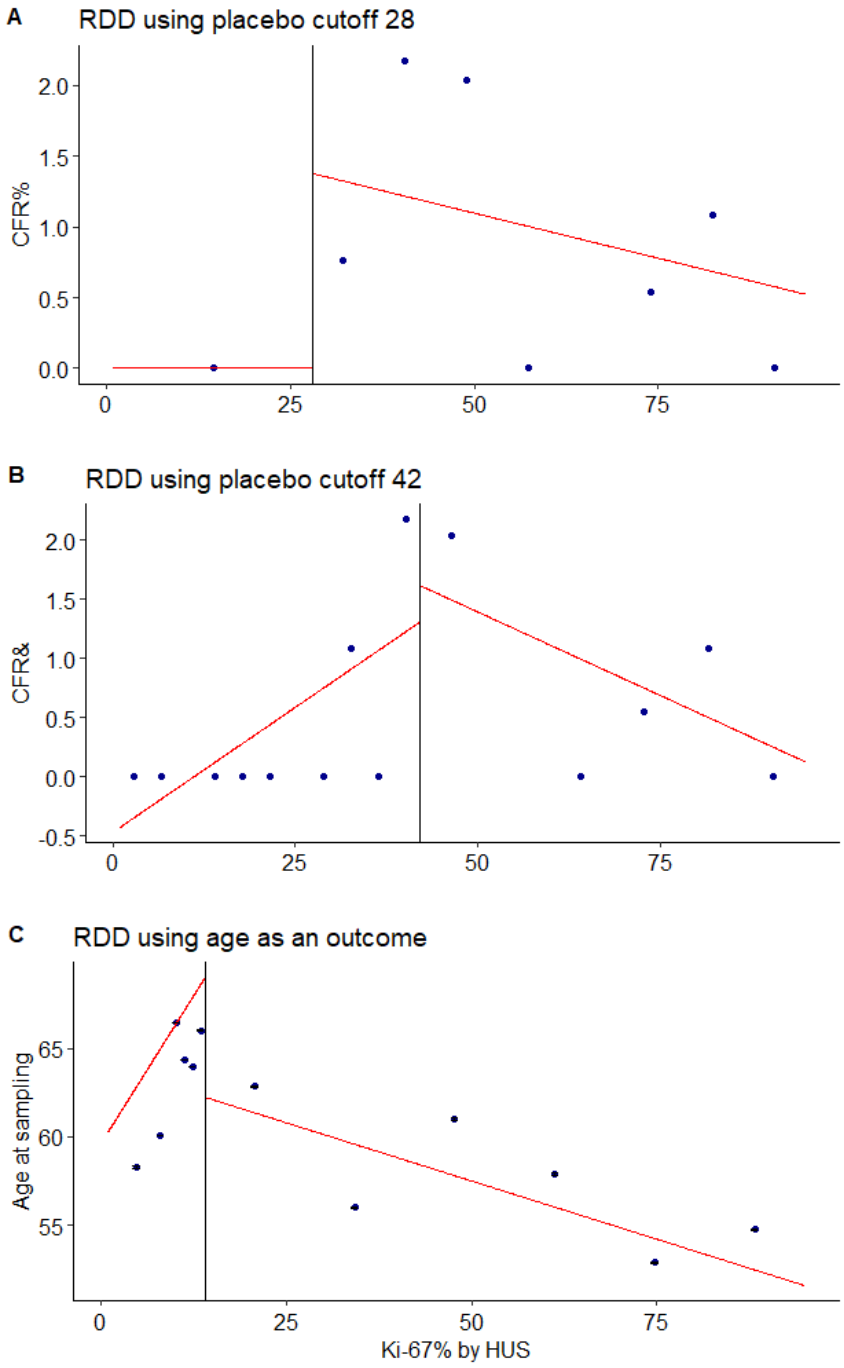

Figure 5.8: **Falsification RDD plots using linear regression and uniform kernel, HUS data.** Figure compiles the falsification Regression Discontinuity Design (RDD) results done for HUS data using placebo cutoff (A-B) and placebo outcome. For robust results, placebo cutoffs and outcomes should now show any significant differences. In the image CFR = Case Fatality Rate.

5.5 RDD for algorithm analysed samples

Similar to of human pathologist, RDD runs were made to WSI samples analysed by algorithm. Reader is advised to read more detailed description of the used methods from the previous chapter 5.4.

5.5.1 Study type and checks for manipulation

The algorithmic analysed WSI was also checked form manipulation similarly as the pathologist dataset previously. In the image 5.9, part A illustrates the sharps vs fuzzy setting and part B bunching in data. From the part A, it is seen that there are no moderates in positive group and vice versa and the division before and after cutoff is sharp. Part B demonstrates that there's some difference in the density of the observations around the cutoff, but the significance of the observations is to be confirmed by the density test (fig 5.10). The table 5.11 confirms the division of the observations: all moderates are in one group and all positives in another, thus the study type is sharp. The density plot in 5.10 shows CIs are overlapping and table 5.12 confirms that there's no statistical evidence of systematic manipulation of the running variable ($p < 0.05$) .

When the algorithm produced data (fig 5.10) is compared to pathologist analysed, it's evident that algorithm uses far more Ki-67% classes. This is probably due to human pathologist using rounding and rough class division, whereas algorithm uses the whole scale from 0 to 100. The improvement in the data quality can also be seen in the density plot and the table, where the curves and the CIs together with used BW are more detailed compared to those in the pathologist data.

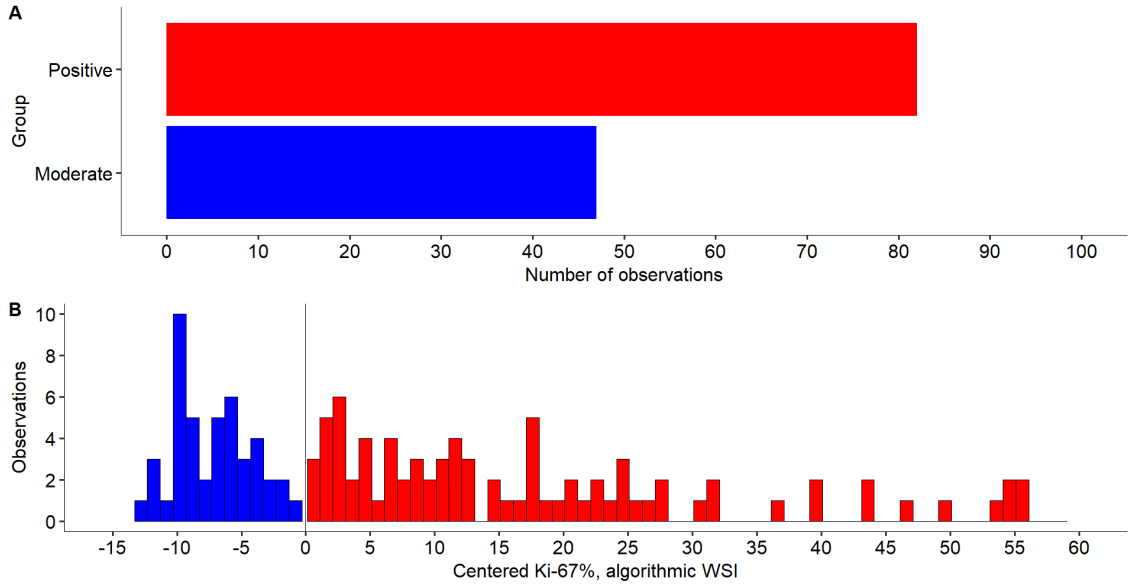

Figure 5.9: **Group overview for algorithmic analysed data.** A) distribution of the observations in the groups and B) the number of observations in each centred Ki-67 bin.

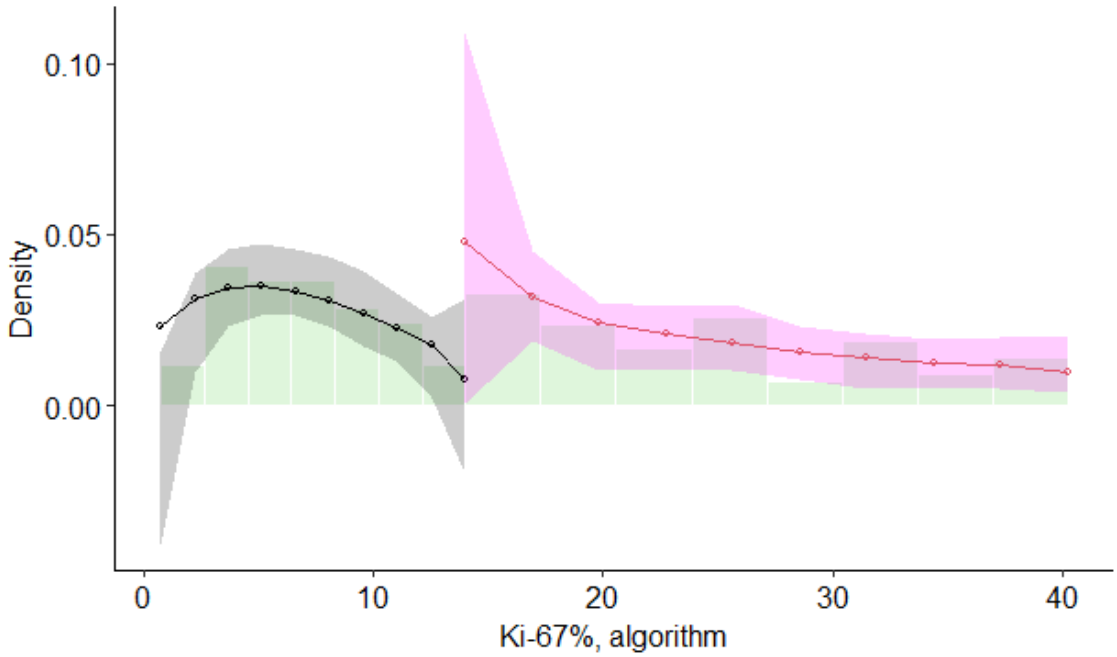

Figure 5.10: **Density plot for algorithmic grouped data.** The figure is the graphical representation of density test to check whether units are sorting. Left (black) the moderate group and right (magenta) the positives. Line represents the density (y-axis) and shaded area is the confidence interval.

Table 5.11: **RDD type in algorithmic data.**

Ki-67 group	Ki-67 $\geq$ 14	n
Moderate	FALSE	47
Positive	TRUE	82

Table confirms the sharp Regression Discontinuity Design (RDD) type in the algorithmic data. Each in the moderate group had Ki-67% under 14 (Ki-67 $\geq$ 14 = FALSE) and each in the positive group had Ki-67% over or equal to 14 (TRUE). I.e. none in moderate group with TRUE and none in positive group with FALSE.

Table 5.12: **Manipulation testing using local polynomial density estimation (algorithmic data)**

Number of obs = 129
Model = unrestricted
Kernel = triangular
BW method = estimated
VCE method = jackknife

c = 14	Left of c	Right of c
Number of obs	47	82
Eff. Number of obs	28	29
Order est. (p)	2	2
Order bias (q)	3	3
BW est. (h)	8.57	8.752

Method	T	P >|T|
Robust	1.6154	0.1062

Table presents the summary of the density test, BW estimated using jackknife consistent estimator (VCE) and a triangular kernel. Final manipulation test result (T) with a p-value (P >|T|) of 0.1062 is >0.05 indicating a fail to reject the null hypothesis.

5.5.2 Parametric RDD for algorithmic data

Three types of linear models (p = 1) were calculated to centralized data, one with whole data, two others with BW of 15 and 20, respectively. The table 5.13 summarizes the results. Compared to results of human pathologist (5.9), the algorithmic data makes linear model calculations possible also with smaller BWs. The linear plot (fig 5.11) visualizes the results from the table and shows the whole data model with a solid line, the BW of 15 in dotted (...) and BW of 20 in long dashed (– –) line.

Based on the observations in table (5.13) and in the plot (5.11) it can be said that when not restricting the BW, there's around 35% increase in CFR if one belongs to the positive group. The negative change in the model are similar observations compared to those in pathologist analysed (5.9), though the difference is approximately 50% larger. The reason for negative difference is unclear, but probably caused by misfit data. Also, it's important to notice that the using whole data is not appropriate either, when of interest is the change around cutoff.

5.5.3 Non-parametric RDD *rdrobust* for algorithmic data

The second approach for the algorithmic results was continued similarly than described in the section 5.4.3: *rdrobust* with three different kernels and four different heteroskedasticity estimates was used to plot and quantify the difference. The figure 5.12 displays the RDplots for centred algorithmic data using uniform, Epanechnikov and triangular kernel and default settings. The changes in the algorithmic data are identical with the plots made using pathologist data (figure 5.6), though the algorithmic data has more bins (blue dots) and the CI (lines on the dots) are smaller. Inspecting the results from the table C.2 reveals that *rdrobust* now manages to use the nearest neighbour VCE in all kernels and now all kernels are capable to measure the size of the jump (coef). Recommended/used bandwidth is now

Table 5.13: **Size of the treatment effect as measured by linear model with differing bandwidths (algorithmic data).**

	Full data	Bandwidth = 20	Bandwidth = 15
(Intercept)	0.108	0.681	0.348
	(0.104)	(0.078)	(0.063)
ki67_centered	0.016	0.099	0.051
	(0.005)	(0.008)	(0.007)
is_positiveTRUE	0.349	-0.958	-0.433
	(0.171)	(0.144)	(0.112)
R2	0.271	0.680	0.438
Num.Obs.	129	100	91

Table illustrates the treatment effect measured with using different bandwidths when the parametric RDD was done with 1st order linear regression. The table is done for centered values and in the table; intercept is the average case fatality risk (see 4.1) at the 14 percentage threshold, ki67_centered is the increase in risk for every point above 14 that subjects Ki-67 value is, is_positiveTRUE is shift in intercept (difference between scores at the threshold) when subject belongs to positive group (has Ki-67 ≥ 14) is true. R2 is a measure of the goodness of fit of a model and the Num.Obs. describe the number of observations the model used. Values below in brackets represent standard error, respectively.

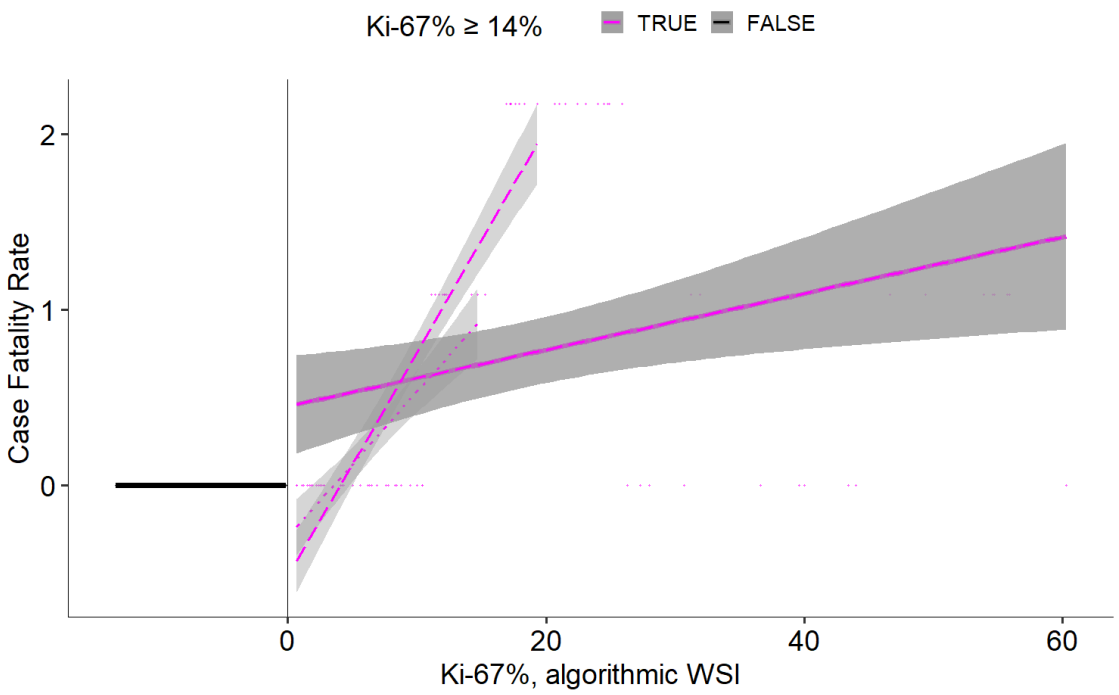

Figure 5.11: **Linear model for algorithm analysed RDD data.** Figure shows the sharp Regression Discontinuity Design (RDD) between Ki-67 groups for whole data (solid line), BW of 20 (dashed) and 15 (dotted), the graph suggests an increase in Case Fatality Rate (see equation 4.1) when Ki-67% increases.

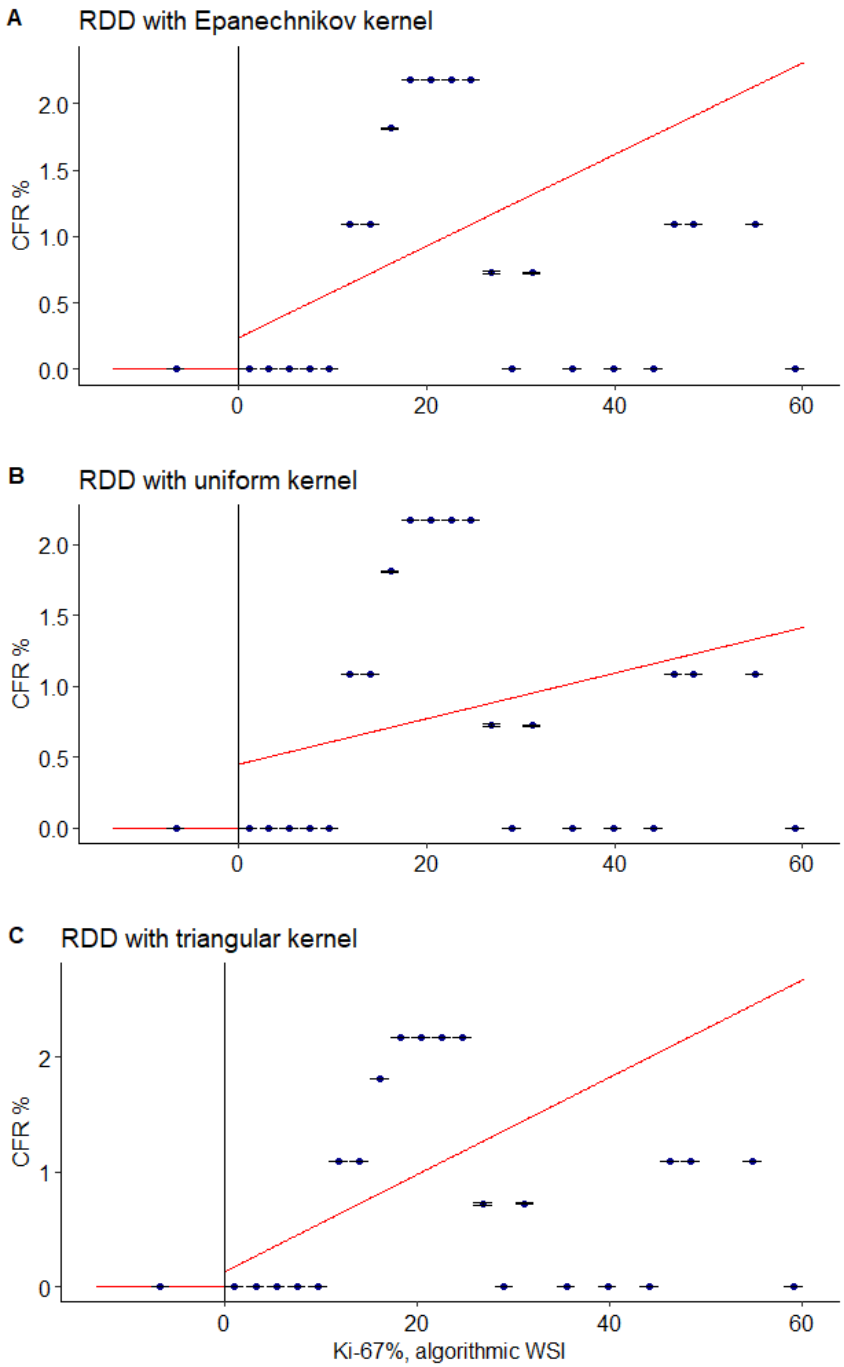

Figure 5.12: **Three non-parametric RD analyses for algorithm analysed Ki-67%.** Figure compiles the results from non-parametric Regression Discontinuity Design (RDD) with the used three different kernels. In the table CFR = Case Fatality Rate and WSI = Whole Slide Image.

constant (60.300/60.300) with all but one kernels and heteroskedasticity estimators. The Epanechnikov kernel with nearest neighbour (nn) estimator stands out having BW of 3.081/3.081 and using only observations around the cutoff as indicated by the effective number of observations 8 in the left of the cutoff and 4 in the right. Every other *k-VCE* combination uses all (47/82) or all but one (47/81) observation. The coef is now constant in triangular (0.131) and in uniform (0.450) kernels, the epan-nn has different coef (-0.149) when compared to rest (0.239) The statistical significance of the observations in the algorithmic data (C.2) follow an opposite logic when compared to human results (C.1): the robust statistics show $p \leq 0.05$, whereas the conventional is insignificant. The epan-nn combination is statistically insignificant in both ways, though results otherwise are relevant.

5.5.4 Local Randomization RDD

Following the implementations to human produced data (chapter 5.4.4), *rdwinselect* function was used to first select a proper window size using uniform kernel for centralized and non-centralized data, see figure 5.13. For non-centralized data (A) recommended window was [3;25] with 77 observations (43 below, 34 above) and for the centralized data (B) recommended window was [-11;11] with 77 observations (43 below, 34 above). The window size for centralized algorithmic data is quite like that of pathologist data [-15;15], but the window in non-centralized algorithmic data differs and is larger than window in the pathologist data [-5;19].

The results from the local randomization RDD – *rdrandinf* – analyses for the algorithmic data are listed in the table 5.14. Table shows that the measured effects (T) are identical to pathologist data (see 5.10), though the used window is now smaller to that in pathologist dataset. The table for the algorithmic *rdrandinf* gives further confirmation that in the current data, measurable effects are present only when the whole width of the data is used.

Table 5.14: **Results from local randomization RDD using algorithmic data.**

c	n	k	stat	win	T	P |T|
14	47/82	uni	d.m	[0.7;74.3]	0.742	0.000
14	47/82	epan	d.m	[0.7;74.3]	0.713	0.000
14	43/34	uni	d.m	[3;25]	0.000	1.000
0	43/34	uni	d.m	[-13.3;60.3]	0.742	0.000
0	43/34	uni	d.m	[-11;11]	0.000	1.000
0	37/39	uni	K-S	[-11;11]	0.000	1.000
0	37/39	uni	r.s.z	[-11;11]	0.000	1.000

The table presents a summary of the local randomization Regression Discontinuity Design (RDD) results using algorithmic data. In the table c = cutoff, n = number of observations, stat = statistical method used, d.m = difference in means, K-S = Kolmogorov-Smirnov, r.s.z = Rank sum z-stat, win = windows size, T = treatment effect, P >|T| = statistical significance of the treatment effect.

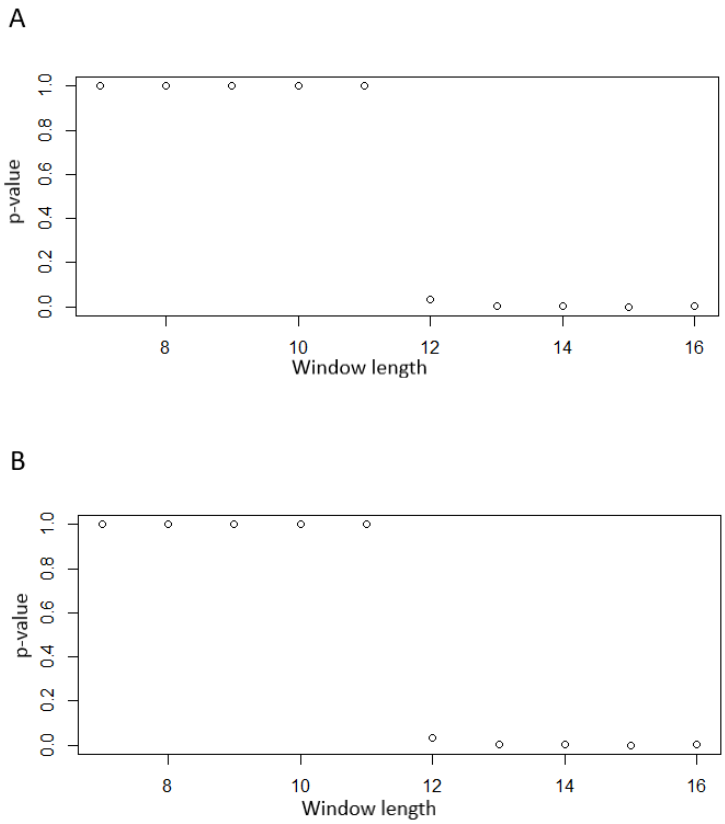

Figure 5.13: **Plots for recommended window size using *rdwinselect*, algorithmic data.** Results from selecting the most optimal window (one before $p < 0.05$) with finite-sample methods near the cutoff where the assumption of randomized treatment assignment is most plausible. In the image windows for non-centralized data (A) and centralized data (B).

5.5.5 Falsification results

To prove/confirm the RDD results, falsification tests/sanity checks were done for the algorithmic data like HUS pathologist data (see 5.4.5). Continuity and manipulation have been checked in chapter 5.5.1 and different bandwidths have been used and discussed in the chapters 5.5.20 & 5.5.4, respectively. The further falsification results were done by plotting RDD for placebo outcomes and for placebo cutoffs. Placebo

cutoffs were chosen to be double (28) and triple (42) compared to the original (14) and the placebo outcome was chosen to be age, as age can't be affected by any treatment whatsoever.

In realistic dataset the placebo cutoff and outcome shouldn't show any effect, but as can be seen from the figure 5.14, this was not the case, despite the algorithmic data being more evenly distribute than the pathologist data. At placebo cutoffs 28 (5.14A) and 42 (5.14B) there are evident jumps appearing at the cutoff, also the plot shows an interesting kink and a clear difference when age is used as an outcome variable (5.14C). The plots clearly illustrate the problems of using a small and unevenly distributed dataset.

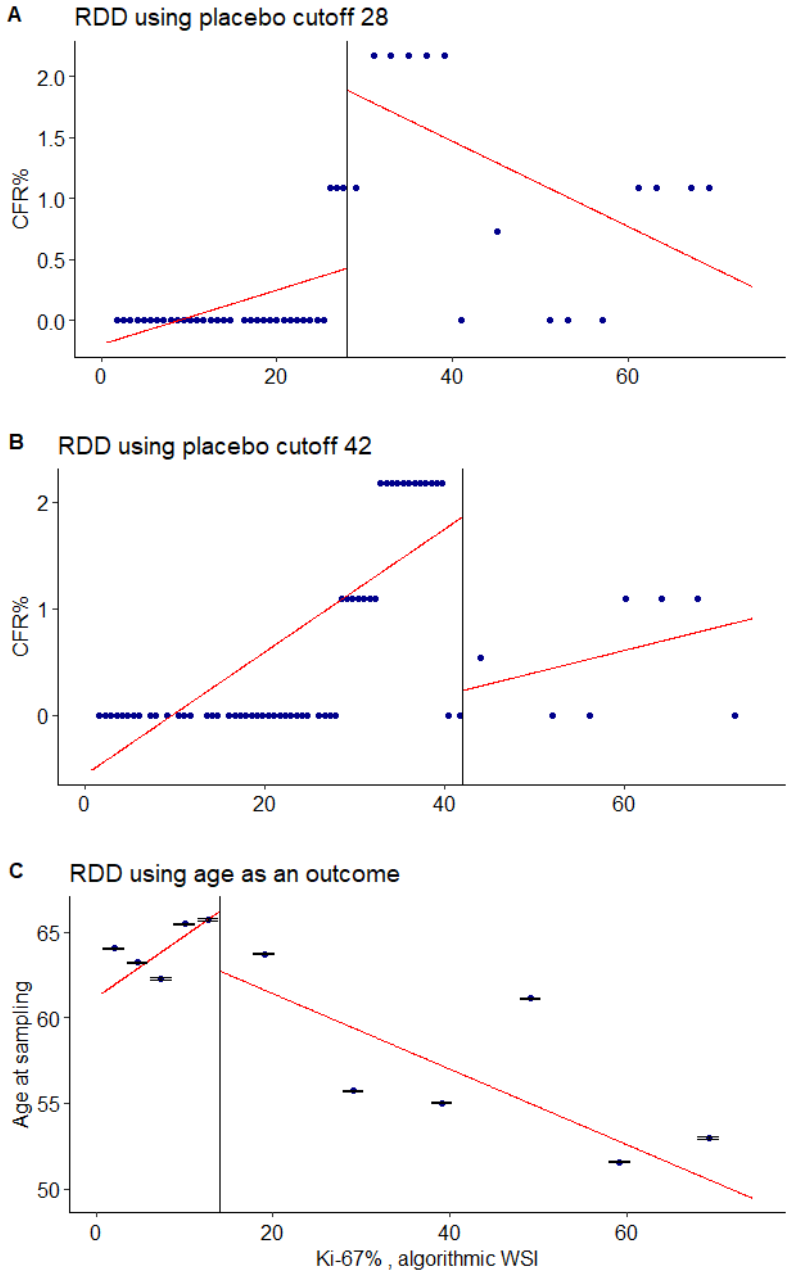

Figure 5.14: **Falsification RDD plots using linear regression and uniform kernel, algorithm data.** Figure compiles the falsification Regression Discontinuity Design (RDD) results done for algorithm data. data using placebo cutoff (A-B) and placebo outcome. For robust results, placebo cutoffs and outcomes should now show any significant differences. In the image CFR = Case Fatality Rate and WSI = Whole Slide Image.

6 Discussion

With literature review and a small-scale proof-of-concept RDD, this thesis aimed to a) review the current usage of RDD in medical sciences and b) to review the possibility of implementing RDD to histopathological data after analysis run in Aiforia Cloud.

The literature reviews indicates that although the RDD has existed for 60 years, the usage of the method in medical research has been a rarity until the 2010's. From since the usage of RDD has had a small but steady growth, but the method continues to be a rare sight in medical research; majority of studies are under medical policy and only a handful of studies use a clinical measure as running variable. Almost as an indication of the rarity, only three reviews – Moscoe et.al. 2015, Venkataramani et. al. 2016 & Hilton Boon et. al. 2021 – could be found reviewing the usage of RDD in medical research [31]–[33].

Out of the cancers currently interested by Aiforia, the RDD was found to be applied only to prostate cancer with PSA as a running variable. When estimating use of RDD with the main assumption of RDD – use of some arbitrary cut-off, continuous running variable and inability to completely manipulate the running variable and sort around the cutoff – any of the reviewed cancers should be compatible for RDD as the conditions should be met in all cases.

When clinical data is used in decision making, cutoffs are often used to guide decision making and to divide patient to different groups [31]. The reviewed cancers

are no exception (see chapter 3.2.2 for details); each reviewed cancer had specific diagnostic classes and treatments were largely based on the diagnostic class. Although the nature of medical treatments relies largely on consideration and lead to imperfect compliance, in those cases it's possible to use fuzzy RD as an approach. Also, all but prostate cancer uses a bio-marker with continuous scale, making them easily implementable for RDD. Though the continuous running variable isn't necessary needed, as the Local Randomization RD can be used for categorical variables [27], thus it could be used for prostate cancer's Grading Group as well. The manipulation of the running variable is measurable (see eg. [24]) but one can also argue with a relative confidence against manipulation in real world medical data. For a subject to manipulate/sort own measurement results would be nearly impossible and for a clinician to manipulate/sort would also be totally unethical and illegal.

The reasons for the rarity of the RDD in medical research are somewhat speculative. Articles reviewing the RDD present that the major reason behind the rarity could be "platinum podium" of the RCT and lack of knowledge of quasi-experiments, particularly from RDD [12], [32], [33], [60]. Quasi-experiments might be shunned by medical researchers as the quasi-experiments usually need "heavy proofs" for the underlying assumptions to make valid causal estimates, which might make using them not so appealing [12]. While RDD has far less restriction and assumptions to make, this seems to be left much unnoticed by researchers [15]. Also there has been a misunderstanding throughout the history of RDD, that the applicability would be limited to only few narrow topics, whilst the RDD can be used to nearly any data that uses cut-off-based divisions and the use cases in economy and sociology show the versatility of the RDD. [15], [31]–[33]

The technical implementation of RDD with the used dataset and outcome (CFR derived from deaths) sets some limitations for interpretation of the results as there wasn't enough observations and the outcome wasn't the best choice for Ki-67.

The survival time calculations were used to assay the underlying group differences between the positive and moderate Ki-67 groups and to inspect if there was enough variation for death to be used as an outcome. Of the 37 subjects in moderate group deceased and in positive group out of 92 subjects, 8 deceased during the five year follow up. The low number of observations (ie. good survival) prevented using median survival values (time it takes to reach 50% survival) and Cox model is a hazard ratio (HR), but the results that could be used showed clear difference between groups. The finding for overall survival rate of 94% is in line with the literature: the average 5y survival ratio in Finland for a woman with a breast cancer is $\sim$ 91% [37]. As the Ki-67 is a cell proliferation marker, the prognosis is usually slightly worse; in the thesis data, Ki-67 positives had 5y survival rate of 91%, which is like findings in German women from 2013 who had a survival rate of 89.3% [61].

The RDD analysis is known to need 3-4 times more observations than a RCT [22] to display any reasonable effect, this meant that a set of 129 observations was far less than perfectly adequate for the study. The scarceness and in the case of pathologist data, the uneven distribution were affecting throughout the study. This can be visualized from the histograms that look uneven and have noticeable jumps, which might tell about manipulation, so to truly separate a discontinuity in the density from noise using a density test, one would need plenty of observations at cutoff [54]. The density plot for pathologist data (fig 5.4) is a good illustration of a plot really not showing anything. For algorithmic data (fig 5.10) the plot is clearly better as the observations are more heterogeneous, but still the displayed discontinuity – though not significant – might be just noise.

In illustrative sense, the scarceness of the data had benefits when different weighted kernels were used. Whereas the uniform rectangular kernel and Epanechnikov used similar amounts of observations to choose bandwidths, the triangular clearly reduced the number of observations. None of the kernels in pathologist data

tried to reduce the number of observations left to cutoff (ie. the moderates). Interestingly the all but one kernel-VCE combinations in algorithmic data used all existing data, though the algorithmic data was smoother than the pathologist data. The Epanechnikov kernel with nearest neighbour VCE ended using a realistic number of observations, 8/4, though it might be just an illustration of the *multiple testing problem* (e.g. see Jensen & Cohen 2000 [62]). Taken together, it can be said that as the kernel choice does make some difference, the results are highly sensitive to the choice of bandwidth and not very credible [30].

The falsification of the RDD results proved also that the data wasn't 100% fit for analysis. In real life one would use only small steps around the cutoff, to create placebo cutoffs, but in thesis case it wasn't possible as the RDD couldn't be calculated for those areas. Instead, more extreme placebo cutoffs were used to illustrate the problems in the data. Also, there's an uncertainty whether age is appropriate placebo outcome, but in the case it would be, it would underscore the data problems.

Also, it needs to be mentioned about the original outcome that using a five-year survival rate isn't likely the best outcome. When any time-to-event study is executed, it needs to capture enough events with long enough follow-up, for breast cancer a 5-year follow-up yields only a short- to-medium-term indication of survival [63]. For those reasons, it might be more beneficial to use some other outcome.

The main goal however was to review the possibility of technical implementation of the RDD after running image analysis on Aiforia Cloud. Regardless of the challenges of not getting actual real-life evidence, the technical possibility can be seen as shown true. Using Aiforia API one could get the data out from the Aiforia Cloud and data could be mangled into appropriate format using R. The RDD could be done using the libraries and online tutorials, the author found especially beneficial to study the work of Calonico et. al [25]. No obstacles were observed of

using algorithmic data for the analysis and histopathological data in general seems to suite well for the analysis as it easily meets the main criteria of RDD

6.1 Future work

The RDD offers a plethora of fields to where it can be implemented on. In the future, one might repeat this study with a proper number of study subjects (several thousands) to get real results. In principle death rate is appropriate for RDD, but one must use consideration when choosing the follow-up time, 5y might suite if there are enough observations. Moreover, one might consider using outcomes such as relapse rate or some hospital visit index and one would need exact information about the treated vs. untreated to make correct estimates.

There shouldn't be any obstacles for implementing RDD to any other breast cancer biomarker. Actually, it might be more beneficial for the medical research to implement RDD to ER, PR or HER2 as they are currently used for treatment decisions [37]. One might even do a RDD where all of these markers are taken into account as multiple cutoffs or one of those as running variable and others as covariates.

Also, in clinical point of view, as the algorithmic WSI results won't be part of patient diagnostics, this study setting should be repeated with CAD results rather than using solely algorithmic results.

Finally, for anyone to continue use the RDD in medical sciences, it might be a good idea to get acquainted to the very recent work of Cattaneo & al. [64] of doing analyses on medical data.

7 Conclusions

This thesis aimed to review the technical feasibility of implementing RDD to histopathological data analysed in Aiforia Cloud and assessing the use cases of RDD in cancer research.

The literature review showed that the RDD has had small growth during the last ten years, but despite growth it still continues to be a rarely used method in medical research. Very few studies implemented RDD in some histopathology related data, while the RDDs assumptions could be easily met by clinical data used in cancer diagnostics. Patients are diagnosed and receive treatment based on some clinical findings that use clear cutoffs and while there's a possibility of being a noncomplier or receiving treatment despite falling of the cutoff, the fuzzy RDD is a possible approach for these types of settings.

Although thesis study has limitations on how generalisable the results are, the study produced knowledge of different aspects of RDD and proved the technical possibility of applying RDD to histopathological data analysed in Aiforia Cloud. Using Aiforia API in R, the data could be exported into R for further analysis and the libraries required to run RDD are well updated and are based on active research of the topic. By following the recommendations from literature, each step from manipulation checks to quality check of the results by falsification could – technically – be made with the exported data. In the future one might reproduce this study with large enough data and more relevant outcome.

References

[1] J. Pearl, *Causality: Models, Reasoning, and Inference.* (Causality: Models, Reasoning, and Inference.). New York, NY, US: Cambridge University Press, 2000, pp. xvi, 384, xvi, 384, ISBN: 0-521-77362-8 (Hardcover). [Online]. Available: https://doi.org/10.1017/CBO9780511803161.

[2] T. J. Kaptchuk, "The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf?", *Journal of Clinical Epidemiology*, vol. 54, no. 6, pp. 541–549, Jun. 1, 2001, ISSN: 0895-4356. DOI: 10.1016/S0895-4356(00)00347-4.

[3] P. G. Shekelle, S. H. Woolf, M. Eccles, and J. Grimshaw, "Developing guidelines", *BMJ*, vol. 318, no. 7183, pp. 593–596, Feb. 27, 1999, ISSN: 0959-8138. pmid: 10037645. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1115034/.

[4] J. Higgins, J. Thomas, J. Chandler, *et al.*, Eds., *Cochrane Handbook for Systematic Reviews of Interventions, Version 6.3 (Updated February 2022)*. Cochrane, Feb. 2022. [Online]. Available: https://training.cochrane.org/handbook/current.

[5] World Health Organization and Regional Office for the Eastern Mediterranean, *Guidelines for Management of Breast Cancer.* Cairo: World Health Organization, Regional Office for the Eastern Mediterranean, 2006, ISBN: 978-92-9021-

405-2. [Online]. Available: https://library.dctabudhabi.ae/sirsi/detail/1237442.

[6] F. Cardoso, S. Kyriakides, S. Ohno, *et al.*, "Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up", *Annals of Oncology*, vol. 30, no. 8, pp. 1194–1220, Aug. 2019, ISSN: 09237534. DOI: 10.1093/annonc/mdz173.

[7] K. H. Allison, M. E. H. Hammond, M. Dowsett, *et al.*, "Estrogen and Progesterone Receptor Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Guideline Update", *Arch. Pathol. Lab. Med.*, vol. 144, no. 5, pp. 545–563, May 1, 2020, ISSN: 0003-9985, 1543-2165. DOI: 10.5858/arpa.2019-0904-SA.

[8] S. Banerji and S. Mitra, "Deep learning in histopathology: A review", *WIREs Data Min. Knowl. Discov.*, vol. 12, no. 1, e1439, 2022, ISSN: 1942-4795. DOI: 10.1002/widm.1439.

[9] G. A. Pignatiello, R. J. Martin, and R. L. Hickman, "Decision Fatigue: A Conceptual Analysis", *J Health Psychol*, vol. 25, no. 1, pp. 123–135, Jan. 2020, ISSN: 1359-1053. DOI: 10.1177/1359105318763510. pmid: 29569950.

[10] L. Pantanowitz, D. Hartman, Y. Qi, *et al.*, "Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses", *Diagn Pathol*, vol. 15, no. 1, pp. 1–10, 1 Dec. 2020, ISSN: 1746-1596. DOI: 10.1186/s13000-020-00995-z.

[11] W. Bulten, H. Pinckaers, H. van Boven, *et al.*, "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study", *The Lancet Oncology*, vol. 21, no. 2, pp. 233–241, Feb. 1, 2020, ISSN: 1470-2045. DOI: 10.1016/S1470-2045(19)30739-9.

[12] T. Bärnighausen, C. Oldenburg, P. Tugwell, *et al.*, "Quasi-experimental study designs series—paper 7: Assessing the assumptions", *Journal of Clinical Epidemiology*, vol. 89, pp. 53–66, Sep. 1, 2017, ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2017.02.017.

[13] A. Luiro and F.-M. Rostedt, *Inderes yhtiöraportti - Aiforia Technologies Oyj, seurannan aloitus.* Jun. 23, 2022. [Online]. Available: https://www.inderes.fi/fi/system/files/company-reports/220623_aiforia_seurannan_aloitus.pdf.

[14] "Aiforia | Aiforia as a company | Aiforia's business". (), [Online]. Available: https://investors.aiforia.com/aiforias-business.

[15] D. S. Lee and T. Lemieux, "Regression Discontinuity Designs in Economics", *J. Econ. Lit.*, vol. 48, no. 2, pp. 281–355, Jun. 2010, ISSN: 0022-0515. DOI: 10.1257/jel.48.2.281.

[16] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin, 2001, 623 pp., ISBN: 978-0-395-61556-0.

[17] G. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York: Cambridge University Press, 2015, 625 pp., ISBN: 978-0-521-88588-1. [Online]. Available: https://doi.org/10.1017/CBO9781139025751.

[18] D. L. Thistlethwaite and D. T. Campbell, "Regression-discontinuity analysis: An alternative to the ex post facto experiment.", *Journal of Educational Psychology*, vol. 51, no. 6, pp. 309–317, 1960, ISSN: 0022-0663. DOI: 10.1037/h0044319.

[19] T. D. Cook, ""Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics", *Journal of Econometrics*, The

Regression Discontinuity Design: Theory and Applications, vol. 142, no. 2, pp. 636–654, Feb. 1, 2008, ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2007.05.002.

[20] G. Imbens and T. Lemieux, "Special issue editors' introduction: The regression discontinuity design—Theory and applications", *Journal of Econometrics*, The Regression Discontinuity Design: Theory and Applications, vol. 142, no. 2, pp. 611–614, Feb. 1, 2008, ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2007.05.008.

[21] A. S. Goldberger, "Selection bias in evaluating treatment effects: Some formal illustrations", in *Modelling and Evaluating Treatment Effects in Econometrics*, ser. Advances in Econometrics, T. Fomby, R. Carter Hill, D. L. Millimet, J. A. Smith, and E. J. Vytlacil, Eds., vol. 21, Bingley: Emerald Group Publishing Limited, Feb. 21, 2008, pp. 1–31, ISBN: 978-0-7623-1380-8. DOI: 10.1016/S0731-9053(07)00001-1.

[22] J. C. Cappelleri, R. B. Darlington, and W. M. Trochim, "Power Analysis of Cutoff-Based Randomized Clinical Trials", *Eval Rev*, vol. 18, no. 2, pp. 141–152, Apr. 1, 1994, ISSN: 0193-841X. DOI: 10.1177/0193841X9401800202.

[23] P. Z. Schochet, "Statistical Power for Regression Discontinuity Designs in Education Evaluations", *Journal of Educational and Behavioral Statistics*, vol. 34, no. 2, pp. 238–266, Jun. 2009, ISSN: 1076-9986, 1935-1054. DOI: 10.3102/1076998609332748.

[24] J. McCrary, "Manipulation of the running variable in the regression discontinuity design: A density test", *Journal of Econometrics*, The Regression Discontinuity Design: Theory and Applications, vol. 142, no. 2, pp. 698–714, Feb. 1, 2008, ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2007.05.005.

[25] S. Calonico, M. D. Cattaneo, R. Chandak, *et al.*, *RD Packages*, Oct. 18, 2022. [Online]. Available: https://rdpackages.github.io/rdrobust/.

[26] A. Gelman and G. Imbens, "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs", *Journal of Business & Economic Statistics*, vol. 37, no. 3, pp. 447–456, Jul. 3, 2019, ISSN: 0735-0015, 1537-2707. DOI: 10.1080/07350015.2017.1366909.

[27] M. D. Cattaneo, N. Idrobo, and R. Titiunik, *A Practical Introduction to Regression Discontinuity Designs: Extensions*, Jul. 14, 2018. [Online]. Available: https://cattaneo.princeton.edu/books/Cattaneo-Idrobo-Titiunik_2018_CUP-Vol2.pdf.

[28] D. Card, D. S. Lee, Z. Pei, and A. Weber, "Regression kink design: Theory and practice", *Adv. Econom.*, vol. 38, pp. 341–382, 2017, ISSN: 0731-9053. DOI: 10.1108/S0731-905320170000038016.

[29] M. D. Cattaneo, N. Idrobo, and R. Titiunik, *A Practical Introduction to Regression Discontinuity Designs: Foundations* (Elements in Quantitative and Computational Methods for the Social Sciences), 1st ed. Cambridge University Press, Nov. 30, 2019, ISBN: 978-1-108-71020-6. DOI: 10.1017/9781108684606.

[30] G. W. Imbens and T. Lemieux, "Regression discontinuity designs: A guide to practice", *Journal of Econometrics*, The Regression Discontinuity Design: Theory and Applications, vol. 142, no. 2, pp. 615–635, Feb. 1, 2008, ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2007.05.001.

[31] A. S. Venkataramani, J. Bor, and A. B. Jena, "Regression discontinuity designs in healthcare research", *BMJ*, vol. 352, p. i1216, Mar. 14, 2016, ISSN: 1756-1833. DOI: 10.1136/bmj.i1216. pmid: 26977086.

[32] M. Hilton Boon, P. Craig, H. Thomson, M. Campbell, and L. Moore, "Regression Discontinuity Designs in Health: A Systematic Review", *Epidemiology*, vol. 32, no. 1, pp. 87–93, Jan. 2021, ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001274.

[33] E. Moscoe, J. Bor, and T. Bärnighausen, "Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: A review of current and best practice", *J Clin Epidemiol*, vol. 68, no. 2, pp. 122–133, Feb. 2015, ISSN: 1878-5921. DOI: 10.1016/j.jclinepi.2014.06.021. pmid: 25579639.

[34] L. M. Smith, E. C. Strumpf, J. S. Kaufman, A. Lofters, M. Schwandt, and L. E. Lévesque, "The Early Benefits of Human Papillomavirus Vaccination on Cervical Dysplasia and Anogenital Warts", *Pediatrics*, vol. 135, no. 5, e1131–e1140, May 1, 2015, ISSN: 0031-4005. DOI: 10.1542/peds.2014-2961.

[35] P. Gild, R. S. Pompe, T. Seisen, *et al.*, "Regression Discontinuity Analysis of Salvage Radiotherapy in Prostate Cancer", *European Urology Oncology*, vol. 4, no. 5, pp. 817–820, Oct. 2021, ISSN: 25889311. DOI: 10.1016/j.euo.2019.08.005.

[36] J. Shoag, J. Halpern, B. Eisner, *et al.*, "Efficacy of Prostate-Specific Antigen Screening: Use of Regression Discontinuity in the PLCO Cancer Screening Trial", *JAMA Oncology*, vol. 1, no. 7, pp. 984–986, Oct. 1, 2015, ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2015.2993.

[37] Suomen Rintasyöpäryhmä Ry, *Rintasyövän Valtakunnallinen Diagnostiikka Ja Hoitosuositus 2022*. Suomen Rintasyöpäryhmä Ry, May 30, 2022. [Online]. Available: https://rintasyoparyhma.yhdistysavain.fi/hoitosuositus/.

[38] M. Sudah, *Rintadiagnostiikan Opas*, 4th ed., Suomen Rintasyöpäryhmä Ry, Ed. Kuopio: Suomen rintaradiologit, Mar. 1, 2019, 157 pp. [Online]. Available:

https://rintasyoparyhma.yhdistysavain.fi/@Bin/183959/Rintadiagnostiikan%20opas%204%20painos.pdf.

[39] B. Keam, S.-A. Im, K.-H. Lee, *et al.*, "Ki-67 can be used for further classification of triple negative breast cancer into two subtypes with different response and prognosis", *Breast Cancer Res*, vol. 13, no. 2, R22, 2011, ISSN: 1465-5411. DOI: 10.1186/bcr2834. pmid: 21366896.

[40] A. Goldhirsch, E. Winer, A. Coates, *et al.*, "Personalizing the treatment of women with early breast cancer: Highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013", *Annals of Oncology*, vol. 24, no. 9, pp. 2206–2223, Sep. 2013, ISSN: 09237534. DOI: 10.1093/annonc/mdt303.

[41] A. C. Wolff, M. E. H. Hammond, K. H. Allison, *et al.*, "Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update", *Arch. Pathol. Lab. Med.*, vol. 142, no. 11, pp. 1364–1382, Nov. 1, 2018, ISSN: 0003-9985, 1543-2165. DOI: 10.5858/arpa.2018-0902-SA.

[42] W. D. Travis, E. Brambilla, A. G. Nicholson, *et al.*, "The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification", *Journal of Thoracic Oncology*, vol. 10, no. 9, pp. 1243–1260, Sep. 1, 2015, ISSN: 1556-0864. DOI: 10.1097/JTO.0000000000000630.

[43] Working group set up by the Finnish Medical Society Duodecim, Suomen Keuhkolääkäriyhdistys ry and Suomen Onkologiayhdistys ry, *Lung Cancer. Current Care Guidelines.* 2017. [Online]. Available: www.kaypahoito.fi.

[44] E. C. Paver, W. A. Cooper, A. J. Colebatch, *et al.*, "Programmed death ligand-1 (PD-L1) as a predictive marker for immunotherapy in solid tumours: A

guide to immunohistochemistry implementation and interpretation", *Pathology*, vol. 53, no. 2, pp. 141–156, Feb. 1, 2021, ISSN: 0031-3025. DOI: 10.1016/j.pathol.2020.10.007.

[45] J. I. Epstein, L. Egevad, M. B. Amin, *et al.*, "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System", *Am. J. Surg. Pathol.*, vol. 40, no. 2, pp. 244–252, Feb. 2016, ISSN: 0147-5185. DOI: 10.1097/PAS.0000000000000530.

[46] P. A. Humphrey, "Gleason grading and prognostic factors in carcinoma of the prostate", *Mod Pathol*, vol. 17, no. 3, pp. 292–306, Mar. 2004, ISSN: 1530-0285. DOI: 10.1038/modpathol.3800054.

[47] A. Erickson, K. Sandeman, K. Lahdensuo, *et al.*, "New prostate cancer grade grouping system predicts survival after radical prostatectomy", *Human Pathology*, vol. 75, pp. 159–166, May 1, 2018, ISSN: 0046-8177. DOI: 10.1016/j.humpath.2018.01.027.

[48] Working group set up by the Finnish Medical Society Duodecim and Finnish Urological Society, *Prostate Cancer. Current Care Guidelines.* 2014. [Online]. Available: www.kaypahoito.fi.

[49] P. Schober and T. R. Vetter, "Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare", *Anesth. Analg.*, vol. 127, no. 3, pp. 792–798, Sep. 2018, ISSN: 0003-2999. DOI: 10.1213/ANE.0000000000003653.

[50] J. M. Last and I. E. Association, Eds., *A Dictionary of Epidemiology*, 4th ed. New York: Oxford University Press, 2001, 196 pp., ISBN: 0-19-514168-7.

[51] A. Heiss, "GSU Program Evaluation for Public Service - Regression discontinuity", Georgia State University – Andrew Young School of Policy Studies,

May 11, 2020. [Online]. Available: https://evalf20.classes.andrewheiss.com/example/rdd/.

[52] R. Titiunik and M. D. Cattaneo, "Regression Discontinuity Designs - NBER", in *Summer Inst. 2021*, National Bureau of Economic Research, Jul. 30, 2021. [Online]. Available: https://www.nber.org/conferences/si-2021-methods-lecture-causal-inference-using-synthetic-controls-and-regression-discontinuity.

[53] M. D. Cattaneo, M. Jansson, and X. Ma, "Simple Local Polynomial Density Estimators", *Journal of the American Statistical Association*, vol. 115, no. 531, pp. 1449–1455, Jul. 2, 2020, ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2019.1635480.

[54] S. Cunningham, *Causal Inference: The Mixtape*. New Haven ; London: Yale University Press, 2021, 572 pp., ISBN: 978-0-300-25168-5.

[55] W. Zucchini, *APPLIED SMOOTHING TECHNIQUES*, Oct. 2003. [Online]. Available: http://staff.ustc.edu.cn/~zwp/teach/Math-Stat/kernel.pdf.

[56] S. Węglarczyk, "Kernel density estimation and its application", *ITM Web Conf.*, vol. 23, W. Zielinski, L. Kuchar, A. Michalski, and B. Kazmierczak, Eds., p. 00 037, 2018, ISSN: 2271-2097. DOI: 10.1051/itmconf/20182300037.

[57] J. A. Rice, *Mathematical Statistics and Data Analysis* (Duxbury Advanced Series), 3rd ed. Belmont, CA: Thomson/Brooks/Cole, 2007, 1 p., ISBN: 978-0-534-39942-9.

[58] S. Date, *Introducing the White's Heteroskedasticity Consistent Estimator*, Oct. 25, 2022. [Online]. Available: https://towardsdatascience.com/introducing-the-whites-heteroskedasticity-consistent-estimator-821beee28516.

[59] A. F. Hayes and L. Cai, "Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation", *Behavior Research Methods*, vol. 39, no. 4, pp. 709–722, Nov. 1, 2007, ISSN: 1554-3528. DOI: 10.3758/BF03192961.

[60] A. D. Harris, J. C. McGregor, E. N. Perencevich, *et al.*, "The Use and Interpretation of Quasi-Experimental Studies in Medical Informatics", *J Am Med Inform Assoc*, vol. 13, no. 1, pp. 16–23, 2006, ISSN: 1067-5027. DOI: 10.1197/jamia.M1749. pmid: 16221933.

[61] E. C. Inwald, M. Klinkhammer-Schalke, F. Hofstädter, *et al.*, "Ki-67 is a prognostic parameter in breast cancer patients: Results of a large population-based cohort of a cancer registry", *Breast Cancer Res Treat*, vol. 139, no. 2, pp. 539–552, 2013, ISSN: 0167-6806. DOI: 10.1007/s10549-013-2560-8. pmid: 23674192.

[62] D. D. Jensen and P. R. Cohen, "Multiple Comparisons in Induction Algorithms", *Machine Learning*, vol. 38, no. 3, pp. 309–338, Mar. 1, 2000, ISSN: 1573-0565. DOI: 10.1023/A:1007631014630.

[63] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival Analysis Part I: Basic concepts and first analyses", *Br J Cancer*, vol. 89, no. 2, pp. 232–238, Jul. 2003, ISSN: 0007-0920, 1532-1827. DOI: 10.1038/sj.bjc.6601118.

[64] M. D. Cattaneo, L. Keele, and R. Titiunik, *A Guide to Regression Discontinuity Designs in Medical Applications*, May 24, 2022. [Online]. Available: https://rdpackages.github.io/references/Cattaneo-Keele-Titiunik_2022_RDIntroMed.pdf.

Appendix A Survival plots

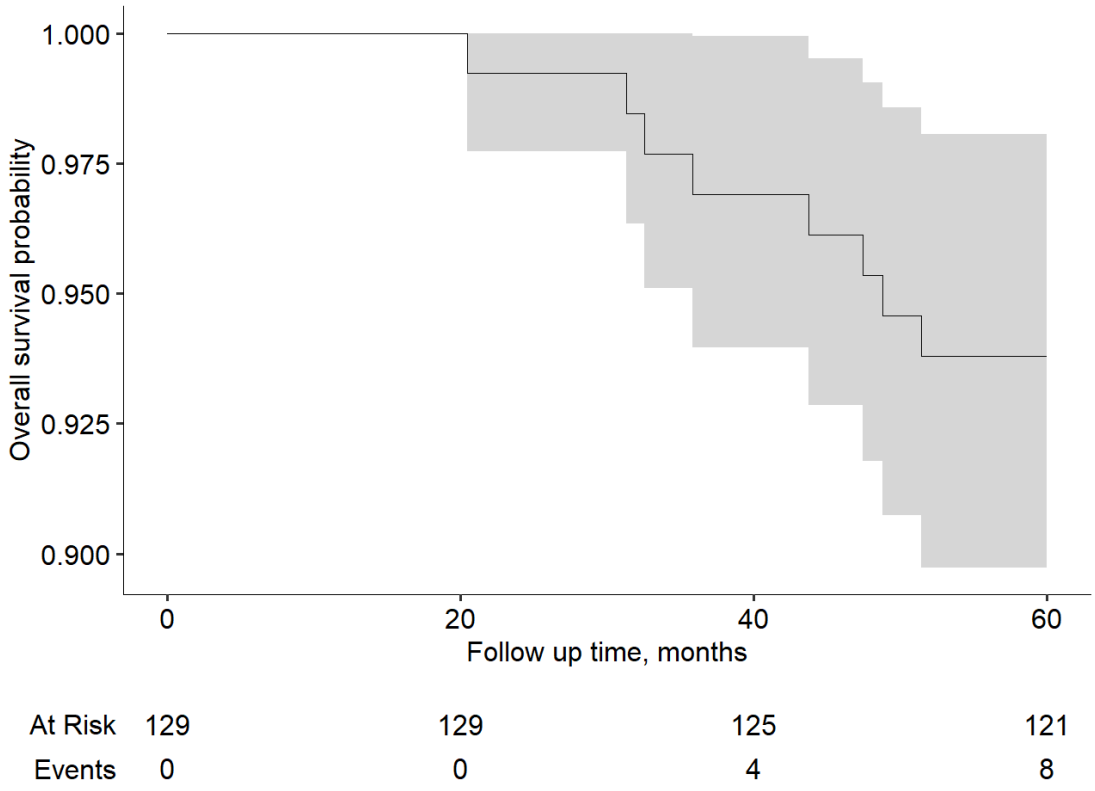

Figure A.1: **Overall survival probability plot.** The figure presents the overall survival (i.e. probability of being alive) for the ungrouped data during 60 months follow up. The "Events" represent deaths and "At risk" individuals at the risk during the observation point. Curve represents the risk and the shaded area is the confidence interval for the risk.

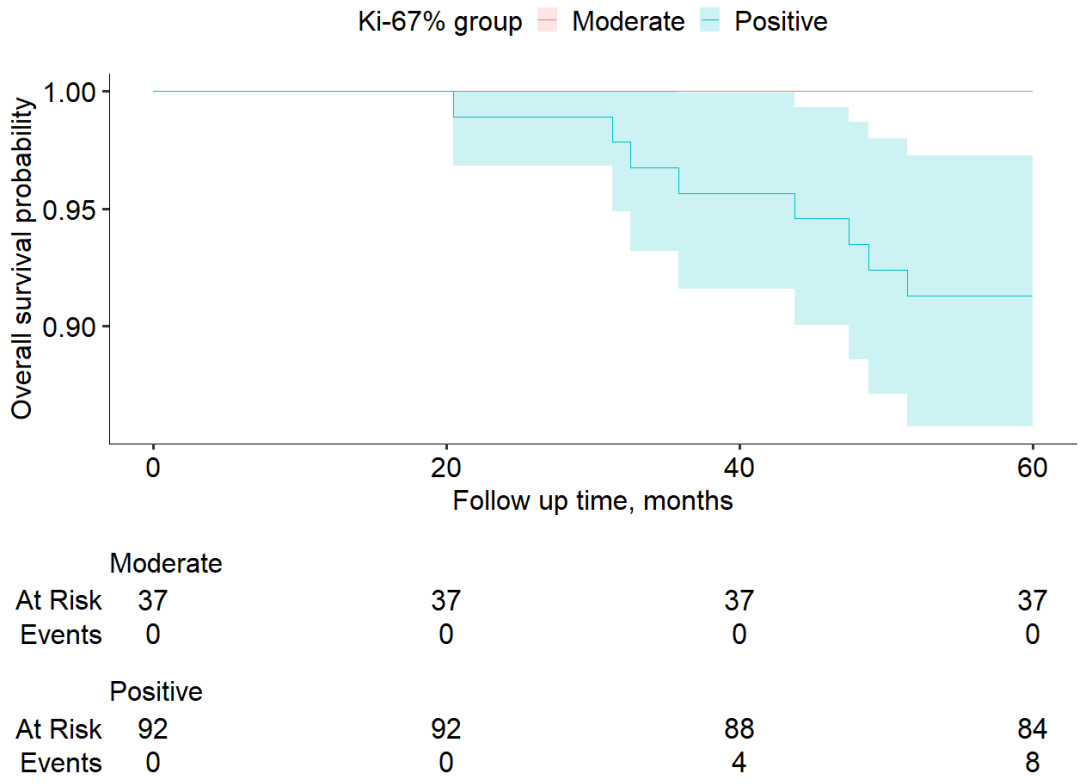

Figure A.2: **Ki-67% grouped survival probability plot.** The figure presents the overall survival (i.e. probability of being alive) for the moderate (0) and positive (1) groups during 60 months follow up. The "Events" represent deaths and "At risk" individuals at the risk during the observation point. Curve represents the risk and the shaded area is the confidence interval for the risk.

Appendix B CFR bins

The table B.1 lists separately for pathologist data and for algorithmic data the bins that were used for the calculation of Case Fatality Rate (CFR)%.

Table B.1: **CFR bins.**

Hus Pathologist	Algorithm
1	0
2	5
3	10
4	15
5	20
7	25
10	30
15	35
20	40
25	45
30	50
35	55
40	60

Continued on next page

Table B.1 continued from previous page

45	65
50	70
60	75
70	
75	
80	
90	
95	

Table B.2: **CFR's based on HUS pathologist Ki-67%.**

Group (by HUS Ki-67)	n	CFR%	Deaths
1	2	0.00	0
2	3	0.00	0
3	2	0.00	0
4	1	0.00	0
5	12	0.00	0
7	1	0.00	0
10	16	0.00	0
15	13	0.00	0
20	15	0.00	0
25	11	0.00	0
30	12	1.09	1
35	5	0.00	0
40	9	2.17	2
45	1	1.09	1
50	7	2.17	2
60	7	0.00	0
70	1	1.09	1
75	1	0.00	0
80	7	1.09	1
90	2	0.00	0
95	1	0.00	0

Table present the calculated Case Fatality Rate (CFR) for each bin, number of subjects in the bin and number of deaths in the bin.

Table B.3: **CFR based on algorithmic Ki-67%.**

Group	n	CFR	Deaths
(0,5]	17	0.00	0
(5,10]	21	0.00	0
(10,14]	9	0.00	0
(14,20]	21	0.00	0
(20,25]	13	0.00	0
(25,30]	11	1.09	1
(30,35]	10	2.17	2
(35,40]	8	2.17	2
(40,45]	4	0.00	0
(45,50]	2	1.09	1
(50,55]	3	0.00	0
(55,60]	2	0.00	0
(60,65]	2	1.09	1
(65,70]	5	1.09	1
(70,75]	1	0.00	0

Table presents the calculated Case Fatality Rate (CFR) for each bin, number of subjects in the bin and number deaths in the bin.

Appendix C Non-parametric RDD results

Tables C.1 & C.2 lists the results non-parametric RD run with the three used kernels (K). In the table VCE = variance-covariance matrix of the estimator, n = number of observation at each side of the cutoff, BW = statistically estimated bandwidth, Coef = effect size, Method = conventional & robust statistics, p >|z| [95% C.I] = statistical significance and 95% confidence interval. Kernels are Epanechnikov, triangular and uniform.

Table C.1: **Results from non-parametric Regression Discontinuity Design (RDD) with different kernels and VCE's, HUS data.**

K	VCE	n	BW est. (h)	Coef	Method	p >|z| [95% C.I]
Epan	HC0	37/51	16.209/16.209	-0.029	Convent	0.004 [-0.048 , -0.009]
					Robust	0.571 [-0.055 , 0.100]
Epan	HC1	37/51	16.300/16.300	-0.039	Convent	0.004 [-0.066 , -0.013]
					Robust	0.571 [-0.069 , 0.095]
Epan	HC2	37/51	16.323/16.323	-0.042	Convent	0.003 [-0.070 , -0.014]
					Robust	0.826 [-0.072 , 0.091]

Continued on next page

Table C.1 continued from previous page

Epan	HC3	37/51	16.388/16.388	-0.048	Convent	0.003 [-0.081 , -0.016]
					Robust	0.988 [-0.084 , 0.085]
Tri	HC0	37/39	15.486/15.486	0.000	Convent	NaN [0.000 , 0.000]
					Robust	0.101 [-0.009 , 0.101]
Tri	HC1	37/39	15.579/15.579	0.000	Convent	NaN [0.000 , 0.000]
					Robust	0.122 [-0.012 , 0.102]
Tri	HC2	37/39	15.606/15.606	0.000	Convent	NaN [0.000 , 0.000]
					Robust	0.126 [-0.012 , 0.101]
Tri	HC3	37/39	15.693/15.693	0.000	Convent	NaN [0.000 , 0.000]
					Robust	0.153 [-0.016 , 0.101]
Uni	HC0	37/51	16.299/16.299	-0.278	Convent	0.000 [-0.380 , -0.176]
					Robust	0.013 [-0.304 , -0.036]
Uni	HC1	37/51	16.325/16.325	-0.278	Convent	0.000 [-0.382 , -0.175]
					Robust	0.015 [-0.307 , -0.032]
Uni	HC2	37/65	26.241/26.241	-0.377	Convent	0.000 [-0.519 , -0.236]
					Robust	0.001 [-0.479 , -0.127]
Uni	HC3	37/51	18.306/18.306	-0.278	Convent	0.000 [-0.386 , -0.171]
					Robust	0.000 [-0.595 , -0.167]
Uni	NN	37/39	15.684/15.684	0.000	Convent	NaN [0.000 , 0.000]
					Robust	0.000 [-0.023 , -0.023]

Table C.2: **Results from non-parametric Regression Discontinuity Design (RDD) with different kernels and VCE's, algorithmic data.**

K	VCE	n	BW est. (h)	Coef	Method	p >|z| [95% C.I]
Epan	HC0	47/81	60.300/60.300	0.239	Convent	0.010 [0.057 , 0.421]
					Robust	0.003 [0.079 , 0.400]
Epan	HC1	47/81	60.300/60.300	0.239	Convent	0.011 [0.055 , 0.423]
					Robust	0.004 [0.076 , 0.403]
Epan	HC2	47/81	60.300/60.300	0.239	Convent	0.011 [0.055 , 0.424]
					Robust	0.004 [0.077 , 0.401]
Epan	HC3	47/81	60.300/60.300	0.239	Convent	0.012 [0.052 , 0.427]
					Robust	0.004 [0.076 , 0.403]
Epan	NN	8/4	3.081/3.081	-0.149	Convent	0.645 [-0.781 , 0.484]
					Robust	0.212 [-1.097 , 0.243]
Tri	HC0	47/81	60.300/60.300	0.131	Convent	0.108 [-0.029 , 0.290]
					Robust	0.044 [0.003 , 0.258]
Tri	HC1	47/81	60.300/60.300	0.131	Convent	0.113 [-0.031 , 0.292]
					Robust	0.048 [0.001 , 0.261]
Tri	HC2	47/81	60.300/60.300	0.131	Convent	0.114 [-0.031 , 0.293]
					Robust	0.046 [0.002 , 0.260]
Tri	HC3	47/81	60.300/60.300	0.131	Convent	0.120 [-0.034 , 0.296]
					Robust	0.048 [0.001 , 0.261]
Tri	NN	47/81	60.300/60.300	0.131	Convent	0.000 [0.093 , 0.169]
					Robust	0.000 [0.093 , 0.169]

Continued on next page

Table C.2 continued from previous page

Uni	HC0	47/82	60.300/60.300	0.450	Convent	0.000 [0.240 , 0.660]
					Robust	0.000 [0.206 , 0.695]
Uni	HC1	47/82	60.300/60.300	0.450	Convent	0.000 [0.237 , 0.663]
					Robust	0.000 [0.201 , 0.699]
Uni	HC2	47/82	60.300/60.300	0.450	Convent	0.000 [0.237 , 0.663]
					Robust	0.000 [0.203 , 0.698]
Uni	HC3	47/82	60.300/60.300	0.450	Convent	0.000 [0.234 , 0.667]
					Robust	0.000 [0.200 , 0.701]
Uni	NN	47/82	60.300/60.300	0.450	Convent	0.000 [0.392 , 0.508]
					Robust	0.000 [0.392 , 0.508]

Appendix D Description of files in GitLab

The project in GitLab `https://gitlab.utu.fi/jojuket/thesiskettunen` holds the codes to replicate the masters thesis project in R. The patient analysis results and the descriptive statistics are under NDA, thus the GitLab project folder doesn't contain any analysis results or descriptive statistics. The code files are structured as following:

- dataRetrieving: fetching the data through API and saving it locally
- dataHandling: cleaning the data and combining additional information from xls
- dataAnalysis: creating descriptive statistics of the data
- survAnalysis: survival analysis calculation, done only for pathologist data
- rdd_for_ki67: regression discontinuity design for pathologist data, includes CFR's
- ai_rdd_for_ki67: regression discontinuity design for algorithmic data, includes CFR's

Appendix E Ethical consent – HBP20200138

Hakemus näytteiden saamiseksi Helsingin Biopankista " Tekoälyyn perustava kuva-analysointi syöpädiagnostiikan kliinisenä apuvälineenä. Image analysis based on artificial intelligence as a clinical tool for cancer diagnostics. Aiforia- AI for Image analysis"-nimistä tutkimusta varten (HBP20200138)

Perustelut

Kyseessä on lääketieteellinen tutkimus, jonka tavoitteena on tutkia, voidaanko tekoälyyn perustuvalla kuva-analyysillä helpottaa kliinisen syöpäpatologian työnkulkuun ja kliiniseen tulkintaan liittyviä rajoitteita. Aluksi tutkitaan, voiko tekoälyyn perustuvalla mallilla vähentää patologien välistä variaatiota eturauhassyövän Gleason luokituksessa ja tarkemmin määrittää immunohistokemiallisen värjäyksen voimakkuutta rinta-ja keuhkosyöpänäytteillä. Tämän lisäksi tutkitaan, voidaanko hematoxylin/eosin (H&E) perusvärjäyksistä ennustaa suoraan biomarkkeritason ilmentymiä ilman työläitä ja kalliita immunohistokemiallisia värjäyksiä. Lopuksi tutkitaan, voidaanko kehitettyjen tekoälymallien avulla paremmin ennustaa potilaan syövän uusimista, siihen liittyvää lääkehoitoa ja/tai potilaan ennustetta.

Tutkimuksessa hyödynnetään tutkittavien Helsingin Biopankkiin siirrettyjä ja biopankkisuostumuksen perusteella Helsingin Biopankkiin talletettuja patologian kudosnäytteitä ja niistä saatuja kuvia.

Näytteisiin yhdistetään tutkittavien kliinisiä tietoja hakemuksen mukaisesti.

Hakemus on käsitelty ja arvioitu Helsingin Biopankin aineistoluovutusprosessin mukaisesti. HUS:in eettinen toimikunta II on antanut tutkimussuunnitelmasta 10.2.2021 puoltavan lausunnon (HUS/415/2021 §24). Hakemus täyttää biopankkitutkimukselle biopankkilaissa (688/2012) asetetut vaatimukset.

Tutkimuksesta vastaavana henkilönä toimii PhD Juuso Juhila, Aiforia Technologies Oy.

Päätös Näytteet ja niihin liittyvät tiedot luovutetaan tutkimussuunnitelman ja hakemuksen edellyttämässä laajuudessa, kuitenkin vain siinä määrin, kun näytteiden ja tietojen saatavuus sen sallii. Aineistojen saatavuuden arvioi Helsingin Biopankki tapauskohtaisesti. Näytteiden ja tietojen luovutuksesta ja käsittelystä laaditaan aina kirjallinen sopimus.

Näytteiden ja tietojen mahdollisesta siirrosta EUn ulkopuolelle laadittava erillinen kirjallinen sopimus GDPR:n mukaisesti.

Lupa näytteiden ja tietojen käsittelyyn myönnetään vuoden 2025 loppuun asti. Patologian aineiston laina-ajassa on huomioitava patologisten arkistojen omat laina-ajat.

Sovelletut oikeusohjeet

Biopankkilaki (688/2012)
Laki viranomaisten toiminnan julkisuudesta (621/1999)
Henkilötietolaki (523/1999)
Laki potilaan asemasta ja oikeuksista (785/1992)

Päätösvallan perusteet

Yhtymähallinnon toimintaohje

Päättäjä

Eero Punkka
HUS-kuntayhtymä biopankin johtaja
Allekirjoitettu koneellisesti

Liitteet

Oikaisuvaatimusohje

Lisätietoja

Theresa Knopp
etunimi.sukunimi@hus.fi

Tiedoksi

Tutkija Juuso Juhlia, juuso.juhila@aiforia.com

Tiedoksianto

Sähköpostiviestillä 22.2.2021

Oikaisuvaatimusohje

Oikaisuvaatimuksen saa tehdä se, johon päätös on kohdistettu tai jonka oikeuteen, velvollisuuteen tai etuun päätös välittömästi vaikuttaa (asianosainen). Kuntayhtymän viranomaisen päätöksestä oikaisuvaatimuksen saa tehdä myös kuntayhtymän jäsenkunta ja sen jäsen. Oikaisuvaatimus tehdään kirjallisena.

Oikaisuvaatimuskielto

Oikaisuvaatimusta ei saa tehdä päätöksestä, joka koskee vain valmistelua tai täytäntöönpanoa, oikaisuvaatimuksen johdosta annetusta päätöksestä eikä päätöksestä, johon haetaan muutosta muun lain kuin kuntalain (410/2015) nojalla.

Oikaisuvaatimusviranomainen

HUSin hallituksen alaisten viranhaltijoiden päätöksistä oikaisuvaatimus osoitetaan hallitukselle.
HYKS-sairaanhoitoalueen ja muiden sairaanhoitoalueiden lautakuntien alaisten viranhaltijoiden päätöksistä oikaisuvaatimus osoitetaan asianomaiselle lautakunnalle.

Oikaisuvaatimusaika

Oikaisuvaatimus on tehtävä 14 päivän kuluessa päätöksen tiedoksisaannista. Asianosaisen katsotaan saaneen päätöksestä tiedon, jollei muuta näytetä, seitsemän päivän kuluttua kirjeen lähettämisestä, kolmantena päivänä sähköisen viestin lähettämisestä, saantitodistuksen osoittamana aikana tai erilliseen tiedoksisaantitodistukseen merkittynä aikana. Kunnan jäsenen katsotaan saaneen päätöksestä tiedon seitsemän päivän kuluttua siitä, kun pöytäkirja on nähtävänä yleisessä tietoverkossa.

Tiedoksisaantipäivää ei lueta oikaisuvaatimusaikaan. Jos oikaisuvaatimusajan viimeinen päivä on pyhäpäivä, itsenäisyyspäivä, vapunpäivä, joulu- tai juhannusaatto taikka

arkilauantai, saa oikaisuvaatimuksen toimittaa perille ensimmäisenä arkipäivänä sen jälkeen.

Oikaisuvaatimuksen sisältö

Oikaisuvaatimuksessa on ilmoitettava
- päätös, johon vaaditaan oikaisua,
- miltä kohdin päätökseen vaaditaan oikaisua ja mitä muutoksia siihen vaaditaan tehtäväksi,
- oikaisuvaatimuksen perustelut,
- mihin oikaisuvaatimusoikeus perustuu, ellei oikaisuvaatimuksen kohteena oleva päätös kohdistu sen tekijään,
- oikaisuvaatimuksen tekijän nimi, kotikunta ja yhteystiedot,
- mahdollisen asiamiehen tai laillisen edustajan yhteystiedot sekä
- postiosoite ja mahdollinen muu osoite, johon asiaan liittyvät asiakirjat voidaan lähettää.

Viranhaltijapäätöksen liitteistä voi tiedustella päätöksessä mainitulta lisätietojen antajalta.
Tiedon luovuttamiseen salassa pidettävistä asiakirjoista sovelletaan viranomaisten toiminnan julkisuudesta annetun lain (621/1999) säännöksiä.

Oikaisuvaatimuksen liitteet

Oikaisuvaatimukseen on liitettävä:
- oikaisuvaatimuksen kohteena oleva päätös oikaisuvaatimusohjeineen,
- selvitys siitä, milloin oikaisuvaatimuksen tekijä on saanut päätöksen tiedoksi, tai muu selvitys oikaisuvaatimusajan alkamisajankohdasta sekä
- asiakirjat, joihin oikaisuvaatimuksen tekijä vetoaa vaatimuksensa tueksi, ellei niitä ole jo aikaisemmin toimitettu viranomaiselle.
-

Oikaisuvaatimuksen perille toimittaminen

Asianosaisen tai hänen valtuuttamansa henkilön on toimitettava
- HUSin hallitukselle osoitetut oikaisuvaatimukset HUS keskuskirjaamoon
- sairaanhoitoalueiden lautakunnille osoitetut oikaisuvaatimukset asianomaisen sairaanhoitoalueen kirjaamoon
-

Oikaisuvaatimus on jätettävä niin ajoissa, että se ehtii perille oikaisuvaatimusajan viimeisenä päivänä ennen kirjaamon asiakaspalveluajan päättymistä. Omalla vastuulla oikaisuvaatimuksen voi lähettää postitse, lähetin välityksellä tai faksilla taikka sähköpostilla.

Yhteystiedot

HUS Keskuskirjaamo, HYKS-sairaanhoitoalueen kirjaamo
Osoite: PL 200, 00029 HUS
Käyntiosoite: Marjaniementie 74, Iiris-keskus, 00930 Helsinki
Puhelinvaihde: 09 4711
Puhelin: 050 428 7837
Faksi: 09 471 75500
Sähköposti: keskuskirjaamo@hus.fi
Asiakaspalvelu arkisin klo 9.00–15.00

Hyvinkään ja Porvoon sairaanhoitoalueiden kirjaamo
Osoite: Sairaalankatu 1, 05850 Hyvinkää
Puhelinvaihde: 019 458 71
Puhelin: 050 568 5518
Faksi: 019 458 72357
Sähköposti: kirjaamo.hyvinkaa@hus.fi
Asiakaspalvelu arkisin klo 9.00 -15.00

Lohjan sairaanhoitoalueen kirjaamo, Lohjan sairaala
Osoite: PL 1010, 00029 HUS
Käyntiosoite: Sairaalatie 8, 08200 Lohja
Puhelin: 040 847 6733
Faksi: 019 388 468
Sähköposti: kirjaamo.lohja@hus.fi
Asiakaspalvelu arkisin klo 9.00-15.00

Länsi-Uudenmaan sairaanhoitoalueen kirjaamo, Raaseporin sairaala
Osoite: PL 1020, 10601 Tammisaari
Käyntiosoite: Itäinen Rantakatu 9, Tammisaari
Puhelin: 019 224 2401
Faksi: 019 224 2299
Sähköposti: registratur.vns@hus.fi
Asiakaspalvelu arkisin klo 9.00-15.00