Turing and van Gogh walk into a bar

Guus Berkelmans

Turing and van Gogh walk into a bar

A computational approach to suicide research

Guus Berkelmans

Invitation

To the public defense of my dissertation

Turing and Van Gogh walk into a bar

Date and Time

24-10-2023 at 11:45

Location

Vrije Universiteit
Amsterdam
De Boelelaan 1105

Contact

guus.berkelmans@gmail.com

VRIJE UNIVERSITEIT

Turing and Van Gogh walk into a bar

A computational approach to suicide research

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op dinsdag 24 oktober 2023 om 11.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Guus Arend Berkelmans

geboren te Amstelveen

promotoren:	prof.dr. R.D. van der Mei
	prof.dr. S. Bhulai
copromotoren:	dr. R. Gilissen
	dr. L.Schweren
promotiecommissie:	prof.dr.ir. F.H. van der Meulen
	prof.dr. B.W.J.H. Penninx
	prof.dr. A. Hemert
	prof.dr. F.E. Scheepers
	prof.dr. M. Liem

Turing and Van Gogh walk into a bar

A computational approach to suicide research

Dissertation Committee

promotors: prof.dr. R.D. van der Mei
Centrum Wiskunde & Informatica, Amsterdam, the Netherlands
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
prof. dr. Sandjai Bhulai
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

copromotors: dr. R. Gilissen
113 Zelfmoordpreventie, Amsterdam, the Netherlands
dr. L. Schweren
113 Zelfmoordpreventie, Amsterdam, the Netherlands

committee: prof.dr.ir. F.H. van der Meulen
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
prof.dr. B.W.J.H. Penninx
Vrije Universiteit Medisch Centrum, Amsterdam, the Netherlands
prof.dr. A. Hemert
Leids Universitair Medisch Centrum, Amsterdam, the Netherlands
prof.dr. F.E. Scheepers
Universitair Medisch Centrum Utrecht, Utrecht, the Netherlands
prof.dr. M. Liem
Universiteit Leiden, Leiden, the Netherlands

ISBN: 978-94-6473-230-6
DOI: http://doi.org/10.5463/thesis.344

Typeset by LaTeX.
Printed by: Ipskamp Printing
Cover design by: Stable Diffusion

Acknowledgments

This thesis has dominated the major part of the past 5 years of my life. Like many things in life, it couldn't have been done without the support of a great many people. So many, in fact that it will be impossible to name everyone by name without making this the longest section in the entire thesis. So if you miss your name here, and I can't stress this enough, it is not personal.

I would like to begin with thanking the absolutely wonderful people at 113 suicide prevention. You provided me with an extremely inspiring place to work on my thesis, and it has been wonderful to see the organisation grow during my time there. The research department was small when I started (around 8 people) but has tripled in size since then. Saskia, Elias, Elke, Kim, Josine, Milou, Jeff, Nikki, Margot, Daniël, Marijn and all the others, it has truly been a delight to work together all these years. I will never forget our wonderful times, both at the office and beyond (and since I am sticking around I look forward to many more). But it wasn't just the research department I would like to thank, also everyone else, including (but definitely not limited to) Rob, Robin, Mattias, Maarten, Bilal, Henrike, Solange, Miriam, Rob, Lilian, Iris, Eline, Manon, Sabine, Evita, and many more.

But 113 wasn't the only organisation I was linked to, since my

project was in fact the collaboration between two parties. The second of which is the National Research Institute for Maths and Computer Science (CWI). I have been privileged to be part of the Stochastics department for the better part of the last 5 years. Since this group has known many members throughout the years I'll limit myself to the ones I had most overlap with (but don't worry, I love all of you, also there might be some recency bias), Robin, Jan, Etienne, Rebekka, Joris S., Jesse, Arwin, Ruurd, Elisabeth, Berend, Marie-Colette, Tim and many more. We had some truly great times and it has been a true delight.

Finally, no PhD project can ever be complete without the support of a university. In my case I had the chance to be part of the Analytics & Optimization group at the Vrije Universiteit Amsterdam. Our Thursdays were always a great place to see what other people were up to, bounce ideas, and most importantly chat nonsense over a cup of coffee. I will always cherish these. It has been my pleasure to work with Yura, Anni, Jesper, Corné, Ger, Jaap, April, Erica, Jeroen, Joost, Jasper, Rik, and many more. And I look forward to trek the alps (or other mountains) with you for many years to come.

My project was not done in isolation. In fact there were 2 PhD positions on my project. Where I worked with population data, the other part of the project focused more on text data from the chatlines. This second position was filled initially by Emil Rijcken, and subsequently by Salim Salmi. They are both wonderful human beings, Emil one of the tallest people I know and unsurprisingly a former basketball player, and Salim one of the most artistic ones I know (I look forward to the cover of your thesis which promises to be something special). It has been a joy working with the both of you.

Speaking of not doing things in isolation, I would like to extend special thanks to Joris Pries, initially my room-mate at CWI, and then my neighbour. Our walls were so thin, I would inevitably get drawn into the numerous discussions that were had in their office (some work related, some less so). It was therefore almost inevitable that one of these discussions would lead to a collaboration. Lo and

behold, the second part of this thesis was born. Working with you on these papers was a true pleasure, and I am still amazed at the TeX-skills on display. That was some truly black magic.

Now we have come to the personal side of things. I would like to start of by thanking Jonna, Jelle, Ivo, and Gijs. Since a school-trip to Rome in the penultimate year of school, we have been a very tight friend group. Whether it was just grabbing a beer in a pub (or during the lockdown years in someone's backyard), or going to Gijs' choir performance (with a surprise cameo), I will treasure our times together forever, and look forward to many more.

I would also like to thank the Woodstock crew, with which I have gone to the middle of nowhere in Poland on more than one occasion to drink beers in a tent camp in the forest, swim in a river, listen to Polish bands none of us had ever heard of, missing the one concert we actually came for, and on one occasion watching two of my friends get engaged. Milan, Gaudi, Marten, Ids, Alique, Patryk and all the others, we should keep this tradition alive. Even though we concluded we might be 'getting too old for this', if there is anything we have learned from movies and television, it just means we will keep doing it.

I cannot omit the France/Belgium crew with which I have gone on multiple holidays, to (quelle surprise) France and Belgium (to a spooky haunted house). Jimmy, Mathilde, Jochem, Hector, Tijmen, Niels, and others. It was a true delight to get out of it all, though on one occasion I did have to run around a French mountain looking for 4G to manage a little crisis. I hope we can do something again soon!

My friends from the UK cannot be left out either. Whether it was the virtual hangs during covid, the Halloween shitty movie night over zoom, the time some of you came down to visit, or Tom's wedding, it always brought me much joy. So thanks to John, Tom, Ruairidh, Ben, Jovan, Sophie, and all the others. And hopefully see you sometime soon.

We are nearing the end, but still have some important people to thank. My day-to-day supervisors (and co-promotors) Renske

Gilissen and Lizanne Schweren. I am incredibly grateful for all that you have done in making this thesis much better than it would have been without all of your wonderful feedback. And it need not be said, but I will do it anyway: it was wonderful working with the both of you.

Penultimately, I would like to thank my supervisors Rob van der Mei and Sandjai Bhulai. Without you this project would have never seen the light of day. I am incredibly grateful for all the support the both of you have given me.

Finally, I would like to thank my family for all the support during these years. Our fortnightly game nights, the discussions about dad's research, the impromptu visits from my mother who was babysitting my sister's kids and decided to pop by with the kids 'since it was only a couple of streets away and she was a bit bored'. They were all a great help, and I couldn't have done it without you.

Contents

1 Introduction

1.1 Motivation behind the thesis

1.1.1 Suicide is a major public health issue

Approximately 2,000 people a day, that is the current estimate of the number of suicide victims worldwide. It accounts for at least 1% of all deaths [165]. It kills more people than malaria, meningitis, alcohol/drug use, falls, drowning, traffic accidents, war (since 1946 [131]), or interpersonal violence [132]. For every person who dies due to suicide, approximately 135 people are affected [45]. This makes suicide a serious public health concern.

Every suicide is a tragic event, not only in the life lost but also in the effect it has on those remaining behind. It is therefore of the utmost importance to reduce the number of suicides. To do this, it helps to have a solid understanding of which people die by suicide, why they do it, and how we can prevent this. This is where the field of suicide (prevention) research comes in.

1.1.2 Current state of the field of suicide research

The field of suicide research is relatively young, but luckily it is growing rapidly. A search on Pubmed for the word suicide, resulted in 879 results up to and including the year 1960, 6,854 results in the years 1961-1980, 23,093 results in the years 1981-2000, and 74,552 results in the years 2001-2020. To illustrate (with some back of the envelope calculations, and hand waving, and unreasonable assumptions) how rapid this growth roughly is: if current trends continue, by the year 2100, there will be a paper on suicide for every 1,000 people on earth, and by the year 2300, there will be approximately 100 papers on suicide for every person on earth.

Research has been done into all three research avenues named above: who dies by suicide, why it happens, and how it can be prevented. Since this thesis focuses on the question of who dies by suicide we will discuss that research avenue last.

On the question of why people die by suicide, several psychological models have been proposed to describe the journey from not being suicidal to suicide attempts or death by suicide (e.g., [82, 105, 113]). In addition, they have studied past victims of suicide through a methodology called a psychological autopsy [14]. This is a framework in which the period leading up to the suicide and the problems that may have contributed to the suicide are analysed in detail. This is done by performing interviews with those left behind. Not only with close relatives, but peers and people in a supervisory role as well, leading to more insights.

Other studies have included ecological momentary assessments (EMAs), where the behaviour of suicidal thoughts is studied throughout a period of time (e.g., [138]) to discover patterns in said behaviour. Another approach used has been modelling the suicidal symptoms as a network (e.g., [25] to see how symptoms influence each other.

On the preventive side of things, many interventions with an effect have been found. For example, preventing access to means has been shown to reduce suicides using said method without transitioning

to other methods [167, 168]. Additionally, school-based awareness programmes, media guidelines, and proper aftercare have been shown to reduce suicides [99, 168]. Also, the training of gatekeepers has been shown to help [33]. These are people who regularly come into contact with high risk groups, can recognise signs of suicidality, are able to start a conversation, and get people the professional help they need.

On the question of who dies by suicide, there have been many studies looking into risk factors for suicide. Many such risk factors have been found, including but not limited to previous suicide attempts or self injury, exposure to suicide or self injury in others, social factors, psychoses, physical illnesses, cognitive problems, demographic factors, and biological factors [59].

1.1.3 Opportunities for improvement

Although some great strides have been made, on the quantitative side of suicide research some serious general limitations remain [32]. First and foremost, due to the relative rarity of the event, most study populations do not contain sufficient suicidal study subjects, unless actively recruited for. Therefore, they tend to be recruited from high risk groups, which are generally not representative of the population as a whole. Second, still due to the rarity of the event, it is often necessary to measure proxies such as suicidal ideation or having suicidal thoughts. Third, for ethical reasons it is often necessary to offer treatment options to people at the highest risk of suicide. This is the only moral thing to do, but the fact you actively influence your study population does have the consequence that the results of said study are less representative.

In the Netherlands, various institutions have databases containing information on the Dutch population, such as, for example, the municipalities, the tax service, the ministry of education, and health insurance companies. Most of these institutions are required by law to provide information to Statistics Netherlands (CBS), a government institution that under tight regulations is allowed to analyse this data, and provide restricted access to it for research purposes [143].

This results in a large database filled with information on, for example, people's age, sex, current residence, education status, employment status, income, social benefits, marital status, household status, healthcare costs, and prescribed medications. Additionally, and crucially, forensic pathologists are required to communicate causes of death directly to CBS.

This wealth of untapped knowledge provides many an opportunity for large scale, robust, and unbiased research. It allows one to ask and answer the important question of 'Who dies by suicide?' in a more robust, representative, and therefore generalisable manner. However, the knowledge of how to effectively gain insights from this data is something that is not necessarily widespread among the field of suicide prevention. On the other hand, the practical experience and knowledge required to interpret and act on the insights gained from the data are not necessarily widespread among data scientists, computer scientists, or mathematicians.

1.1.4 Added value of interdisciplinary research

This is where interdisciplinary research plays a key role.

Some promising studies have been performed applying methods from machine learning and mathematics to the field of suicide research [88]. However, though the predictive power of these models may have improved, the interpretability of these models has suffered. This would not be much of a problem if our goal was to identify suicidal people and (crucially) we were actually able to do so. However the predictive power of these models has not nearly improved enough to identify these people with any reasonable degree of certainty. And even if we could, most of these models cannot be applied in practice due to the detailed level of information it would require on an individual level.

Therefore, what we are interested in are models focusing on finding specific risk groups instead. What has been lacking are insightful models which are more detailed than conventional methodology allows, yet are designed in such a way that the results of these models lead to insights that might be useful in practice. This is a

gap that this thesis aims to fill.

There are generally three levels of possible interventions. The first level concerns universal interventions. These types of intervention are targeted at the whole population, and as such do not require an answer to the question of 'who?'. The second level concerns selective interventions. These types of interventions are targeted at high risk groups. The third level concerns indicated interventions. These interventions are targeted at specific individuals.

For the first type of intervention, no answer to the question of 'who?' is required. To help with the third type of intervention, we would need near to real-time prediction models with a very high degree of accuracy. These are as of yet infeasible. So the second level remains, that of targeting high-risk groups.

In this thesis we will dive into how machine learning can aid suicide research in finding these high-risk groups, so that selective interventions might be deployed as effectively as possible. Due to the data we have access to, it will be limited to suicides of Dutch inhabitants, but the results will be reasonably transferable. Additionally, though the data is region-bound the same does not hold for the developed methodologies.

1.2 From application to theory and back to application

The thesis consists of two major parts. In the first part we will use a conventional method, namely logistic regression, to find risk groups from the data of Statistics Netherlands (CBS [44]), and propose an extension to gain information about possible risk groups corresponding to interactions of risk factors which produce an especially high risk. This extension of logistic regression showed that a methodological gap existed. Though there were ways to measure the information gained from single variables, no clear way existed to measure information gained from combinations of variables. Additionally, due to the relative rarity of the event of suicide most model assumptions that rely on normal distributions are not necessarily

satisfied.

This leads to the second part of this thesis in which we introduce a framework which measures these kinds of dependencies, and show that it satisfies certain theoretical requirements. Additionally, it makes no assumptions on the structure of the underlying distributions or the nature of the dependency. We use this notion of dependency to propose a measure of feature importance, a concept which is very important to the field of data science and machine learning. We also apply the dependency measure to the domain of suicide research alongside conventional methodology and compare the performance.

We will give a short overview of what each chapter is about. Both of the major parts can be read separately. Each chapter is based on a research article and thus designed to be able to be read as a stand alone chapter as much as possible. However, Chapter 2 to Chapter 4 follow a gradual progression, and Chapter 7 builds upon the concepts introduced in Chapter 6. It is therefore recommended to read the chapters within each part in order. Additionally, Chapter 8 is a short chapter that is not based on a paper, and instead serves to introduce an extension of the main concept introduced in Chapter 6, which is then tested on the data from Chapter 5.

1.3 Part 1: Risk groups

1.3.1 Risk factors among youths

In July 2018 it was announced by CBS that in 2017 there was a sharp increase in suicides among youths up to 20 years of age. In previous years there were quite consistently around 50 suicides in this age group, 58 in 2013, 55 in 2014, 48 in 2015, and 48 in 2016. However, in 2017 there were 81 suicides. This was the primary motivation behind Chapter 2 in which we investigated which youths up to 23 died by suicide and compared these group differences with the population in general. We considered suicides in the period 2013-2017, and took sex, age, different regions, migration background, and place in household into account. We also looked at when the

suicide occurred (day of the week and month) and by which method it occurred.

1.3.2 Identifying risk factors using population data on the level of individuals

Most studies into risk factors looked at risk factors in isolation, or only corrected for sex and age which allows for the possibility of proxy effects. This makes it hard to draw conclusions from the data.

Additionally, data on such things as social benefits, income, and healthcare costs is usually not available. Because we have access to the rich database of CBS we have all this data. This allows us to consider a broader range of features in Chapter 3 and run a logistic regression model to account for internal proxy effects. The results both confirmed, earlier international results, but also showed where the priorities should lie for suicide prevention in the Netherlands specifically, including but not limited to males, those of middle age, those living alone, and people on benefits.

1.3.3 Interactions of risk factors

The main limitation of a logistic regression model is the inherent assumption that the relative effect of a risk factor is the same regardless of the presence of other risk factors. It is possible to counter this assumption in part by considering interaction terms (i.e., 'male' and 'is on unemployment benefits'). However, then the next problem crops up: which interaction terms should one add to the model? Sometimes, such decisions are made based on preconceived notions such as 'I think this risk factor differs across the different sexes'. The problem with this approach, is that if the interactions that matter are not yet known, you will not find them either. In other words, you already need to think you know that which you want to know. There are other methods to add interactions, however these scale badly when the number of features becomes large. Additionally, these are usually limited to interactions with two components (i.e., of the form 'living alone'

and 'low income'), not allowing higher level interactions such as for example 'aged between 40 and 54 years old' and 'on unfit for work benefits' and 'with high healthcare costs'.

In Chapter 4 we try to circumvent this problem by introducing a novel heuristic algorithm. This (hypothesis-free) approach incrementally grows the logistic regression model with interaction terms based on a subset of the data, thus obtaining interaction terms we might not necessarily have considered to matter. Finally we estimate this logistic regression model on another disjoint part of the data to find results that can be interpreted in the conventional manner, thus sacrificing nothing in terms of interpretability compared to the case where we knew the interactions in advance. This allowed us to both find high risk groups, and also to find risk groups we would not have necessarily found otherwise, including but not limited to widowed males and people between the ages of 25 and 39 with a low level of education.

1.3.4 Medications

In Chapter 5 we investigate associations between medication prescriptions and suicide risk, whilst taking individuals age, sex, and mental healthcare usage into account. The scale on which we were able to do this was unprecedented. We find that there are strong associations between medication prescriptions and suicide risk in general, and very strong associations for a number of specific medication classes.

1.4 Part 2: Dependency and Feature Importance

1.4.1 Dependency

One of the main problems that we ran into, was the lack of a clear quantification of how much two observed variables depended on each other. A thorough review of existing literature showed that none yet existed. A great number of demands were hoisted upon

what such a quantification should satisfy, and none of the existing proposals satisfied all of them, with most failing even the most basic requirements. In Chapter 6 we list all the previously made requirements, and by combining or generalising them, retaining eight core properties such a quantification should have. We show none of the ones we have found satisfy all of these requirements. We propose our own quantification, the Berkelmans-Pries Dependency (or BP Dependency), and show it satisfies all of the eight core properties.

1.4.2 Feature Importance

In Chapter 7 we combine the fields of probability and game theory. We use the BP dependency defined above, combined with concepts from cooperative game theory, to assign a score of the importance of a feature in prediction. We then list a number of properties that we feel a feature importance score should have, and prove that ours satisfies all of them. We then test using synthetic data, for the 468 existing feature importance scores included, whether the properties hold in that case. We show that of the 18 tests devised, the best existing feature importance method passes 11, thus showing our method outperforms existing methods by a wide margin.

1.4.3 From theory back to practice

When considering the BP dependency, we note that it is possible to add information, whilst retaining a constant dependency. This is due to the fact that it measures 'distance from independence'. However, this does mean that we cannot draw the conclusion that something is conditionally independent if the dependency does not change when it is added. To be able to draw conclusions on conditional dependence, we need to extend our notion of dependency. In Chapter 8 we propose such an extension for the discrete setting. We show that it has a number of nice theoretical properties which can be seen as conditional versions of the core properties of Chapter 6. We then look at its performance on the real live data set on medication prescription and suicide also used in Chapter 5. We compare this performance to the conventional methodology used in that chapter.

Both methods run reasonably fast, and appear to complement one another nicely with the conventional method performing better when its assumptions are satisfied, and the conditional BP dependency performing better when they were violated.

Both approaches resulted in a list of medications which were associated with a higher risk of suicide. Though most of the medications which resulted in the highest risks were associated with psychiatric applications, this does not hold for all of them.

1.5 Key message

The main take away from this thesis will depend on who you are. If you are someone working in suicide prevention that focuses on interventions targeting risk groups, the take-home message should be the various risk groups found in Chapters 2 to 5. If you are someone in suicide prevention sitting on a load of data you do not know what to do with, the take-home message should be that there are plenty of people in the machine learning, data science, and mathematics fields who would love to apply their skills to something worthwhile. If you are a clinician your take-home message should be the medications found in Chapter 5. And for a more general audience the take away is that wonderful results can come out of projects combining expertise from various different fields.

1.6 Publications

This thesis is based on the following publications:

- Chapter 2 is based on [20]: G. Berkelmans, R.D. van der Mei, S. Mérelle, R. Gilissen. 'Demographic risk factors for suicide among youths in The Netherlands'. In: International Journal of Environmental Research and Public Health 17.4 (2020), pages 1-11.
- Chapter 3 is based on [19]: G. Berkelmans, R.D. van der Mei, S. Bhulai, R. Gilissen. 'Identifying socio-demographic risk

factors for suicide using data on an individual level'. In: BMC Public Health 21.1 (2021), pages 1-8.

- Chapter 4 is based on [17]: G. Berkelmans, L.J. Schweren, S. Bhulai, R.D. van der Mei, R. Gilissen. 'Identifying populations at ultra-high risk of suicide using a novel machine learning method'. To appear in Comprehensive Psychiatry 123 (2023), page 152380.
- Chapter 5 is based on [18]: G. Berkelmans, L.J. Schweren, S. Bhulai, R.D. van der Mei, R. Gilissen, A. Beekman. 'On the relation between medication prescriptions and suicide'. Submitted for publication.
- Chapter 6 is based on [16]: G. Berkelmans, J. Pries, R.D. van der Mei, S. Bhulai. 'The BP Dependency Function: a Generic Measure of Dependence between Random Variables'. To appear in Journal of Applied Probability 60.4 (2023).
- Chapter 7 is based on [121]: J. Pries, G. Berkelmans, S. Bhulai, R.D. van der Mei. 'The Berkelmans-Pries Feature Importance Method: a Generic Measure of Informativeness of Features'. Submitted for publication.

1.7 About the title and cover of the thesis

The title of this thesis was born from a simple question: 'Which concepts or people best express the connection between the computational side of this thesis and the suicide research side?'. The answer that was reached was the two individuals mentioned in the title: Alan Turing and Vincent van Gogh.

Alan Turing was the mathematician who cracked Enigma by building one of the very first computers. He is widely considered to be the most influential figure in computing. Sadly, in 1954 he passed away due to suicide at the young age of 42. Vincent van Gogh was a very influential painter, though unfortunately not during his lifetime. Regrettably, he too passed away due to suicide, way back in 1890

at the even younger age of 37. His connection to the computational side is at first glance not as obvious as Turing's. However, where Turing was influential on the theoretical side of machine learning, van Gogh's role provided the other requirement: data.

Van Gogh produced a large collection of paintings throughout his life, and had a very distinctive style. This allowed machine learning researchers to use his paintings to develop new pattern recognition models. One might, for example, consider the models underlying the mobile phone apps that allow you to modify photos so they are in the style of a certain painter. Or one might consider the current generation of text to image generators. These models allow someone to go from a description of a scene or subject, along with certain modifiers, to an image. An example of this last type of model can be seen on the front of this thesis which used the modifier 'painted by Vincent van Gogh'.

Part I

Risk Groups

2 Demographic risk factors for suicide among youths in The Netherlands

2.1 Introduction

Suicide is the number one cause of death among youths from the age of 10 till the age of 30 in the Netherlands. In July 2018, CBS announced that the number of suicides among youths from age 10 up to (not including) 20 had risen to 81 in 2017. In previous years, the number had always been around 50 and below 60: in 2013 there were 58, in 2014 there were 55, and in each of 2015 and 2016 there were 48 suicides among youths from 10 up to 20.

A number of risk factors have been identified that lead to youth suicidal behaviour, such as previous suicide attempts, feeling hopeless or depressed, alcohol abuse, social isolation and others [10, 13, 23, 130, 142]. However, most of these risk factors are psychological and behavioural in nature and thus require a more in-depth look at the individual, and even then might be hard to observe. In addition

Based on [20]: G. Berkelmans, R.D. van der Mei, S. Mérelle, R. Gilissen. 'Demographic risk factors for suicide among youths in The Netherlands'. In: International Journal of Environmental Research and Public Health 17.4 (2020), pages 1-11.

these risk factors are in part derived from psychological autopsy studies where recall bias and a small sample size limit the results [23]. It would be useful to know more about the risk of suicide from less in-depth, easier to observe, and more accurately measurable factors such as socio-demographic characteristics. A substantial number of studies into demographic characteristics of suicidal behaviour have been done. However, these generally had limited non-random samples and yielded limited results [77]. Also only a few looked at the demographic characteristics of young suicide victims, and most of these were focused on the United States [54, 63]. To get a better understanding of socio-demographic risk factors, we look at suicides among all the youths from 10 up till 23 (not including 23) in the Netherlands. The rationale for selecting this age group is that the Dutch government considers this the youth population for policy purposes. Because we included all the Dutch youths, we have a large data-set without selection bias. We separated out the suicides in the period 2013-2017 by gender, age, region, immigration background and place in household and compared them to the corresponding sub-populations of the general population between 10 and 23. Our second aim is to give insight in possible differences in demographic risk factors between youth suicides and suicides in the entire population (including the youths under 23).This could hopefully allow us to find sub-populations among youth suicides that would allow for targeted interventions among youth that would complement interventions targeted among general sub-populations of all ages. A third aim is to see whether there are months with a significantly higher amount of suicides. This could indicate temporal clustering effects and be cause for a further qualitative study.

2.2 Methodology

The data used was micro-data of CBS [143]. This data contains information on all inhabitants of the Netherlands (among others: dates of birth, municipality they live in (and thus province and Public Health Service region (GGD)), type of household, their role in said household, immigration background, social welfare, and in

case of death they include cause of death, date of death, and more) on a yearly basis from various sources which are required to provide this information by law.

Due to the privacy sensitive nature of the data, it is not freely accessible or the data itself allowed to be published. Access has to be granted by the CBS on project to project basis (which was granted for this project) and it is only possible to work with the data via remote connection to their secure servers and any output checked on whether it satisfies the privacy regulations before it is released for publication.

From the data-set those individuals who died by suicide in the years 2013-2017 were extracted on the basis of their cause of death as established by coroners of the GGD (ICD10 codes for external causes: intentional self-harm (X60-X84)) [58]. The coroner is contacted when a person dies and there is any doubt as to whether they died of natural causes. The coroner is always contacted when the deceased is underage (in the Netherlands this means younger than 18 years old). Since the cause of death is provided both privately and anonymously to CBS there is no cause for concern over discrepancies between what the coroner believes the cause of death to be and that which is reported to CBS.

For the reference population (for relative suicide rates and significance checks) we looked at the population at the end of 2017 and included only inhabitants who were listed in the GBA (Municipal Personal Records Database), who were at that time 10 years or older (a minimum age standard used by CBS) and who were at that time registered as being a part of a household (all inhabitants of the Netherlands are in both databases and removed upon death or emigration, but occasionally records are not removed from one of the databases due to an administrative error).

For immigration background we use the classification used by CBS. Being of Dutch descent means having both Dutch parents. If exactly one of the parents is an immigrant, we say the youth has an immigration background corresponding to the country of origin of said parent. If both parents are immigrants, we consider only the country of origin of the mother. Lastly, if the youth is an immigrant them-

selves we say they have an immigration background corresponding to the country of origin. Countries classified as Western are countries from Europe (Turkey excluded), North-America, Oceania, and the countries Indonesia and Japan. Countries classified as non-Western are countries from South-America, Africa, and Asia (Indonesia and Japan excluded) and additionally Turkey.

For tests of significant differences between sub-populations we used the chi-square test of homogeneity with a significance level of 0.05. We compared the frequencies of the sub-population within the suicide victims to the frequencies of the sub-population within the corresponding reference population. In the case where significant differences were found to be present we subsequently looked at normalised residuals and used thresholds of -2 for significantly lower and 2 for significantly higher. We did not correct for multiple comparisons since this is not desirable in an explorative study [7].

2.3 Results

Disclaimer

Due to privacy concerns, numbers strictly lower than 10 could not be reported. In addition, to prevent those numbers to be able to be deduced from the remaining numbers, some other numbers also had to be hidden. All hidden numbers have been replaced by * in the tables. They are still taken into account when doing tests of significance, however chi-squared values, residuals and p-values have not been reported since it might be possible to deduce some of the hidden numbers from these values.

2.3.1 Gender

From the data, we observe that yearly among youths under 23 roughly 1.5 to 2 times as many males than females died by suicide in the period 2013-2017 (Table 2.1), 331 male youths and 170 female youths, with the number of males varying more than the number of females. When compared to the entire population (Table 2.2), we observe that this ratio is even higher: males consistently died

by suicide more than twice as often as females with 6421 male suicide victims and 2956 female suicide victims during the entire period.

Table 2.1: Number of male and female suicides among Dutch youths under 23 years old in the years 2013 to 2017.

Year	Male	Female	Total
2013	73	38	111
2014	56	39	95
2015	65	30	95
2016	59	34	93
2017	78	39	117

Table 2.2: Number of male and female suicides among the entire Dutch population in the years 2013 to 2017.

Year	Male	Female	Total
2013	1308	549	1857
2014	1250	589	1839
2015	1280	591	1871
2016	1279	614	1893
2017	1304	613	1917

2.3.2 Age

Looking at the age of the suicide victims under 23 (Table 2.3), we observe that older youths are more likely to die by suicide with the number of suicides increasing until we get to 19 years old with 77 suicides, 73 suicides at 20 years old, 76 suicides at 21 years old and 74 suicides at 22 years old during the period 2013-2017. There was no statistically significant difference in the number of suicides among youths under 23 in the years in the study period.

Table 2.3: Number of suicides by age of youths under 23 in the period 2013 to 2017.

Age	Number of suicides
10-13	19
14	17
15	24
16	37
17	50
18	64
19	77
20	73
21	76
22	74

2.3.3 Province and healthcare regions

Looking at provinces (Table 2.4), we see substantial differences with the highest provincial suicide rates among youths, Groningen and Noord-Brabant at 5.47 and 5.15 per 100,000 youths per year respectively, being more than twice that of the lowest, Zuid-Holland with 2.50 per 100,000 per year. The provinces Groningen, Noord-Brabant and Gelderland had significantly higher suicide rates among youths than the rest of the country, whereas Zuid-Holland had significantly lower suicide rates among youths than the rest of the country. When looking at the whole population, the provinces Groningen (13.92 per 100,000), Noord-Brabant (12.45 per 100,000), Friesland (13.18 per 100,000), Drenthe (12.76 per 100,000) and Limburg (11.97 per 100,000) have significantly high suicide rates while Overijssel (9.89 per 100,000), Utrecht (9.36 per 100,000) and Zuid-Holland (9.34 per 100,000) have significantly low suicide rates.

Table 2.4: Number of suicide victims under 23 in the period 2013-2017 by province (RS = Relative Suicide Rate per 100,000 per year).

Province	Suicides youths (N)	RS Youths	Suicides whole pop. (N)	RS whole pop.
Netherlands	511	3.86	9377	12.27
Groningen	25	5.47	346	13.92
Friesland	16	3.21	396	13.18
Drenthe	14	3.60	314	12.76
Overijssel	28	2.83	569	9.89
Flevoland	14	3.64	199	9.67
Gelderland	78	4.51	1143	11.10
Utrecht	32	2.93	606	9.36
Noord-Holland	76	3.32	1491	10.54
Zuid-Holland	78	2.50	1745	9.34
Zeeland	*	*	227	11.88
Noord-Brabant	106	5.15	1547	12.45
Limburg	31	3.61	669	11.97

Among so-called Municipal Health Service Regions (regions where municipalities organise healthcare together, also know as GGD regions) even larger differences can be observed with the lowest observed rate of suicides for youths being 2.14 per 100,000 and the highest 5.73 per 100,000 (Table 2.5). The lowest observed rate for the population as a whole is 8.32 per 100,000 in South Holland South and the highest being Groningen with 13.92 per 100,000. However, this is to be expected due to the fact that we are dealing with more regions and even smaller population sizes, which causes the variability on the relative suicide rates to increase. We see that generally high suicide rates among youths coincide with high suicide rates among the population as a whole. What is interesting to note is that both the suicide rates of youths and the suicide rates of the population as a whole are relatively low in the Municipal Health Service Regions containing the four largest cities of the Netherlands: Amsterdam, Rotterdam, Utrecht and the Hague (collectively known as the 'Randstad'). The GGD regions with significantly high suicide rates among youths are GGD Groningen, Security & Health Region (SHR) Middle Gelderland,

GGD North Holland North, GGD Heart for Brabant and GGD Brabant Southeast and the ones with significantly low suicide rates are GGD Amsterdam, GGD Rotterdam-Rijnmond, and Health & Youth Service (HYS) South Holland South. While the GGD regions with significantly high suicide rates among the whole population are GGD Groningen, GGD Drenthe, GGD West-Brabant, GGD Heart for Brabant, GGD Limburg South and GGD Fryslân and the ones with significantly low suicide rates among the entire population are GGD Region Twente, GGD Region Utrecht, GGD Kennemerland, GGD Hollands-Midden, GGD Rotterdam-Rijnmond, HYS South Holland South, GGD Haaglanden.

Table 2.5: Number of suicide victims under 23 and among the population as a whole in the period 2013-2017 by Municipal Health Service (GGD) region (RS = Relative Suicide rate per 100,000 per year).

GGD Region	Suicides Youths	RS Youths	Suicides Full Pop.	RS Full Pop.
Groningen	25	5.47	346	13.92
Fryslân	16	3.21	396	13.18
Drenthe	14	3.60	314	12.76
IJsselland	14	3.14	274	10.47
Region Twente	14	2.57	295	9.40
North- and East-Gelderland	26	3.93	443	10.79
Middle Gelderland	32	5.46	369	10.77
Gelderland South	20	4.14	331	11.96
Flevoland	14	3.64	199	9.67
Region Utrecht	32	2.93	606	9.36
North Holland North	30	5.73	373	11.39
Kennemerland	13	2.99	239	8.83
Amsterdam	20	2.32	541	10.34
Gooi en Vechtstreek	*	*	152	12.02
Middle Holland	23	3.43	370	9.32
Rotterdam-Rijnmond	23	2.14	626	9.61
South Holland South	*	*	204	8.32
Zeeland	*	*	227	11.88
West-Brabant	25	4.44	456	12.98
Heart for Brabant	46	5.30	669	12.64
Brabant SouthEast	35	5.59	449	11.71
Limburg North	18	4.59	297	11.46
Limburg South	13	2.79	372	12.42
Haaglanden	25	2.59	545	9.97
Zaanstreek/Waterland	*	*	186	11.11

When considering regions it is important to note that suicide rates among in-patients of psychiatric institutions are many times higher than the average suicide rates [161] and these institutions are not spread homogeneously across the country, so high regional suicide rates could be due to the in-patients of said institutions. Also the effect possible suicide clusters might have will also affect the suicide rate heavily (since the number of suicides in most regions are relatively small).

2.3.4 Immigration background

When looking at the immigration background of the youths who died by suicide (Table 2.6), we observe that 75% were of Dutch descent, 10% had a western immigration background and 15% had a non-western immigration background. The suicide ratio among those of a non-western immigration background was significantly lower than the average suicide ratio in the youth population as a whole. However, neither the suicide rate among youths of Dutch descent or the suicide rate among youths with a western immigration background can be shown to be significantly higher than the suicide rate among all youths. When considering the entire population (Table 2.7), we observe that not only is the suicide rate among people with a non-western immigration background significantly lower, the suicide rate among people of Dutch descent and the suicide rate among people with a western migration background are both significantly higher than the population as a whole which is consistent with findings in Belgium [15]. The fact that non-western immigrant youth had lower suicide rates than other youth was consistent with findings from Ontario and Switzerland [135, 157]. And although we only had data on fatal attempts it has been previously reported that young female non-western immigrants were more likely to attempt suicide [156].

Table 2.6: Number of suicides among youths under 23 of Dutch descent, youths with a western migration background and youths with a non-western immigration background.

Year	Dutch	Western	Non-western
2013	89	*	*
2014	70	12	13
2015	68	*	*
2016	70	*	*
2017	86	11	20
Total	383	52	76

Table 2.7: Number of suicides among people of Dutch descent, people with a western migration background and people with a non-western immigration background among the entire population.

Year	Dutch	Western	Non-western
2013	1570	162	114
2014	1507	216	104
2015	1539	206	117
2016	1536	232	112
2017	1561	207	142

2.3.5 Month and day of the week

If we look at how suicides were distributed among youths in the various months in the period 2013-2017, none of the months have a significantly high or low number of suicides (Table 2.8) Among youth in the United states a significant increase was found in the number of suicides in March and April of 2017, which was associated with the release of the Netflix series '13 Reasons Why' [31]. However no such increase was found in suicides among the youth in the Netherlands in those same months.

Table 2.8: Number of suicides a month (M) among youths under 23 in the years (Y) 2013 to 2017.

M \ Y	2013	2014	2015	2016	2017
Jan	10	*	*	10	11
Feb	*	*	*	*	12
Mar	13	13	*	*	*
Apr	13	*	*	*	*
May	11	*	11	*	10
Jun	14	*	*	*	12
Jul	*	*	*	*	13
Aug	*	*	*	*	*
Sep	*	*	12	*	*
Oct	*	*	11	14	11
Nov	*	12	*	11	13
Dec	*	*	*	*	14

There appears to be a difference in suicide rates among youths on the various days of the week with 14% on Sunday, 15% on Monday, 15% on Tuesday, 16% on Wednesday, 16% on Thursday, 13% on Friday, and 11% on Saturday, but this is not significantly different (Table 2.9). This is noteworthy since in Ireland a significant difference was found in which days of the week young people died by suicide [10] which saw suicide concentrated in the period from Saturday till Monday. They theorised that this could be due to increased alcohol consumption in the weekend, however the fact that Dutch youths tend to drink mostly on Friday and Saturday [152] which have the lowest rates of suicide (although not statistically significant) seems to indicate that there is no clear relation between alcohol use and youth suicide in the Netherlands. This is different among the whole population, we do see a significant difference in the suicide rates throughout the days of the week with 13% on Sunday, 17% on Monday, 16% on Tuesday, 15% on Wednesday, 14% on Thursday, 14% on Friday, and 11% on Saturday (Table 2.10). Also note that the lower number of suicides on Saturdays is consistent throughout the examined period. The difference in distribution of the youths and the Dutch population as a whole is not significant,

so it cannot be concluded that there are differences in distribution of weekdays between youths and the population as a whole. The fact Monday shows a significantly higher number of suicides among the whole population is consistent with recent studies in the UK, Australia and Korea [41, 85, 94].

Table 2.9: Number of suicides among the Dutch population under 23 for each day (D) of the week over the years (Y) 2013 to 2017.

D \ Y	2013	2014	2015	2016	2017	Total
Sunday	21	13	14	13	11	72
Monday	19	11	11	15	19	75
Tuesday	14	16	16	16	17	79
Wednesday	16	12	15	20	20	83
Thursday	19	16	13	13	20	81
Friday	*	10	16	*	19	67
Saturday	*	17	10	*	11	54

Table 2.10: Number of suicides among the general Dutch population for each day of the week over the period 2013 to 2017.

D \ Y	2013	2014	2015	2016	2017	Total
Sunday	243	255	236	222	254	1210
Monday	317	288	295	311	353	1564
Tuesday	291	283	303	315	292	1484
Wednesday	278	261	311	280	278	1408
Thursday	265	270	264	269	264	1332
Friday	266	246	260	280	254	1306
Saturday	197	236	202	216	222	1073

2.3.6 Place in household

When considering the place the youths occupy within a household, we observe that youths living with their parents are significantly less likely to die by suicide than youths not living with their parents (Table 2.11) Although they make up over 60% of youth suicides,

they make up an even larger proportion of the youth population as an entirety. Within the group of youths not living with their parents, we observe that youths living on their own are significantly more likely to die by suicide. The group least likely to die by suicide are non-married youths living with their partner who do not have any children.

Table 2.11: Number of suicides among youths under 23 separated out by place in household.

Year	Living with parents	Living Alone	Partner non-married couple without children	Member of institutional household	Other
2013	82	20	*	*	*
2014	64	20	*	*	*
2015	65	19	*	*	*
2016	63	23	*	*	*
2017	83	26	*	*	*
Total	357	108	13	23	10

2.3.7 Method of suicide

Among youths who die by suicide, we see that the most common method of suicide (47%) is strangulation or suffocation (which also includes hanging), followed by jumping or lying in front of a moving object (33%) (Table 2.12) In the general population, strangulation and suffocation is also responsible for 47% of suicide deaths (Table 2.13). However jumping or lying in front of a moving object is responsible for 11% of suicide deaths which is substantially lower than the 33% among youths. We see that 21% of deaths among the general population is due to self-poisoning (this includes drugs, both medicinal and recreational, alcohol, gas, bleach and others), whereas among youths it accounts for 8% of suicide deaths. The disparity between methods is possibly in part due to the fact that adults are more likely to have access to the means required for auto-intoxication. This could also explain the high rates among youths for jumping or lying in front of moving objects since the

rail is relatively easily accessible and does not require any other means. The fact drowning is a more common method of suicide for adults seems to be consistent with a Norway study which found that drowning was mostly used by older women [122].

2.3.8 Limitations

When interpreting the results it is important to note that even though we observe various statistical differences between the various sub-populations obtained from our socio-demographic characteristics, the individual effects of said characteristics are harder to measure due to the heavily correlated nature of the characteristics. The youths under 18, for example, are way more likely to live with their parents than to live on their own compared to the youths older than 18, so it becomes difficult to measure whether or not the suicide rate is higher among youths who live alone due to an isolation factor or due to the fact that these youths are usually the older ones. Similarly, the various geographical regions will have a different demographic makeup thus making it hard to separate out the various effects. We also do not know how the various effects interact and stack. In addition, due to privacy concerns the amount of suicides in some sub-populations could not be reported leading to an incomplete view. However these unreported values were taken into account for tests of significance. Also totals over the entire period 2013-2017 could often be reported so the impact of not being able to report these specific values was limited.

2.4 Discussion

We have managed to obtain unbiased frequencies of suicide in various sub-populations of both youths and the population as a whole. This showed us that there was a higher risk of suicide among older youths, male youths, youths living alone, those of Dutch descent and those living in certain regions (Groningen etc.). The lowest risks are seen among youths who live with their parents, younger youths, female youths and youths living in or around the largest cities in the Netherlands.

The most common causes of death among young suicide victims were strangulation or suffocation and jumping or lying in front of moving objects. There were no significant difference in suicide rates of young people among months or days of the week.

The most common method of suicide among both young suicide victims and adults were strangulation or suffocation. The second most common was jumping or lying in front of moving objects for youths but self-poisoning for adults. We do not see any significant changes in causes of death. However this could be due to the period only being five years as trends might occur slowly over a longer period of time [122]. There were no significant differences in suicide rates of young people among months or days of the week. In the population as a whole however we do see significant differences in days of the week with a peak at Monday and a trough at Saturday.

We found that the main differences between the risk factors of youth and the general population is one of effect size. Males have higher risk than females and this effect is greater in the general population than in the youth population. Similarly the protective factor of being a non-western immigrant is larger in the general population than in the youth population. This suggests that these effects accumulate as one ages, for example through continued exposure to certain expectations or to a certain culture. Sadly this makes focusing on risk groups less effective than it would be for adults.

The results agreed with some results from earlier studies done in other populations in some respects such as immigration background [15, 135, 157], suicide being more common on Mondays [41, 85, 94], and drowning being more common among adults. On the other hand there are also some results that contrast with studies done in other populations such as there not being more youth suicides in the weekend [10] or there not being an increase in suicides in the period surrounding the release of '13 Reasons Why' [31]. This suggests some results might be generalisable to other countries whereas some others are not due to there possibly being cultural elements at play. Thus additional evidence from other countries is advised before trying to generalise these results.

We also found differences in methods, youths die more often due

to jumping or lying in front of moving objects whereas adults die more often to self-intoxication or drowning. Restricting access to means for hanging or strangulation is infeasible unless the individual is restricted to a closed institution. Therefore the best restriction to means for youths would be to focus on hot-spots for train suicides.

In Chapter 3 we look at decorrelating effects, and in Chapter 4 examine the way various effects interact and whether there are combinations of risk factors that are especially dangerous.

Table 2.12: Number of suicides among youths under 23 separated out by method of suicide (percentage of total in year in brackets).

Year	Self-Poisoning	Strangulation or Suffocation	Drowning	Jumping from high place	Jumping or lying in front of moving object	Other
2013	*	49 (44%)	*	*	38 (34%)	*
2014	*	46 (48%)	*	*	29 (31%)	*
2015	*	39 (41%)	*	*	32 (34%)	*
2016	*	42 (45%)	*	*	36 (39%)	*
2017	*	65 (56%)	*	*	35 (30%)	*
Total	41 (8%)	241 (47%)	10 (2%)	36 (7%)	170 (33%)	13 (3%)

Table 2.13: Number of suicides among the general Dutch population separated out by method of suicide (percentage of total in year in brackets).

Year	Self-Poisoning	Strangulation or Suffocation	Drowning	Jumping from high place	Jumping or lying in front of moving object	Other
2013	345 (19%)	926 (50%)	105 (6%)	136 (7%)	201 (11%)	144 (8%)
2014	397 (22%)	877 (48%)	111 (6%)	138 (8%)	188 (10%)	128 (7%)
2015	432 (23%)	859 (46%)	111 (6%)	123 (7%)	212 (11%)	134 (7%)
2016	428 (23%)	864 (46%)	111 (6%)	147 (8%)	220 (12%)	123 (6%)
2017	401 (21%)	909 (47%)	116 (6%)	136 (7%)	219 (11%)	136 (7%)
Total	2003 (21%)	4435 (47%)	554 (6%)	680 (7%)	1040 (11%)	665 (7%)

3 Identifying socio-demographic risk factors for suicide using data on an individual level

3.1 Introduction

Suicide is a complex issue that involves multiple factors. Many researchers have looked into risk factors for suicide. However, much of this research looks at risk factors in isolation, or corrected only for age or gender [12, 21, 48, 59, 116]. As a consequence, risk factors found in these studies could simply be a proxy for other risk factors due to the fact that they are correlated (for example, education level and income). Additionally, many studies are of limited size, and are usually non-representative of the population as a whole due to the way the selection procedure was set up, for example, a clinical setting [59].

Knowing that suicide is rarely related to just one risk factor, this study quantifies the effect of individual characteristics as accurately

Based on [19]: G. Berkelmans, R.D. van der Mei, S. Bhulai, R. Gilissen. 'Identifying socio-demographic risk factors for suicide using data on an individual level'. In: BMC Public Health 21.1 (2021), pages 1-8.

as possible by correcting for correlation of characteristics. Furthermore, this study uses all suicide cases in the Netherlands (around 1,900 suicides are reported every year) and a large randomly selected sample of control cases drawn from the full population. This avoids issues of small sample size and selection bias.

To our knowledge, only Gradus et al. [67] used such an approach before in Denmark. They found sex-specific risk profiles for suicide, focusing their risk profiles mainly on medical data. However, in this paper, we focus on socio-demographic risk factors.

This study decorrelates the effects of the risk factors to obtain odds ratios which take into account the proxy effects to the other risk factors. Moreover, we look across multiple years (2014-2017) and at a large number of socio-demographic factors. In this way, we obtain risk factors that are both robust to inter-correlation as well as to events that raise the risk among a certain sub-population.

3.2 Methodology

The primary aim of the study is to find risk factors for suicide that are robust to inter-correlation. In this way we can be sure that the risk factors are not proxies for the numerous other risk factors that are included in the study. Additionally, a secondary aim is to make sure that we can be sure that the risk factors found are based on a large unbiased sample.

The data used was the micro-data of CBS [143]. We limited ourselves to the period of 2014-2017 since some of the databases for 2018 and later were still undergoing data quality checks. Additionally, some databases had a different format prior to 2014 so did not include all of the characteristics of interest prior to 2014. Therefore, we could not analyse data from before 2014 alongside data from the period 2014-2017 while retaining all characteristics of interest. From the data-set of the years 2014-2017, those individuals who died by suicide were identified based on their cause of death, as established by coroners (ICD-10 codes for external causes: intentional self-harm (X60-X84)). The coroner is contacted when there is doubt as to whether a person died of natural causes. The coroner is always

contacted when the deceased is underage (in the Netherlands, this means younger than 18 years old).

Statistical analysis

The binomial logit model was used (commonly referred to as logistic regression) to decorrelate effects. Socio-demographic characteristics of each inhabitant aged 10 and up on the 31st of December (of 2013, 2014, 2015, or 2016) were categorised. We limited ourselves to ages 10 and up since CBS does not report on suicides among youths under 10 years old, due to it being an extremely rare event.

We then modelled the probability of suicide according to a binomial logit model such that

$$\mathbb{P}(S_n|\vec{x}_n) = \frac{e^{V(\vec{x}_n)}}{1 + e^{V(\vec{x}_n)}},$$

where S_n is the event that individual n dies of suicide in the year following observation of the vector of one-hot encoded characteristics $\vec{x}_n$. And

$$V(\vec{x}_n) = \beta_0 + \vec{x}_n \cdot \vec{\beta},$$

where $\vec{\beta}$ is a vector of parameters.

This results in the odds being

$$O(\vec{x}_n) = \frac{\mathbb{P}(S_n|\vec{x}_n)}{1 - \mathbb{P}(S_n|\vec{x}_n)} = e^{V(\vec{x}_n)}.$$

Then define $\vec{x}_{n,k,i}$ to be the same as $\vec{x_n}$ but with the entry corresponding to characteristic k set to value i. This results in the odds-ratio of characteristic k becoming

$$\frac{O(\vec{x}_{n,k,1})}{O(\vec{x}_{n,k,0})} = e^{V(\vec{x}_{n,k,1}) - V(\vec{x}_{n,k,0})} = e^{\beta_k}.$$

The main advantage of such a model is that proxy effects are corrected for as long as the original effect is also included in the model. Therefore, risk groups that are heavily related with for example, age, gender, or income are corrected for. Though there

is still an underlying assumption that risk factors increase risk independently to a certain degree, this assumption is significantly weaker than if one considered the risk factors in isolation or if corrected for a small number of risk factors.

Estimation was done using the Python package *biogeme* [22]. This package estimates the model parameters using maximum likelihood estimation by gradient descent. It has been proven that in the case of the binomial logit model, this always converges to the optimal model with regards to the training error. This means we do not have to worry about local optima. Additionally, the package provides us with standard errors on the parameter estimation, allowing us to form confidence intervals and do tests of significance. The tests of significance done are t-tests (which show how many standard deviations of the estimator it is distanced from 0).

First, estimation was done on a training set. This training set consisted of both people who died by suicide as well as a group of people who did not die by suicide. The people who died by suicide were included with independent probability 0.8 (ended up being 5,854 cases). The people who did not die due to suicide were included with independent probability 0.01 (ended up being 596,416 cases). Due to the way the sampling was done, all bias introduced is introduced into the β_0 parameter. We, therefore, do not report this parameter. The selection procedure of the training set does not introduce any bias into the other parameters.

Secondly, we generated a test set. This test set contained the remaining suicide cases (1,425 cases). Additionally, it contained cases of people who did not die by suicide. These cases were again included with probability 0.01, in such a way that it contains no cases included in the training set. We then estimated the predicted risk of suicide for this test set. From these predictions, we calculated the sensitivity (the proportion of correctly classified cases among suicide victims) and specificity (the proportion of correctly classified cases among those who did not die due to suicide) for various risk thresholds. We then plotted the sensitivity and specificity against each other. In this way, we obtained the receiver operating characteristics curve (ROC curve). We then calculated the area

under the ROC curve (AUC) to estimate model performance. The AUC is also the probability that a random case of death by suicide gets a higher predicted risk than a random case of someone who does not die due to suicide.

3.3 Results

The parameters we estimated (i.e., the β_j parameters and associated standard errors, t-tests, and odds-ratios) for the binomial logit model are shown in Table 3.1. When we talk about increased risk we are talking about increases to the odds of suicide.

Taking the effect of possible correlating risk factors into account, significant increases in risk in all age groups were observed compared to those aged 10 to 19. We see large increases in particular among people aged between 40 and 49 (OR 5.70, 95% CI [4.57,7.24]), between 50 and 59 (OR 6.69, 95% CI [5.37,8.33]), and between 60 and 69 (OR 4.76, 95% CI [3.82,5.93]).

The fact that males die more often due to suicide than females (OR 2.60, 95% CI [2.46,2.77]) still holds when corrected for other characteristics. Furthermore, having mental health problems (OR 7.69, 95% CI [7.24,8.17]) as well as physical health problems as measured through healthcare costs (up to OR 2.23, 95% CI [2.01,2.46]) are major risk factors. Additionally, living alone (OR 1.75, 95% CI [1.49,2.05]), and all forms of unemployment, especially those that have been found unfit for work (UFW; having an OR of 1.89, 95% CI [1.75,2.05]), increase the risk of suicide.

Looking at protective factors, the analyses show that people with a high level of education have a low risk (OR 0.82, 95% CI [0.74,0.90]). Low-risk people are also those with a non-western immigration background (OR 0.63, 95% CI [0.57,0.69]) and 1st generation immigrants (OR 0.72, 95% CI [0.66,0.78]). Also being married or having children is a protective factor for a couple living together (OR 0.64, CI 95% [0.54.0.75] for a married couple without kids, OR 0.63, 95% CI [0.52,0.77] for a non-married couple with kids). These effects are weaker when the other effect is already present (OR 0.58, 95% CI [0.48,0.69]).

Table 3.1: Results of the logistic regression model.

Categories	Characteristics	Beta Parameters	Standard Errors	t-tests	Odds-ratio	N(%) training set	N(%) suicides training set
Age	10 to 19	0	fixed	fixed	1	79,525(13%)	195(3%)
	20 to 29	0.95	0.10	9.40***	2.58	80,131(13%)	541(9%)
	30 to 39	1.39	0.11	12.52***	4.01	79,243(13%)	677(12%)
	40 to 49	1.74	0.11	15.82***	5.70	97,348(16%)	1,159(20%)
	50 to 59	1.9	0.11	17.27***	6.69	97,423(16%)	1,487(25%)
	60 to 69	1.56	0.11	13.68***	4.76	82,917(14%)	945(16%)
	70 to 79	1.40	0.12	11.57***	4.06	53,098(9%)	533(9%)
	80 or older	1.13	0.13	8.83***	3.10	32,585(5%)	317(5%)
Gender	Female	0	fixed	fixed	1	305,867(51%)	1,887(32%)
	Male	0.96	0.03	30.00***	2.60	296,403(49%)	3,967(68%)
Personal income/year	Less than €10,000	0	fixed	fixed	1	170,265(28%)	963(16%)
	€10,000 to €20,000	-0.12	0.05	-2.42*	0.89	133,646(22%)	1,878(32%)
	€20,000 to €30,000	-0.21	0.06	-3.63***	0.81	96,273(16%)	1,120(19%)
	€30,000 to €40,000	-0.17	0.07	-2.49*	0.85	75,794(13%)	816(14%)
	€40,000 to €50,000	-0.31	0.08	-3.89***	0.73	48,697(8%)	426(7%)
	€50,000 to €75,000	-0.31	0.09	-3.52***	0.74	51,114(8%)	431(7%)
	€75,000 to €100,000	-0.27	0.12	-2.23*	0.76	14,750(2%)	124(2%)
	€100,000 to €150,000	-0.27	0.15	-1.72	0.77	8,003(1%)	66(1%)
	More than €150,000	-0.44	0.22	-1.97*	0.64	3,728(1%)	30(1%)
Household income/year	Less than €20,000	0	fixed	fixed	1	52,404(9%)	1,147(20%)
	€20,000 to €40,000	-0.18	0.05	-3.49***	0.84	129,459(21%)	1,712(29%)
	€40,000 to €60,000	-0.31	0.07	-4.65***	0.74	105,090(17%)	958(16%)
	€60,000 to €80,000	-0.31	0.08	-4.11***	0.73	95,590(16%)	711(12%)
	€80,000 to €100,000	-0.32	0.08	-3.77***	0.73	75,645(13%)	508(9%)
	€100,000 to €150,000	-0.43	0.09	-4.85***	0.65	98,135(16%)	557(10%)
	€150,000 to €200,000	-0.46	0.12	-3.86***	0.63	28,459(5%)	151(3%)
	More than €200,000	-0.26	0.14	-1.88	0.77	17,488(3%)	110(2%)
Household wealth	Less than €-100,000	0.09	0.11	0.84	1.10	13,279(2%)	98(2%)
	€-100,000 to €-80,000	0.45	0.13	3.38***	1.57	6,008(1%)	63(1%)
	€-80,000 to €-60,000	0.09	0.12	0.72	1.09	10,561(2%)	77(1%)
	€-60,000 to €-40,000	-0.00	0.09	-0.01	1.00	18,992(3%)	135(2%)
	€-40,000 to €-20,000	-0.08	0.08	-1.08	0.92	29,613(5%)	204(3%)
	€-20,000 to €0	-0.01	0.05	-0.15	0.99	65,468(11%)	709(12%)
	€0 to €20,000	0	Fixed	Fixed	1	132,799(22%)	1,753(30%)
	€20,000 to €40,000	0.05	0.06	0.84	1.05	38,994(6%)	356(6%)
	€40,000 to €60,000	-0.03	0.08	-0.33	0.98	26,999(4%)	213(4%)
	€60,000 to €80,000	0.06	0.08	0.69	1.06	21,684(4%)	184(3%)
	€80,000 to €100,000	0.12	0.08	1.37	1.12	19,393(3%)	170(3%)

Categories	Characteristics	Beta Parameters	Standard Errors	t-tests	Odds-ratio	N(%) training set	N(%) suicides training set
	€100,000 to €150,000	0.08	0.06	1.31	1.09	43,924(7%)	362(6%)
	€150,000 to €200,000	0.05	0.07	0.72	1.05	36,455(6%)	289(5%)
	More than €200,000	0.25	0.05	5.45***	1.29	138,101(23%)	1,241(21%)
Education level	Unknown	0	fixed	Fixed	1	243,871(40%)	2,622(45%)
	Low	-0.04	0.05	-0.96	0.96	135,166(22%)	1,138(19%)
	Middle	-0.05	0.04	-1.31	0.95	126,604(21%)	1,338(23%)
	High	-0.20	0.05	-4.28***	0.82	96,629(16%)	756(13%)
Immigration background	Dutch	0	Fixed	Fixed	1	479,538(80%)	4,861(83%)
	Western non-Dutch	0.15	0.04	4.23***	1.17	55,507(9%)	618(11%)
	Non-Western	-0.47	0.05	-10.28***	0.63	67,225(11%)	375(6%)
	1st generation immigrant	-0.33	0.04	-8.42***	0.72	66,276(11%)	490(8%)
	2nd generation immigrant	0.02	0.04	0.37	1.02	56,456(9%)	503(9%)
Urbanicity	Less than 10,000 people	0.16	0.15	1.05	1.17	5,125(1%)	52(1%)
	10,000 to 100,000 people	0.10	0.04	2.67**	1.10	259,489(43%)	2,465(42%)
	More than 100,000 people	0	fixed	fixed	1	332,825(55%)	3,285(56%)
	Low address density	-0.00	0.04	-0.04	1.00	175,387(29%)	1,689(29%)
	Medium address density	-0.00	0.04	-0.18	0.99	100,540(17%)	974(17%)
	High address density	0	Fixed	Fixed	1	321,512(53%)	3,139(54%)
Place in household	Kid living at home	0	Fixed	Fixed	1	111,592(19%)	464(8%)
	Living alone	0.56	0.08	7.15***	1.75	114,055(19%)	2,592(44%)
	Part non married couple without kids	-0.01	0.09	-0.14	0.99	41,826(7%)	378(6%)
	Part non married couple with kids	-0.46	0.10	-4.47***	0.63	132,559(22%)	990(17%)
	Part married couple without kids	-0.45	0.08	-5.66***	0.64	32,483(5%)	192(3%)
	Part married couple with kids	-0.55	0.08	-6.82***	0.58	128,513(21%)	768(13%)
	Member institutional household	0.02	0.11	0.18	1.02	10,339(2%)	183(3%)
	Parent of single parent household	-0.10	0.11	-0.95	0.91	22,115(4%)	211(4%)
	Reference person other household	0.16	0.26	0.61	1.17	1,578(0%)	17(0%)
	Other place household	-0.07	0.15	-0.45	0.94	7,210(1%)	59(1%)
Healthcare costs/year	Less than €1000	0	fixed	Fixed	1	407,142(68%)	2,834(48%)
(excl. mental health care)	€1000 to €5,000	0.33	0.03	9.85***	1.39	137,936(23%)	1,816(31%)
	€5,000 to €10,000	0.59	0.05	11.41***	1.80	29,758(5%)	535(9%)
	More than €10,000	0.80	0.05	16.56***	2.23	27,434(5%)	669(11%)
Social benefits	Unfit for work benefits (UFW)	0.64	0.04	14.51***	1.89	25,196(4%)	966(17%)
	Long term unemployment benefits (LTU)	0.23	0.06	3.94***	1.26	22,089(4%)	577(10%)
	Short term unemployment benefits	0.19	0.06	3.33***	1.21	31,920(5%)	406(7%)
	Both UFW and LTU	-0.30	0.20	-1.51	0.74	699(0%)	32(1%)
Province	Groningen	0	Fixed	Fixed	1	17,791(3%)	232(4%)
	Drenthe	-0.06	0.10	-0.63	0.94	17,727(3%)	194(3%)
	Utrecht	-0.21	0.08	-2.52*	0.81	44,708(7%)	388(7%)
	Noord-Holland	-0.23	0.07	-3.13**	0.79	98,542(16%)	902(15%)

Categories	Characteristics	Beta Parameters	Standard Errors	t-tests	Odds-ratio	N(%) training set	N(%) suicides training set
	Zuid-Holland	-0.21	0.07	-2.94**	0.81	127,000(21%)	1,095(19%)
	Noord-Brabant	-0.01	0.07	-0.17	0.99	89,342(15%)	978(17%)
	Limburg	-0.21	0.08	-2.61**	0.81	40,603(7%)	409(7%)
	Overijssel	0.23	0.08	2.72**	1.26	40,315(7%)	354(6%)
	Flevoland	-0.11	0.12	-0.92	0.90	13,721(2%)	116(2%)
	Zeeland	-0.11	0.11	-0.98	0.90	13,449(2%)	141(2%)
	Gelderland	-0.09	0.07	-1.19	0.92	72,524(12%)	725(12%)
	Friesland	0.03	0.09	0.34	1.03	21,717(4%)	268(5%)
Year	2014	0	Fixed	Fixed	1	149,977(25%)	1,406(24%)
	2015	0.10	0.04	2.59**	1.10	150,595(25%)	1,468(25%)
	2016	0.10	0.04	2.62**	1.11	151,830(25%)	1,475(25%)
	2017	0.14	0.04	3.70***	1.15	149,868(25%)	1,505(26%)
Other	Self Employed	-0.05	0.06	-0.83	0.95	34,287(6%)	335(6%)
	Main Earner of Household	-0.06	0.05	-1.29	0.94	310,036(51%)	4,244(72%)
	Has a legal debt repayment plan	-0.30	0.19	-1.58	0.74	1,428(0%)	31(1%)
	In mental health care	2.04	0.03	64.97***	7.69	32,921(5%)	2,134(36%)

Having a higher income is also a protective factor. This holds for both personal income (up to OR 0.64, 95% CI [0.41,1]) as well as household income (up to OR 0.63, 95% CI [0.50,0.80]). Interestingly, household wealth does not appear to be a protective factor. It even increases risk in the wealthiest category (Table 3.1) We observe urbanity and regional differences being mostly non-significant.

3.4 Discussion

To our knowledge, this is the first study done into suicide on socio-demographic factors with such a large and unbiased sample, where, due to the level of detail of the data, analyses could be done to control for many characteristics, giving us very robust risk factors. We found that previously discovered risk factors for suicide (middle-age, male gender, and unemployment (as measured through benefits)) remain elevated even when corrected for a wide array of socio-demographic characteristics. The same holds for commonly found protective factors for suicide, like having a higher income or immigration background.

Most increased risk came from being a recipient of mental health care (which includes being an inpatient as well as being an outpatient), which can be expected knowing that approximately 87% of people who die by suicide have mental health problems [11]. Additionally, physical healthcare being a risk factor could be explained due to hospitalisations for previous suicide attempts. However, due to the fact that the risk keeps increasing as physical health care costs increase, it is unlikely this would account for all of the increased risk.

This study did not observe significant differences between rural and urban municipalities. However, it is important to note that due to the high population density in the Netherlands, most rural areas in the Netherlands might still be considered urban compared to rural areas in other countries. Looking at raw frequencies, we see regional differences in the Netherlands [20]. These differences became much less when the effects of possible correlating risk factors were considered. This seems to indicate that the regional differences

are primarily caused by the differences in the demographic makeup of the regions as opposed to specific local causes.

When we look at level of education, we see that being highly educated remains a protective factor. However, this only holds for the highest level of education and is not particularly protective. Especially when compared to the results of Phillips and Hempstead [118] who found large differences between the suicide rates among people with a high school degree and those among people with a college degree in the United States. Combined with the protective factor of income and the high correlation between level of education and income, this seems to suggest a proxy effect. The level of education might only be a protective factor due to the associated increase in income.

Our model has a reasonable fit with AUC 0.77, which is high for a model predicting suicide death [59] and comparable to the recent results of Zheng et al. [170] who used deep neural networks on electronic health records to predict suicide attempts (AUC 0.769). It could be used to identify low, regular, or high-risk groups. However, the model is not usable to predict suicide risk in individuals. Suicide is a rare event that on average occurs in about 1 in 10,000 people a year. This means that even if you have a tenfold increase in predicted risk, you will still have 1,000 false positives for each true positive.

Although then not useful for prediction on an individual level, the results from this study allow for targeted prevention measures at certain risk groups. For example, it would be possible to train social workers that are in regular contact with recipients of social benefits to be gatekeepers. Alternatively, high risk groups may be specifically targeted to raise awareness of suicide prevention hot-lines within these groups. The authors also recommend that this study is repeated at regular intervals to see whether changes in public policy coincide with changes in risk groups.

The methodology used in this study allowed us to find robust risk and protective factors for suicide. However, with this methodology it is not possible to discover which specific combinations of risk factors or protective factors are especially dangerous or safe. Research has shown that the interactions of risk factors play a substantial role in

suicide prediction and greatly improves model performance [170]. Therefore, having a proper understanding of such interactions will be of great importance in future research. In the next chapter we will describe and estimate a new machine learning model that allows us to find significant interactions in a data-driven and hypothesis-free manner. Since we are doing this in a data driven and hypothesis-free manner, it both limits bias on which interactions to include and allows us to discover interactions that have not even been considered before.

4 Identifying populations at ultra-high risk of suicide using a novel machine learning method

4.1 Introduction

In the Netherlands alone, an average of five people die by suicide each day. Every case of suicide marks a personal tragedy, both for the victim and for those left behind. Therefore, it is of utmost importance to implement effective suicide prevention programmes at multiple levels, including interventions directed at the entire population (e.g., public awareness campaigns), interventions targeting high-risk groups or sub-populations (e.g., training gatekeepers among professionals encountering individuals with financial difficulties) and interventions targeting at-risk individuals (e.g., cognitive behavioural therapy for individuals with suicidal thoughts) [165].

Interventions at the second level, targeting sub-populations, require adequate identification and detection of groups at elevated risk of

Based on [17]: G. Berkelmans, L.J. Schweren, S. Bhulai, R.D. van der Mei, R. Gilissen. 'Identifying populations at ultra-high risk of suicide using a novel machine learning method'. In: Comprehensive Psychiatry 123 (2023), page 152380.

suicide. Multiple studies have been performed to detect risk factors for suicide [12, 21, 48, 59, 116]. Not unexpectedly, the most important predictor of death by suicide is a prior non-fatal suicide attempt or prior psychiatric hospitalisation [59]. Experiencing stressful life events and mental health problems including depression and substance use problems substantially increase the risk for suicide attempts and suicidal ideation, which in turn increases the risk of suicide [59]. In addition, certain socio-demographic groups are at elevated risk, including but not limited to men, people of middle age, people of lower socio-economic status and people living alone [19, 59].

In complex and multi-factorial outcomes such as mental illness, risk factors are known to interact or accumulate. For instance, stressful live events may trigger a depressive episode in persons with a genetic vulnerability to depression [155]. To our knowledge, however, little is known about interacting socio-demographic risk factors for suicide. In a hypothetical example, one might expect that unemployment might increase the risk of suicide more for men living alone than for the rest of the population. The detection of relevant interacting socio-demographic risk factors will allow the identification of more specific sub-populations at elevated risk of suicide. This may increase the efficacy of targeted preventive interventions and has the potential to reduce suicide rates.

Machine learning methods offer new possibilities for flexible, data-driven, hypothesis-free and robust investigation of accumulating risk factors for suicide. A recent study performed such analyses using predominantly healthcare data and succeeded in identifying multiple relevant interactions [67]. Risk of suicide was higher, for instance, in men and women who had recently attempted suicide and were not being treated with pharmacotherapy. In a second study, including over 15,000 features (including but not limited to: demographics, diagnostic codes, procedure codes, and medication prescriptions) in the initial model and retaining 117 of them, researchers were able to develop a risk prediction model with acceptable performance parameters to stratify hospital patients by suicide risk [170].

An important limitation of the above studies is their complexity,

hampering translation of their results to actionable recommendations for clinical practice. Moreover, as Kirtley et al. have recently emphasised [88], current machine learning methods have limited capabilities to support decisions and interventions at the individual level, as false-positive rates as well as false-negative rates are typically high. Thus, there is a need for more actionable and transparent machine-learning models to aid detection of high-risk subgroups rather than individuals.

In this paper, we present a new machine learning model that allows for investigation of complex interactions of socio-demographic risk factors whilst retaining interpretability. This model is applied to predict suicide risk groups in a data-set spanning the entire population of the Netherlands over a period of nine years, thereby mitigating sampling bias and sample size limitations. Our model yields detailed and interpretable results to aid the identification of sub-populations of individuals at relatively high risk for suicide, which may aid targeted preventive interventions.

4.2 Methodology

4.2.1 Data

CBS is a national administrative authority aiming to collect and provide reliable information that advances the understanding of social issues. CBS maintains a high-quality database containing, among others, socio-demographic and medical information regarding every inhabitant of the Netherlands. For privacy reasons, the data can only be accessed via a remote access connection to their computational servers. Prior to releasing results for publication, compliance with privacy laws is ensured.

For the current paper, we included data regarding all inhabitants of the Netherlands on the 31st of December of nine consecutive years (2011 to 2019), adding up to a total of 137,666,515 person years. Of those, 16,417 person years ended by suicide in the year following observation and 137,650,098 person years did not end by suicide in the year following observation.

4.2.2 Features of interest

The following socio-demographic predictor variables were measured on the 31st of December of the year preceding the outcome: sex, age, immigration background, household income, personal income, household wealth/debts, level of education, physical healthcare costs, place in household, marital status, short-term unemployment benefits, long-term unemployment benefits and unfit for work benefits. For details, see Table 4.1. Categorical variables were one-hot-encoded for use in machine learning analyses, meaning that for each category a new variable was introduced which has value 1 if the individual was in said category and has value 0 otherwise. Continuous variables were split into mutually exclusive response categories (e.g., quartiles) and also one-hot-encoded.

4.2.3 Model

A heuristic algorithm was devised to obtain interacting features which provide additional risk of suicide or reduce the risk. The obtained interaction features were prioritised on statistical significance as well as model improvement. The algorithm comprises four steps.

Step 1: the data is divided into three disjoint partitions: a training set, a validation set and a test set. The training set includes fifty percent of person years ending in suicide (N=8,214) and one percent of all other person years (N=1,377,055) and is used to detect significant interactions between features of interest. The validation set includes forty percent of person years ending in suicide (N=6,512) and one percent of all other person years (N=1,377,870) and is used to estimate the final logistic regression model. The test set includes ten percent of person years ending in suicide (N=1,691) and one percent of all other person years (N=1,375,966) and is used to evaluate the performance of the final model.

Step 2: the algorithm identifies significant interactions between features of interest in the training data-set. For details, see Section 4.A.1. In short, the algorithm defines a main-effects logistic regression model including all features listed in Table 1 (hereafter

Table 4.1: Predictor variables or 'features of interest' included in the machine learning model, after sampling (all person years resulting in suicide were included and 3% of the person years not resulting in suicide were included, see model section), (ref) means the reference category.

Features	Response categories	N	%
Sex	Male (ref)	2050131	49.4
	Female	2097177	50.6
Age in years	10-24	835473	20.1
	25-39 (ref)	856591	20.7
	40-54	999010	24.1
	55-69	879303	21.2
	70+	576931	13.9
Immigration background	Dutch (ref)	3231078	77.9
	1st generation western	213524	5.1
	2nd generation western	207883	5.0
	1st generation non-western	314951	7.6
	2nd generation non-western	179868	4.3
Personal income	1st quartile (ref)	1007657	24.3
	2nd quartile	1027422	24.8
	3rd quartile	1016962	24.5
	4th quartile	1015324	24.5
	Unknown	79943	1.9
Household income	1st quartile (ref)	1019868	24.6
	2nd quartile	1016622	24.5
	3rd quartile	1016383	24.5
	4th quartile	1014626	24.5
Household wealth/debts	1st quartile (ref)	1017399	24.5
	2nd quartile	1017837	24.5

Features	Response categories	N	%
	3rd quartile	1016503	24.5
	4th quartile	1015760	24.5
Level of education	Low	892702	21.5
	Middle (ref)	859185	20.7
	High	684749	16.5
	Unknown	1710672	41.3
Physical healthcare costs	€0 (ref)	59793	1.4
	€1- €5000,	3635734	87.7
	€5001-€10000	201167	4.9
	€10001+	183200	4.4
	Unknown	67414	1.6
Place in household	Child living at home	760069	18.3
	Living alone	802714	19.4
	Partner in couple with children	1201518	29.0
	Partner couple without children (ref)	1102279	26.6
	Other	280728	6.8
Marital status	Never married/registered partner (ref)	1714362	41.3
	Married/registered partner	1834896	44.3
	Divorced	348547	8.4
	Widowed	232123	5.6
Unfit for work benefits	Yes	196522	4.7
	No (ref)	3950786	95.3
Short-term unemployment benefits	Yes	215734	5.2
	No (ref)	3931574	94.8
Long-term unemployment benefits	Yes	171810	4.1
	No (ref)	3975498	95.9

referred to as basic features). Next, interaction terms are added in an iterative manner. The algorithm looks at combinations of the form "X and Y", where X is a feature already present in the model, and Y is a basic feature. So the new combination feature "X and Y" would have value 1 if both feature X and feature Y have value 1. For each of these combinations, it calculates the rate at which it would improve the log-likelihood. Then we corrected for sub-population size, since larger sub-populations without an underlying effect on suicide risk will still have a large effect on log-likelihood simply due to variance. The significant interactions that came out of this analysis were listed and for the further analyses we focused on interactions of features which had the largest effects and also included at least 200 suicides. This was done because for suicide prevention interventions the primary interest is in sub-populations with a substantial number of suicides. After this, a check was performed whether this (interaction of) feature(s) truly improved the model. If it did not, it was removed. The process was stopped when the ratio at which removals needed to be performed exceeded 10% and at least 30 interactions were tested.

Step 3: a logistic regression model was estimated on the validation data-set including all significant interactions detected in step two. As the data in the validation set is disjoint from the training set, the notion of over-fitting is removed and regular test statistics such as t-tests and p-values can be interpreted.

Step 4: the following performance statistics were computed on the test set: log-likelihood as an indicator of model fit, and area under the receiver operating characteristics curve (AUC) as an indicator of the model's ability to distinguish between those who died by suicide and those who did not.

4.2.4 Statistics

For each significant feature and each interaction between two or more features, we report the logistic regression model β parameters, odds ratios and corresponding confidence intervals. For interaction terms, we also report the compound odds ratios (COR's) and their confidence intervals, reflecting the summed effect of features when com-

bined (e.g., exp $(\beta_{male} + \beta_{widowed} + \beta_{male\ and\ widowed})$). Also reported are the number of suicides in the corresponding sub-populations for the validation set as well as the relative rate in said sets (per 100,000 inhabitants per year), which are corrected for the sampling procedure (number of suicides is scaled up by a factor of 2.5, and number of non-suicides by a factor of 100).

4.3 Results

4.3.1 Main effects

For a complete list of main effects see Section 4.A.2. Most important risk factors for suicide were middle age ($\beta_{40-54\ vs\ 25-39}$ = 0.48, 95% CI = [0.39, 0.57]; $\beta_{55-69\ vs\ 25-39}$ = 0.37, 95% CI = [0.22, 0.52]), living alone ($\beta_{living\ alone\ vs\ couple\ without\ children}$ = 0.88, 95% CI = [0.77, 0.98]), high healthcare costs ($\beta_{5-10k/year\ vs\ none}$ = 0.87, 95% CI = [0.64, 1.11]; $\beta_{>10k/year\ vs\ none}$ = 1.53, 95% CI = [1.26, 1.80]), being divorced ($\beta_{divorced\ vs\ never\ married}$ = 0.51, 95% CI = [0.39, 0.62]), and receiving benefits ($\beta_{short-term\ unemployment\ vs\ not}$ = 0.19, 95% CI [0.08, 0.30]; $\beta_{long-term\ unemployment\ vs\ not}$ = 0.54, 95% CI [0.42, 0.67]; $\beta_{unfit\ for\ work\ vs\ not}$ = 1.30, 95% CI [1.16, 1.44]). Most important protective factors for suicide were female sex ($\beta_{female\ vs\ male}$ = -0.83, 95% CI = [-0.9, -0.76]), younger age ($\beta_{10-24\ vs\ 25-39}$ = -0.85, 95% CI = [-1, -0.71]), non-western migration background ($\beta_{first\ generation\ non-western\ vs\ Dutch}$ = -1.02, 95% CI = [-1.15, -0.89], $\beta_{second\ generation\ non-western\ vs\ Dutch}$ = -0.53, 95% CI = [-0.70, -0.35]) and higher income (e.g. $\beta_{personal\ income\ in\ 4th\ quartile\ vs\ 1st\ quartile}$ = -0.62, 95% CI = [-0.73, -0.50]). For confidence intervals of the differences between non-reference groups (i.e., 40-54 vs 10-24), see Section 4.A.3. Among the general population there is a suicide rate of 11.8 per 100,000. When considering relative suicide rates among the sub-populations corresponding to the various features, the highest rate among the basic features is among the people who are unfit for work with a suicide rate of 47.0 per 100,000 on the validation set, with the second highest rate being among the long-term unemployed with a suicide rate of 32.1 per 100,000 on the validation set, and the rest of the sub-populations having rates below 30.0 per

100,000.

4.3.2 Interaction effects

Table 4.2 lists all twenty interaction terms included in the final logistic regression model. Of those, seventeen yielded significant effects in the validation data-set ($p < 0.05$). Among the interaction features there are ten sub-populations identified with relative risks higher than 30.0 per 100,000 on the validation set.

Table 4.2: Interaction terms found by the algorithm as tested on the validation set. With corresponding Beta parameters, Odds-Ratios, Compound Odds Ratios, absolute and relative number of suicides within the sub-population within the validation set.

Interaction term	Beta (95% CI)	Odds-Ratio (95% CI)	Compound Odds Ratio (95%CI)	Number of suicides	Relative number of suicides
Aged 25-39 and low level of education	0.46 ([0.30, 0.62])	1.58 (1.35, 1.86])	1.63 ([1.38, 1.93])	259	20.07
Aged 40-54 and long-term unemployment	-0.22 ([-0.41, -0.04])	0.80 ([0.67, 0.96])	2.23 ([1.90, 2.61])	234	35.58
Aged 55-69 and living alone	-0.42 ([-0.67,-0.17])	0.66 ([0.51, 0.84])	2.27 ([1.78, 2.9])	833	35.54
Aged 55-69 and living alone and Dutch immigration background	0.18 ([-0.04, 0.39])	1.20 ([0.96, 1.48])	2.71 ([2.30, 3.19])	728	39.37
Aged 55-69 and living alone and household income in the 1st quartile and never married	-0.21 ([-0,43, 0.01])	0.81 ([0.65, 1.01])	3.44 ([2.60, 4.55])	229	57.22
Aged 55-69 and never married	0.32 ([0.15, 0.5])	1.38 ([1.16, 1.65])	2.00 ([1.64, 2.44])	427	34.81
Aged 55-69 and part of couple without child at home	-0.46 ([-0.63, -0.29])	0.63 ([0.53, 0.75])	0.91 ([0.79, 1.05])	622	9.38
Aged 55-69 and healthcare costs of €10001 or more	-0.44 ([-0.63, -0.25])	0.64 ([0.53, 0.78])	4.30 ([3.16, 5.86])	238	30.70
Aged 70 or older and healthcare costs of €10001 or more	-0.66 ([-0.88, -0.44])	0.52 ([0.41, 0.64])	2.14 ([1.58, 2.9])	175	15.59
Male and unfit for work	-0.39 ([-0.54, -0.24])	0.68 ([0.59, 0.78])	2.48 ([2.21, 2.79])	642	58.56
Male and part of couple with child at home	0.64 ([0.48, 0.8])	1.90 ([1.61, 2.22])	0.82 ([0.73, 0.92])	801	10.94

Interaction term	Beta (95% CI)	Odds-Ratio (95% CI)	Compound Odds Ratio (95%CI)	Number of suicides	Relative number of suicides
Male and widowed	0.54 ([0.33, 0.74])	1.72 ([1.40, 2.09])	1.56 ([1.31, 1.86])	218	31.31
Male and healthcare costs of €10001 or more	-0.30 ([-0.46, -0.14])	0.74 ([0.63, 0.87])	3.42 ([2.64, 4.43])	456	27.48
Never married and unfit for work	-0.03 ([-0.26, 0.19])	0.97 ([0.77, 1.21])	3.54 ([2.77, 4.53])	441	88.48
Never married and unfit for work and physical healthcare costs between €1 and €5000	0.54 ([0.31, 0.78])	1.72 ([1.36, 2.18])	6.45 ([4.83, 8.61])	321	83.01
Never married and household income in the 1st quartile	0.30 ([0.18, 0.43])	1.35 ([1.19, 1.54])	1.35 ([1.19, 1.54])	1438	25.69
Never married and average level of education	0.25 ([0.12, 0.37])	1.28 ([1.13, 1.45])	1.28 ([1.13, 1.45])	871	13.59
Never married and personal income in the 2nd quartile	0.27 ([0.15, 0.4])	1.31 ([1.16, 1.49])	1.04 ([0.93, 1.17])	259	20.07
Unfit for work and personal income in the 2nd quartile	-0.38 ([-0.53, -0.23])	0.68 ([0.59, 0.8])	1.98 ([1.65, 2.38])	234	35.58
Education unknown and physical healthcare costs between €1 and €5000	0.28 ([0.16, 0.41])	1.32 ([1.17,1.51])	1.21 ([0.95, 1.54])	833	35.53

Broadly, three categories of interacting risk factors can be distinguished (with minor crossover): (1) interactions related to age, (2) interactions related to sex, and (3) interactions related to marital status. Two significant interactions did not fit any of these categories.

Interactions involving age: among people of young working age (25-39 years old), but not in the other age groups, a low level of education is an important risk factor for suicide (OR = 1.58 (95% CI OR [1.35,1.86], COR [1.38,1.93])). In contrast, being unemployed is an important risk factor for suicide in the general population but not among people of middle age (40-54 years old; OR = 0.80 (95% CI OR [0.67,0.96], COR [1.90,2.61])). Among those aged between 55-69, having never been married is an important risk factor (OR = 1.38 (95% CI OR [1.16,1.65], COR [1.64,2.44])), while high healthcare costs (OR = 0.64 (95% CI OR [0.53,0.78], COR [3.16,5.86])) and living alone (OR = 0.66 (95% CI OR [0.51,0.84], COR [1.78,2.9])) are less of a risk factor in this age group compared to in other age groups (though they do remain risk factors). High healthcare costs are also less important for persons aged 70 or older (OR = 0.52 (95% CI OR [0.41,0.64], COR [1.58,2.09])).

Interactions involving sex: although being widowed is not a risk factor in general (OR = 0.91 (95% CI OR [0.76,1.10])) it is a major one for males (OR = 1.72 (95% CI OR [1.4,2.09], COR [1.31,1.86])). Being a part of a couple with a child at home is very protective in general (OR = 0.43 (95% CI OR [0.37,0.51])), however this effect is greatly reduced for males (OR = 1.90 (95% CI OR [1.61,2.22], COR [0.73,0.92])) although it does remain a protective factor.

Being on unfit for work benefits is a larger risk factor for females (OR = 3.67 (95% CI OR [3.18,4.23])) than it is for males (OR = 0.68 (95% CI OR [0.59,0.78], COR [2.21,2.79])). Having higher healthcare costs (€10,001 or more) is a larger risk factor for females (OR = 4.62 (95% CI OR [3.54,6.05])) than it is for males (OR = 0.74 (95% CI OR [0.63,0.87], COR [2.64,4.43])).

Interactions involving marital status: although never being married is protective in general, in specific groups it is a risk factor: those unfit for work with low healthcare costs (OR = 1.72 (95% CI

OR [1.36,2.18], COR [4.83,8.61])), those with the 25% lowest household incomes (OR = 1.35 (95% CI OR [1.19,1.54], COR [1.19,1.54])), and those with an average level of education (OR = 1.28 (95% CI OR [1.13,1.45], COR [1.13,1.45])).

Other interactions: finally, there are two interaction features who fit into none of the three major groups. Personal income being in the 2nd quartile is most protective for those who are unfit for work, though not so protective as to completely mitigate the risk associated with being unfit for work (OR = 0.68 (95% CI OR [0.59,0.8], COR [1.65,2.38])). Lastly though education being unknown is a protective factor in general (OR = 0.86 (95% CI OR [0.75,0.98])) this protective effect disappears for those with low healthcare costs (OR = 1.32 (95% CI OR [1.17,1.51], COR [0.95,1.54])).

4.3.3 Model performance

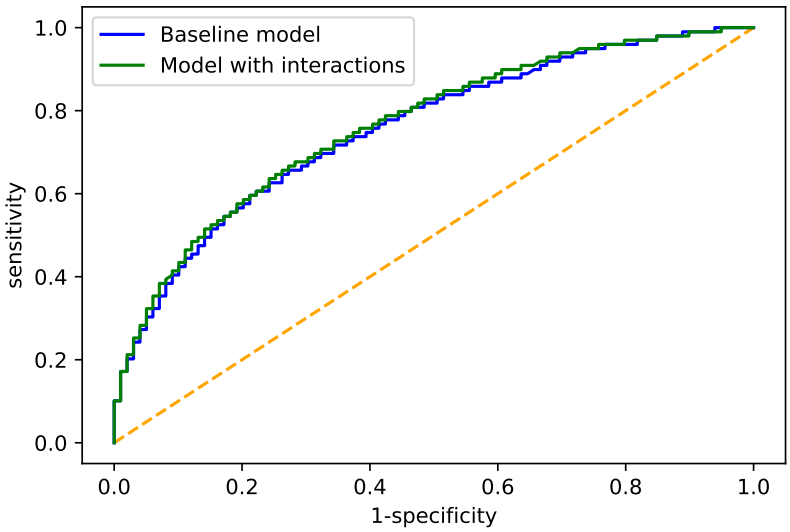
The baseline logistic regression model without interaction terms had a log-likelihood of -12,184.54 and AUC 0.75. In comparison the logistic regression model with interaction terms had a log-likelihood of -12,119.24 and AUC 0.76. See Figure 4.1 for the curves themselves.

4.4 Discussion

Effective suicide prevention programs include, among others, interventions targeting subgroups of people at particularly high-risk of suicide. Here, we designed a heuristic model to detect such subgroups based on interactions between risk factors, and applied it to data covering the entire population of the Netherlands. We identified three sub-populations at ultra-high risk for suicide, with relative suicide rates of 50/100,000 person years or higher. In addition, we identified several factors that when combined increase the risk of suicide, while in isolation they do not increase the risk of suicide. These risk factors would not be detected using traditional prediction models.

We identified three sub-populations at ultra-high risk of suicide,

Figure 4.1: Receiver Operating Characteristics curve for the baseline and the interaction models, sensitivity is the true positive rate while 1-specificity is the false positive rate. The plot shows their values for a range of thresholds.



with social isolation and socio-economic hardship as common denominators. Compared to suicide rates in the general population of the Netherlands (11.8 suicides per 100,000 person years), people who were never married and unfit for work - and among them those with low healthcare costs - were up to 7.4 times more likely to die by suicide (88 suicides per 100,000 person years). Despite the relatively small size of this group in the Dutch population, in 2012-2020 more than 100 suicides (7% of all suicides within that period) occurred in this group each year. The second ultra-high risk group concerns males who are unfit for work, with 59 suicides per 100,000 person years. These findings urge professionals in regular contact with individuals receiving unfit for work benefits, including occupational healthcare professionals, community service providers and municipal workers, to pay particular attention to males and people who were never married. The third ultra-high risk group comprises individuals aged 55-69, who were never married, are living alone and have a

relatively low income, with 57 suicides per 100,000 person years. Further studies, including longitudinal and qualitative studies, are needed to investigate how the combination of these specific risk factors culminates in extreme high-risk profiles.

In addition to the extreme high-risk group, we identified several risk factors that increase the risk of suicide only in the presence of other risk factors. First, while neither young age (25-39 years old) nor lower level of education was found to be a risk factor in itself, together they constituted a major risk profile. Among individuals of young adult age, those with a lower level of education presented with a relative suicide rate more than double that of their peers with a medium or higher level of education (20.1 vs. 8.8 suicides per 100,000 person years). Our data does not provide insights into mechanisms that might underlie the elevated risk of suicide among young adults with lower education. In keeping with our prior observation that socioeconomic hardship may be a common denominator, we speculate that, among many factors, job insecurity might play a role: young adults in the Netherlands, and especially those with lower levels of education, are more likely than other age groups to be offered temporary employment [1]. Job insecurity has been linked to poorer mental health [92], which in turn is linked to a higher suicide risk [21]. To substantiate this hypothesis or find alternative explanations, we recommend research into risk factors for suicide in this group, including socio-economic factors, external stressors, psycho-social circumstances and psychological vulnerabilities.

Second, widowhood did not increase the risk of suicide in the general population in our study, yet it did when combined with the known risk factor male sex. Among widowed males, the suicide rate is more than twice the rate observed in general male population. Previous studies including males only have reported a higher risk of suicide among widowed individuals [27, 129, 166], but to our knowledge the combined risk of widowhood and male gender has not previously been reported. The current study does not allow characterisation of the suicidal process within male widowed individuals. A recent study showed that male widows, compared to female widows, are generally protected from income loss yet are more likely to experience negative

emotional consequences such as loneliness and depression [145]. Our findings underline the need for social support for males who lost their partner, and urge training of gatekeepers among professionals encountering these males.

Finally, we wish to draw the readers attention to two risk factors that each appear in a large number of significant interaction terms: (1) being of middle age (55-69 years old) and (2) having never been married. The large number of significant interactions involving these factors suggests risk profiles within the sub-populations of middle-aged individuals and individuals who were never married that differ from risk profiles in the general population.

Several limitations to our approach should be considered when interpreting our findings. First, death by suicide is a relatively rare event, limiting our statistical power to find associations with risk factors. To achieve reliable model performance, we included all suicides that occurred in the Netherlands between 2012 and 2020. We are unable to assess whether results are stable over time. Second, the model is constructed bottom-up. A top-down approach starting with all possible highest-level interactions might allow detection of more high-risk subgroups, however such approaches are also known to generate more false-positives. Third, adding interaction terms to the model improved model performance only slightly (AUC = 0.76 vs. AUC = 0.75). While the validity of the identification of high-risk groups is not affected (AUC between 0.7 and 0.8 is generally deemed 'acceptable'), it does suggest that even with highly complex statistical modelling predicting death by suicide remains challenging.

Our approach has many strengths. First, since we sampled from the entire population in a controlled manner, we avoid sampling bias. Second, our model is hypothesis-free, allowing identification of previously unidentified risk groups. Third, our model has flexible settings, allowing the user to adjust the trade-off between good model performance and statistically robust results. Finally, and in contrast to existing machine learning methods such as artificial neural networks, our model is open and readily interpretable.

In summary, we performed a heuristic machine learning method

to find interactions. We found disproportionately high suicide rates among people who were never married and received unfit for work benefits, among males who received unfit for work benefits, and among those aged 55-69 who lives alone, were never married and whose household income was low. Additionally, we found high suicide rates among those aged 25-39 with a low level of education and among males who lost their partner. Our findings may have important implications for suicide prevention policies and are generalisable to other (similar) countries.

4.A Appendix

4.A.1 Full explanation of step 2 of the algorithm

In what follows we will outline the full details of every step within Figure 4.2, further splitting the 'Add interaction' step into the sub-steps shown in Figure 4.3.

Start: To start with we specify our hyper-parameters N_{added}, θ, t, and S_{min} whose functions shall be explained as they become relevant. Additionally, we initialise $n_{added} = n_{removed} = 0$ and T as an empty list. These will be updated throughout the procedure.

We define $\vec{x}_i$ for $i \in \{1, 2, \ldots, N\}$ to be our one-hot encoded basic features. We define $\vec{y}_i$ for $i \in \{1, 2, \ldots, L\}$ to be all the features in our model. The number of basic features, N, is fixed. However, since we will be adding features throughout our model, the total number of features, L, will vary.

Split data: We split our training set into two subsets: a *searching* set (80% of cases), and a *control* set (containing the remaining 20%).

Init. model: Using the searching set we estimate an initial logistic regression model specified by

$$\mathbb{P}((\vec{s})_k = 1 | \vec{y_1}, \ldots, \vec{Y_L}) = \frac{e^{V_k}}{1 + e^{V_k}},$$

Figure 4.2: Flowchart that shows the substeps of step 2 of the algorithm.

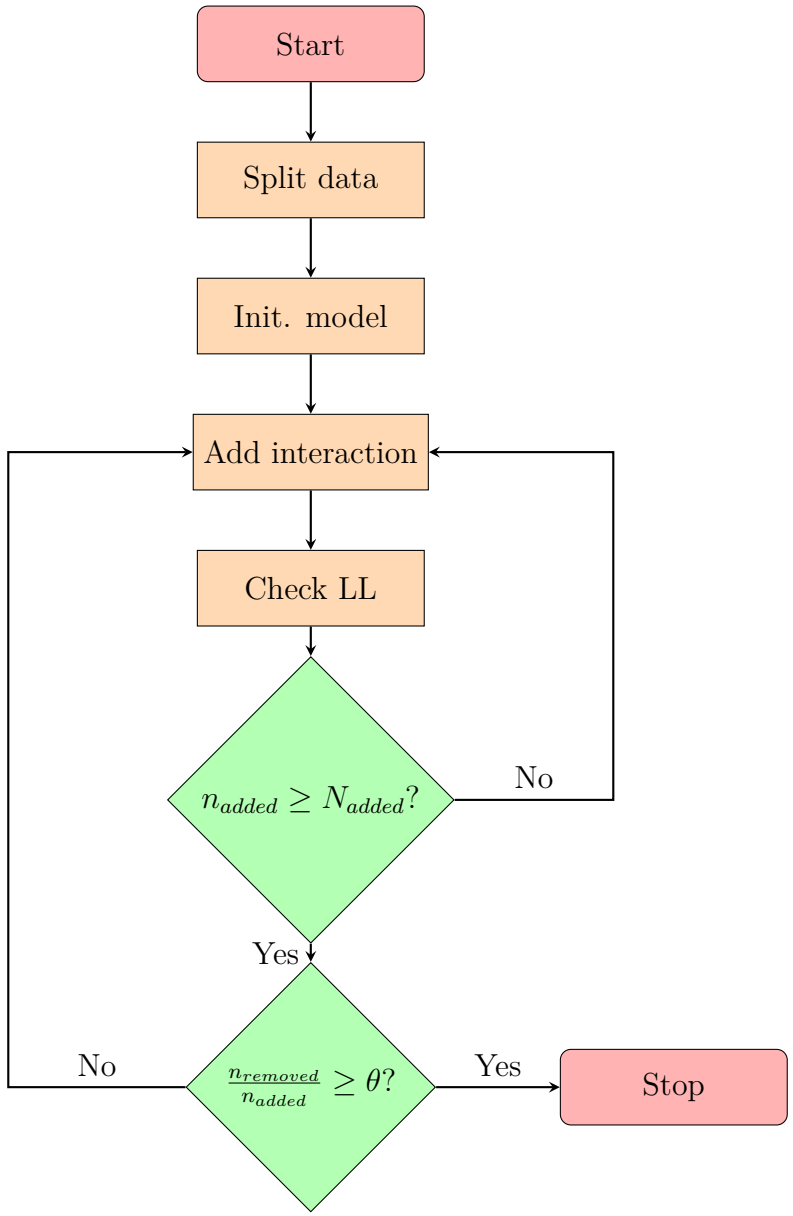
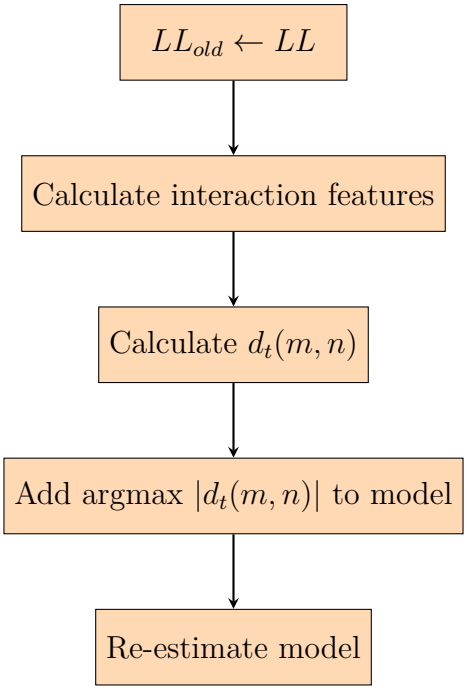


Figure 4.3: Flowchart that shows the substeps of the 'Add interaction' step of the algorithm.



where $\vec{s}$ is the feature corresponding to 'died by suicide' and

$$V_k(\vec{y}_1, \ldots, \vec{y}_L) = \beta_0 + \sum_{i=1}^{L} \beta_i (\vec{y}_i)_k,$$

with the β_i being the parameters to be estimated. Estimation is done through log-likelihood maximization via gradient descent methods. Set LL to be equal to the log-likelihood of the model on the control set.

Add interaction: *$LL_{old} \leftarrow LL$:* We set the value of LL_{old} to the current value of LL.

Calculate interaction features: For each $m \in \{1, \ldots, N\}$ and $n \in \{1, \ldots, L\}$ define $\vec{z}_{m,n} = \vec{x}_m * \vec{y}_n$ where $*$ denotes the element-wise product.

Let $\vec{u}$ be the all ones vector and $N_{\vec{z}_{m,n}} = \langle \vec{z}_{m,n}, \vec{u} \rangle$ be the number of people possessing both characteristic m *and* n. Let $S_{\vec{z}_{m,n}} = \langle \vec{z}_{m,n}, \vec{s} \rangle$ be the number of people possessing both characteristic m *and* n who died by suicide.

Let $s_{\vec{z}_{m,n}} = \mathbb{1}(S_{\vec{z}_{m,n}} \geq S_{min})$. Here S_{min} functions as a lower bound on the number of suicides in the sub-population corresponding to the interaction feature for us to consider it for the model. We used $S_{min} = 200$.

Calculate $d_t(m, n)$*:* Let $LL_{m,n}(\beta_{m,n})$ be the log-likelihood corresponding to the logistic regression model specified as

$$V_k = \beta_0 + \sum_{i=1}^{L} \beta_i (\vec{y}_i)_k + \beta_{m,n} (\vec{z}_{m,n})_k,$$

then

$$\frac{dLL_{m,n}}{\beta_{m,n}} = \sum_{k=1}^{N_p} (\vec{z}_{m,n})_k (s_k - \frac{e^{V_k}}{1 + e^{V_k}}),$$

where N_p is the total number of cases in our searching set. Note that under the assumption that the 'true' value of $\beta_{n,m}$ on the underlying probability process is 0 (i.e., feature $\vec{z}_{m,n}$ is irrelevant) the value of this expression scales to the order of $\sqrt{N_{\vec{z}_{m,n}}}$. Therefore, if we do not correct for this, large values of $|\frac{dLL_{m,n}}{\beta_{m,n}}|$ will simply end up corresponding to large sub-populations. As such we define

$$d_t(m, n) = \frac{1}{N^t_{\vec{z}_{m,n}}} \frac{dLL_{m,n}}{\beta_{m,n}} s_{\vec{z}_{m,n}},$$

where hyper-parameter t describes the trade-off between optimization of the log-likelihood and statistical significance, with a value of 0 completely prioritizing the former, and a value of 0.5 completely prioritizing the latter. We used $t = 0.3$.

Add $\arg\max |d_t(m, n)|$ *to model:* We then select

$$(m^*, n^*) = \arg\max_{m,n} |d_t(m, n)|,$$

and add the corresponding feature to our model by setting $\vec{y}_{L+1} = \vec{z}_{m^*,n^*}$ and set $L \leftarrow L + 1$. We add (m^*, n^*) to the list T. We also set $n_{added} \leftarrow n_{added} + 1$

Re-estimate model: We re-estimate the model with the new feature and set LL to the log-likelihood of this new model on the control set.

Check LL: We check whether or not the performance on the control set has improved by looking at $LL - LL_{old}$. If this is negative we once again remove the added feature from our model and set $n_{removed} \leftarrow n_{removed} + 1$.

$\mathbf{n_{added} \geq N_{added}}$**:** Here N_{added} functions as a minimum number of iterations before stopping. If we have not yet run that many iterations, we return to the 'Add interaction' step. If we have we move on to the next step. We used $N_{added} = 30$.

$\frac{\mathbf{n_{removed}}}{\mathbf{n_{added}}} \geq \theta$**:** Here θ functions as a minimum amount of false positives before terminating. If the proportion of false positives is less than θ we return to the 'Add interaction' step. If it is at least θ we end our algorithm. We used $\theta = 0.1$.

4.A.2 Full results of logistic regression

Table 4.3 gives the full results of our final model including both the basic as well as the interaction features.

4.A.3 Confidence intervals of the differences between non-reference groups

It is interesting to not only know whether or not sub-populations have an increased risk of suicide with respect to a reference sub-population, but also with respect to the other sub-populations. Therefore, confidence intervals for $\beta_A - \beta_B$ for sub-populations corresponding to the same original categorical variable are provided in Tables 4.4 to 4.12.

Table 4.3: Full results logistic regression on validation set including both basic features and interaction terms. With corresponding Beta parameters, Odds-Ratios, Compound Odds Ratios, absolute and relative number of suicides within the sub-population within the validation set as well as the training set. With N(val)=absolute number of suicides within validation set, N(train)=absolute number of suicides within training set, Rel(val)=relative number of suicides within the validation set (corrected for sampling procedure, per 100,000), Rel(train)=relative number of suicides within the training set (corrected for sampling procedure, per 100,000).

Features	β estimates	95% C.I. β	95% C.I. OR	95% C.I. COR	N (val)	Rel (val)	N (train)	Rel (train)
β_0 / Full population	-5.42	[-5.7,-5.13]	[0,0.01]	[0,0.01]	6512	11.7598	8214	11.8591
Male	0.00	[0,0]	[1,1]	[1,1]	4397	16.0445	5565	16.2555
Aged 25-39	0.00	[0,0]	[1,1]	[1,1]	1151	10.0758	1467	10.2593
Dutch Immigration Background	0.00	[0,0]	[1,1]	[1,1]	5378	12.4660	6756	12.5191
Part couple without child at home	0.00	[0,0]	[1,1]	[1,1]	1510	9.4118	1857	9.2573
Personal income in first quartile	0.00	[0,0]	[1,1]	[1,1]	941	6.9806	1228	7.3130
Household income in first quartile	0.00	[0,0]	[1,1]	[1,1]	2784	20.3763	3401	19.9394
Household wealth/debts in first quartile	0.00	[0,0]	[1,1]	[1,1]	1814	13.3128	2176	12.8107
Average level of education	0.00	[0,0]	[1,1]	[1,1]	1448	12.5832	1896	13.2622
No physical healthcare costs	0.00	[0,0]	[1,1]	[1,1]	86	10.8723	123	12.2103
Never married	0.00	[0,0]	[1,1]	[1,1]	2821	12.3053	3508	12.2679
Female	-0.83	[-0.9,-0.76]	[0.41,0.47]	[0.41,0.47]	2115	7.5616	2649	7.5623
Aged 10-24	-0.85	[-1,-0.71]	[0.37,0.49]	[0.37,0.49]	512	4.5826	720	5.1680
Aged 40-54	0.48	[0.39,0.57]	[1.48,1.76]	[1.48,1.76]	1956	15.7218	2403	15.4614
Aged 55-69	0.37	[0.22,0.52]	[1.24,1.68]	[1.24,1.68]	1796	15.3007	2231	15.1825
Aged 70 or older	-0.11	[-0.24,0.03]	[0.79,1.03]	[0.79,1.03]	928	12.0496	1202	12.4417
1st generation western immigration background	-0.21	[-0.33,-0.09]	[0.72,0.92]	[0.72,0.92]	331	11.6500	396	11.0959
1st generation non-western immigration background	-1.02	[-1.15,-0.89]	[0.32,0.41]	[0.32,0.41]	297	7.0322	359	6.8358
2nd generation western immigration background	-0.06	[-0.17,0.06]	[0.84,1.06]	[0.84,1.06]	363	13.0852	493	14.2122
2nd generation non-western immigration background	-0.53	[-0.7,-0.35]	[0.5,0.7]	[0.5,0.7]	143	5.9703	210	6.9805
Child living at home	0.08	[-0.08,0.24]	[0.93,1.27]	[0.93,1.27]	508	4.9926	756	5.9679

Features	β estimates	95% C.I. β	95% C.I. OR	95% C.I. COR	N (val)	Rel (val)	N (train)	Rel (train)
Living alone	0.88	[0.77,0.98]	[2.17,2.66]	[2.17,2.66]	2943	27.4229	3652	27.2016
Part couple with child at home	-0.84	[-1,-0.68]	[0.37,0.51]	[0.37,0.51]	1052	7.1662	1341	7.2863
Other member household	0.14	[0.01,0.27]	[1.01,1.32]	[1.01,1.32]	499	13.3264	608	12.9204
Personal income in the 2nd quartile	-0.23	[-0.35,-0.12]	[0.71,0.89]	[0.71,0.89]	2184	15.9142	2734	15.9101
Personal income in the 3rd quartile	-0.42	[-0.52,-0.32]	[0.6,0.73]	[0.6,0.73]	1847	13.6120	2305	13.5711
Personal income in the 4th quartile	-0.62	[-0.73,-0.5]	[0.48,0.61]	[0.48,0.61]	1407	10.3917	1782	10.5031
Personal income unknown	0.20	[-0.03,0.42]	[0.97,1.53]	[0.97,1.53]	133	12.5132	165	12.3466
Household income in the 2nd quartile	0.00	[-0.1,0.09]	[0.91,1.1]	[0.91,1.1]	1588	11.6848	2057	12.1188
Household income in the 3rd quartile	-0.04	[-0.16,0.07]	[0.86,1.07]	[0.86,1.07]	1142	8.4459	1384	8.1440
Household income in the 4th quartile	-0.20	[-0.32,-0.07]	[0.72,0.94]	[0.72,0.94]	865	6.3886	1207	7.1401
Household net wealth in the 2nd quartile	-0.05	[-0.12,0.02]	[0.89,1.02]	[0.89,1.02]	1848	13.5832	2387	14.0547
Household net wealth in the 3rd quartile	-0.02	[-0.1,0.06]	[0.9,1.06]	[0.9,1.06]	1336	9.8571	1657	9.7626
Household net wealth in the 4th quartile	0.10	[0.02,0.19]	[1.02,1.21]	[1.02,1.21]	1381	10.2071	1829	10.7673
Low level of education	0.03	[-0.09,0.14]	[0.92,1.15]	[0.92,1.15]	1248	10.4582	1478	9.9127
High level of education	0.03	[-0.08,0.14]	[0.92,1.16]	[0.92,1.16]	893	9.8205	1065	9.2930
Level of education unknown	-0.15	[-0.29,-0.02]	[0.75,0.98]	[0.75,0.98]	2923	12.7969	3775	13.2008
Physical healthcare costs between €1 and €5000	0.06	[-0.17,0.28]	[0.84,1.33]	[0.84,1.33]	5053	10.4067	6374	10.5056
Physical healthcare costs between €5001 and €10000	0.87	[0.64,1.11]	[1.89,3.02]	[1.89,3.02]	587	21.8782	727	21.5478
Physical healthcare costs of €10001 or more	1.53	[1.26,1.8]	[3.54,6.05]	[3.54,6.05]	786	23.4980	990	23.5258
Physical healthcare costs unknown	-1.40	[-1.69,-1.11]	[0.18,0.33]	[0.18,0.33]	71	7.9273	78	6.9189
Married or register partnership	0.26	[0.14,0.37]	[1.15,1.45]	[1.15,1.45]	2096	8.5726	2608	8.5051
Divorced	0.51	[0.39,0.62]	[1.48,1.86]	[1.48,1.86]	1155	24.6854	1489	25.5291
Widowed	-0.09	[-0.27,0.09]	[0.76,1.1]	[0.76,1.1]	440	14.2491	609	15.6787
Short-term unemployment	0.19	[0.08,0.3]	[1.09,1.35]	[1.09,1.35]	395	15.8520	503	16.1755
Unfit for work	1.30	[1.16,1.44]	[3.18,4.23]	[3.18,4.23]	1048	46.9534	1262	44.8177
Long-term unemployment	0.54	[0.42,0.67]	[1.52,1.95]	[1.52,1.95]	609	32.0567	746	31.2095
Aged 25-39 and low level of education	0.46	[0.3,0.62]	[1.35,1.86]	[1.38,1.93]	259	20.0663	296	18.2429
Aged 40-54 and long-term unemployment	-0.22	[-0.41,-0.04]	[0.67,0.96]	[1.9,2.61]	234	35.5796	262	31.7326

Features	β estimates	95% C.I. β	95% C.I. OR	95% C.I. COR	N (val)	Rel (val)	N (train)	Rel (train)
Aged 55-69 and living alone	-0.42	[-0.67,-0.17]	[0.51,0.84]	[1.78,2.9]	833	35.5369	1040	35.6329
Aged 55-69 and living alone and Dutch immigration background	0.18	[-0.04,0.39]	[0.96,1.48]	[2.3,3.19]	728	39.3718	892	38.8586
Aged 55-69 and living alone and household income in the 1st quartile and never married	-0.21	[-0.43,0.01]	[0.65,1.01]	[2.6,4.55]	229	57.2214	250	50.4134
Aged 55-69 and never married	0.32	[0.15,0.5]	[1.16,1.65]	[1.64,2.44]	427	34.8185	506	33.1695
Aged 55-69 and part of couple without child at home	-0.46	[-0.63,-0.29]	[0.53,0.75]	[0.79,1.05]	622	9.3768	753	9.0842
Aged 55-69 and healthcare costs of €10001 or more	-0.44	[-0.63,-0.25]	[0.53,0.78]	[3.16,5.86]	238	30.7018	280	29.0080
Aged 70 or older and healthcare costs of €10001 or more	-0.66	[-0.88,-0.44]	[0.41,0.64]	[1.58,2.9]	175	15.5938	260	18.4981
Male and unfit for work	-0.39	[-0.54,-0.24]	[0.59,0.78]	[2.21,2.79]	642	58.5574	764	55.5414
Male and part of couple with child at home	0.64	[0.48,0.8]	[1.61,2.22]	[0.73,0.92]	801	10.9391	979	10.6842
Male and widowed	0.54	[0.33,0.74]	[1.4,2.09]	[1.31,1.86]	218	31.3128	304	34.5278
Male and healthcare costs of €10001 or more	-0.30	[-0.46,-0.14]	[0.63,0.87]	[2.64,4.43]	456	27.4831	596	28.4100
Never married and unfit for work	-0.03	[-0.26,0.19]	[0.77,1.21]	[2.77,4.53]	441	88.4831	495	79.0293
Never married and unfit for work and physical healthcare costs between €1 and €5000	0.54	[0.31,0.78]	[1.36,2.18]	[4.83,8.61]	321	83.0144	362	74.6546
Never married and household income in the 1st quartile	0.30	[0.18,0.43]	[1.19,1.54]	[1.19,1.54]	1438	25.6896	1715	24.6509
Never married and average level of education	0.25	[0.12,0.37]	[1.13,1.45]	[1.13,1.45]	871	13.5912	1144	14.3792
Never married and personal income in the 2nd quartile	0.27	[0.15,0.4]	[1.16,1.49]	[0.93,1.17]	1008	24.7583	1245	24.5072
Unfit for work and personal income in the 2nd quartile	-0.38	[-0.53,-0.23]	[0.59,0.8]	[1.65,2.38]	382	48.5758	470	47.5203
Education unknown and physical healthcare costs between €1 and €5000	0.28	[0.16,0.41]	[1.17,1.51]	[0.95,1.54]	2165	11.5392	2808	11.9722

Table 4.4: Differences of beta parameters of the age groups with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	Age 10-24	Age 25-39	Age 40-54	Age 55-69	Age 70+
Age 10-24	N/A	-0.85 [-1.00,-0.71]*	-1.33 [-1.48,-1.18]*	-1.22 [-1.41,-1.03]*	-0.74 [-0.92,-0.56]*
Age 25-39	0.85 [0.71,1.00]*	N/A	-0.48 [-0.57,-0.39]*	-0.37 [-0.52,-0.22]*	0.11 [-0.03,0.24]
Age 40-54	1.33 [1.18,1.48]*	0.48 [0.39,0.57]*	N/A	0.11 [-0.02,0.24]	0.59 [0.47,0.71]*
Age 55-69	1.22 [1.03,1.41]*	0.37 [0.22,0.52]*	-0.11 [-0.24,0.02]	N/A	0.48 [0.32,0.64]*
Age 70+	0.74 [0.56,0.92]*	-0.11 [-0.24,0.03]	-0.59 [-0.71,-0.47]*	0.48 [0.32,0.64]*	N/A

Table 4.5: Differences of beta parameters of the migration backgrounds with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	Dutch	1st gen Western	1st gen non-Western	2nd gen Western	2nd gen non-Western
Dutch	N/A	0.21 [0.09,0.33]*	1.02 [0.89,1.15]*	0.06 [-0.06,0.17]	0.53 [0.35,0.7]*
1st gen Western	-0.21 [-0.33,-0.09]*	N/A	0.81 [0.65,0.97]*	-0.15 [-0.31,0.01]	0.32 [0.12,0.52]*
1st gen non-Western	-1.02 [-1.15,-0.89]*	-0.81 [-0.97,-0.65]*	N/A	-0.96 [-1.12,-0.80]*	-0.49 [-0.69,-0.29]*
2nd gen Western	-0.06 [-0.17,0.06]	0.15 [-0.01,0.31]	0.96 [0.80,1.12]*	N/A	0.47 [0.27,0.67]*
2nd gen non-Western	-0.53 [-0.7,-0.35]*	-0.32 [-0.52,-0.12]*	0.49 [0.29,0.69]*	-0.47 [-0.67,-0.27]*	N/A

Table 4.6: Differences of beta parameters of place in household with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	Child living at home	Living alone	Partner couple without kids	Partner couple with kids	Other
Child living at home	N/A	-0.80 [-0.94,-0.66]*	0.08 [-0.08,0.24]	0.92 [0.72,1.12]*	-0.06 [-0.22,0.10]
Living alone	0.80 [0.66,0.94]*	N/A	0.88 [0.77,0.98]*	1.72 [1.56,1.88]*	0.74 [0.63,0.85]*
Partner couple without kids	-0.08 [-0.24,0.08]	-0.88 [-0.98,-0.77]*	N/A	0.84 [0.68,1.00]*	-0.14 [-0.27,-0.01]*
Partner couple with kids	-0.92 [-1.12,-0.72]*	-1.72 [-1.88,-1.56]*	-0.84 [-1,-0.68]*	N/A	-0.98 [-1.15,-0.81]*
Other	0.06 [-0.10,0.22]	-0.74 [-0.85,-0.63]*	0.14 [0.01,0.27]*	0.98 [0.81,1.15]*	N/A

Table 4.7: Differences of beta parameters of personal income with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	1st quartile	2nd quartile	3rd quartile	4th quartile	Unknown
1st quartile	N/A	0.23 [0.12,0.35]*	0.42 [0.32,0.52]*	0.62 [0.5,0.73]*	-0.20 [-0.42,0.03]
2nd quartile	-0.23 [-0.35,-0.12]*	N/A	0.19 [0.10,0.28]*	0.39 [0.28,0.50]*	-0.43 [-0.65,-0.21]*
3rd quartile	-0.42 [-0.52,-0.32]*	-0.19 [-0.28,-0.10]*	N/A	0.20 [0.12,0.28]*	-0.62 [-0.84,-0.40]*
4th quartile	-0.62 [-0.73,-0.5]*	-0.39 [-0.50,-0.28]*	-0.20 [-0.28,-0.12]*	N/A	-0.82 [-1.05,-0.59]*
Unknown	0.20 [-0.03,0.42]	0.43 [0.21,0.65]*	0.62 [0.40,0.84]*	0.82 [0.59,1.05]*	N/A

Table 4.8: Differences of beta parameters of household income with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	1st quartile	2nd quartile	3rd quartile	4th quartile
1st quartile	N/A	0.00 [-0.09,0.10]	0.04 [-0.16,0.07]	0.20 [0.07,0.32]*
2nd quartile	0.00 [-0.10,0.09]	N/A	0.04 [-0.04,0.12]	0.20 [0.10,0.30]*
3rd quartile	-0.04 [-0.16,0.07]	-0.04 [-0.12,0.04]	N/A	0.16 [0.07,0.25]*
4th quartile	-0.20 [-0.32,-0.07]*	-0.20 [-0.30,-0.10]*	-0.16 [-0.25,-0.07]*	N/A

Table 4.9: Differences of beta parameters of net household wealth with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	1st quartile	2nd quartile	3rd quartile	4th quartile
1st quartile	N/A	0.05 [-0.02,0.12]	0.02 [-0.06,0.10]	-0.10 [-0.19,-0.02]*
2nd quartile	-0.05 [-0.12,0.02]	N/A	-0.03 [-0.11,0.05]	-0.15 [-0.23,-0.07]*
3rd quartile	-0.02 [-0.10,0.06]	0.03 [-0.05,0.11]	N/A	-0.12 [-0.20,-0.04]*
4th quartile	0.10 [0.02,0.19]*	0.15 [0.07,0.23]*	0.12 [0.04,0.20]*	N/A

Table 4.10: Differences of beta parameters of education level with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	Low	Mid	High	Unknown
Low	N/A	0.03 [-0.09,0.14]	0.00 [-0.10,0.10]	0.18 [0.05,0.31]*
Mid	-0.03 [-0.14,0.09]	N/A	-0.03 [-0.14,0.08]	0.15 [0.02,0.29]*
High	0.00 [-0.10,0.10]	0.03 [-0.08,0.14]	N/A	0.18 [0.05,0.31]*
Unknown	-0.18 [-0.31,-0.05]*	-0.15 [-0.29,-0.02]*	-0.18 [-0.31,-0.05]*	N/A

Table 4.11: Differences of beta parameters of physical healthcare costs with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	€0	€1-5000	€5001-10000	€10001+	Unknown
€0	N/A	-0.06 [-0.28,0.17]	-0.87 [-1.11,-0.64]*	-1.53 [-1.80,-1.26]*	1.40 [1.11,1.69]*
€1-5000	0.06 [-0.17,0.28]	N/A	-0.81 [-0.92,-0.70]*	-1.47 [-1.63,-1.31]*	1.46 [1.07,1.85]*
€5001-10000	0.87 [0.64,1.11]*	0.81 [0.70,0.92]*	N/A	-0.66 [-0.83,-0.49]*	2.27 [1.88,2.66]*
€10001+	1.53 [1.26,1.80]*	1.47 [1.31,1.63]*	0.66 [0.49,0.83]*	N/A	2.93 [2.49,3.37]*
Unknown	-1.40 [-1.69,-1.11]*	-1.46 [-1.85,-1.07]*	-2.27 [-2.66,-1.88]*	-2.93 [-3.37,-2.49]*	N/A

Table 4.12: Differences of beta parameters of marital status with corresponding 95% confidence intervals (significant differences are marked with a *).

A \ B	Never married	Married	Divorced	Widowed	Unknown
Never married	N/A	-0.26 [-0.37,-0.14]*	-0.51 [-0.62,-0.39]*	0.09 [-0.09,0.27]	1.40 [1.11,1.69]*
Married	0.26 [0.14,0.37]*	N/A	-0.25 [-0.35,-0.15]*	0.35 [0.18,0.52]*	1.46 [1.07,1.85]*
Divorced	0.51 [0.39,0.62]*	0.25 [0.15,0.35]*	N/A	0.60 [0.44,0.76]*	2.27 [1.88,2.66]*
Widowed	-0.09 [-0.27,0.09]	-0.35 [-0.52,-0.18]*	-0.60 [-0.76,-0.44]*	N/A	2.93 [2.49,3.37]*
Unknown	-1.40 [-1.69,-1.11]*	-1.46 [-1.85,-1.07]*	-2.27 [-2.66,-1.88]*	-2.93 [-3.37,-2.49]*	N/A

5 On the relation between medication prescriptions and suicide

5.1 Introduction

Interventions at the second level, targeting sub-populations, require identification of groups at elevated risk of suicide. Prior studies have identified multiple high-risk groups, including people who survived a previous suicide attempt, people who experienced adverse childhood events, and people diagnosed with depression and/or substance use problems [59]. Men, people of middle age, people receiving income benefits and people living alone are also at higher risk to die by suicide as seen in Chapter 3.

Besides socio-demographic and mental health characteristics, poor physical health may also indicate groups at high risk of suicide, as frequently reported using medical registry data. In a representative US-based sample of over 11 million individuals, 64% of those who died by suicide made primary care health visits for reasons other

Based on [18]: G. Berkelmans, L.J. Schweren, S. Bhulai, R.D. van der Mei, R. Gilissen, A. Beekman. 'On the relation between medication prescriptions and suicide'. Submitted for publication.

than their mental health within one year prior to their death [6]. In Sweden, 0.3% of individuals suffering from non-communicable diseases (such as chronic respiratory diseases, cardiovascular diseases and diabetes) died by suicide within five years after diagnosis, compared to a five-year cumulative risk of 0.1% in the general population [134]. Ahmedani et al. recently showed that people who had been diagnosed with cancer, and among them especially those not diagnosed with any mental health condition, were at higher risk of suicide as well [5]. In another study, the relative risk of suicide was found to be as high as 12.6 within one week following a diagnosis of cancer compared to cancer-free individuals [57].

Several classes of medication are known to be involved in the pathogenesis of mental illness, such as depression, and thereby contribute to an elevated risk of suicide [26, 42, 72]. In a Norwegian study, 47.2% of males and 64.4% of females who died by suicide had been prescribed medication within twelve months prior to their death, compared to 23.3% of males and 34.4% of females in the general population [125]. Analysis of suicides in Northern Ireland suggests that almost half (45.2%) of those who died by suicide had been prescribed medication for one or more physical conditions [114]. At the same time, however, studies have not consistently identified prescription classes or groups of pharmacologically treated individuals – other than those treated prescribed psychotropic substances such as antidepressants – with elevated suicide risks. Several studies have found no associations between cardiovascular medications (including lipid-lowering drugs, calcium-channel blockers, beta-blockers and ACE-inhibitors) and suicide risk [34, 66, 109]. For other common medications such as albuterol inhalers, antibiotics, non-steroidal anti-inflammatory drugs and corticosteroids, however, evidence is either scarce and inconclusive or absent [66, 160, 169].

Identifying high-risk groups based on medication use may have multiple advantages. Most importantly, it may guide selective prevention efforts. Most patients taking medications for physical health problems are in regular contact with their general practitioner, medical specialist and/or pharmacist. Training these medical professionals to be more alert to changes in the mental health of their patients, especially of those patients being prescribed medications

associated with an elevated risk of suicide, may create an additional layer of protection around vulnerable individuals. Second, associations between suicide and pharmacological treatment may guide future research into biopsychosocial processes resulting in suicidal thoughts and behaviours. By design, identification of associations between medication use and suicide will not elucidate causal mechanisms. Nonetheless, identifying substances associated with higher suicide risks may point to relevant starting points for future research, and as such contribute to our understanding of biological processes contributing to suicidality and vice versa.

In the current study we investigate which pharmacological groups of medications confer a high risk of suicide, with particular focus on those medication types prescribed primarily for somatic rather than psychiatric conditions. By analysing data regarding the entire population of the Netherlands, we have sufficient statistical power to test the groups of medication that are most associated with suicide.

5.2 Methodology

5.2.1 Data

The data used in this study were provided by CBS. For this study, all inhabitants of the Netherlands aged 10 and older were included as of the 31st of December of the years 2009 to 2019, resulting in a total of 173,496,482 person years. Of these person years, a total of 19,657 (11.33 per 100,000 person years) were followed by suicide within 12 months after observation (set to 31 December unless otherwise specified).

5.2.2 Predictor: medication use

All inhabitants of the Netherlands are required by law to register with a healthcare insurance company, ensuring universal access to healthcare including pharmaceutical care. Each provision of healthcare (e.g., consultation, medication, laboratory testing) covered by the insurance policy is registered by the insurance company and data

thereof are provided to CBS. Medicines dispensed during admission to hospitals and nursing homes are not included, nor are medicines not covered by the healthcare insurance.

Medication use is encoded using 4 positions ATC-codes (anatomical main group [A-Z], therapeutic subgroup [01-99] and pharmacological subgroup [A-Z]), yielding 268 unique medication classes. Excluding classes not prescribed during the study period (N = 48), a non-informative class labelled "unknown" (N = 1), and classes with less than 10 suicides over the 11 year period (N = 77) resulted in 142 unique medication classes included. For a list of all included medication classes, see the supplementary material. Per year and per medication class, each individual is categorised as a user (1) or a non-user (0).

5.2.3 Outcome: suicide

The outcome of interest used is whether the individual died by suicide within one calendar year following observation (for example, suicide [yes/no] in 2014 after using medication A in 2013). For each deceased inhabitant of the Netherlands, CBS registers cause of death for statistical purposes. In the Netherlands, all suicide cases are confirmed by a forensic pathologist.

5.2.4 Covariates

First, unadjusted odds ratios (per 100,000 person years) were computed within each medication class by performing univariate logistic regression. These analyses identify patient groups at increased and decreased risk of suicide. We applied correction for multiple testing using Bonferroni correction, indicating an appropriate alpha of 0.05/142=0.000352. Next, all significant predictors of suicide risk with odds ratios greater than 3.00 were evaluated together in a single multivariate prediction model, to account for confounding by polypharmacy. The same analyses are performed for all medications showing a significant reduction (though without an odds ratio threshold).

Second, we evaluate which patient groups are at increased and

decreased risk of suicide while adjusting for age, sex and mental healthcare use, by using univariate conditional logistic regression models. As before, alpha is set to 0.000352 and significant predictors with odds ratios above 3.00 are evaluated together in a multivariate model. As before, the same analyses were performed for all medications showing a significant reduction (without an odds ratio threshold).

5.2.5 Sensitivity analyses

We performed three sensitivity analyses to test the validity of our findings, to see whether they hold generally or are limited to certain sub-populations based on our covariates. First, the conditional logistic regression models were re-estimated with an extra interaction term between medication and sex. Second, the conditional logistic regression models were re-estimated with an extra interaction term between medication and receiving mental healthcare. Third, the conditional logistic regression models were re-estimated with extra interaction terms between medication and age. In the latter analyses, age groups were merged into younger age (10-30), average age (30-60), and older age (60+).

5.3 Results

5.3.1 Higher risk of suicide

The estimated Odds-Ratios for each medication group are shown in Figure 5.1. After correction for multiple testing, risk of suicide was significantly elevated among individuals prescribed 75/142 different classes of medication compared to individuals not prescribed these medications. In univariate analyses, six classes showed an odds-ratio >10 .00 and an additional five classes showed an odds ratio >3.00 (Section 5.3.1). Among those, eight medication classes target the nervous system (anti-cholinergic agents, antipsychotic agents, anxiolytics, hypnotics and sedatives, drugs used in addictive disorders, antidepressants, anti-epileptics, and other nervous system drugs), two target the alimentary tract and digestive system (vitamin

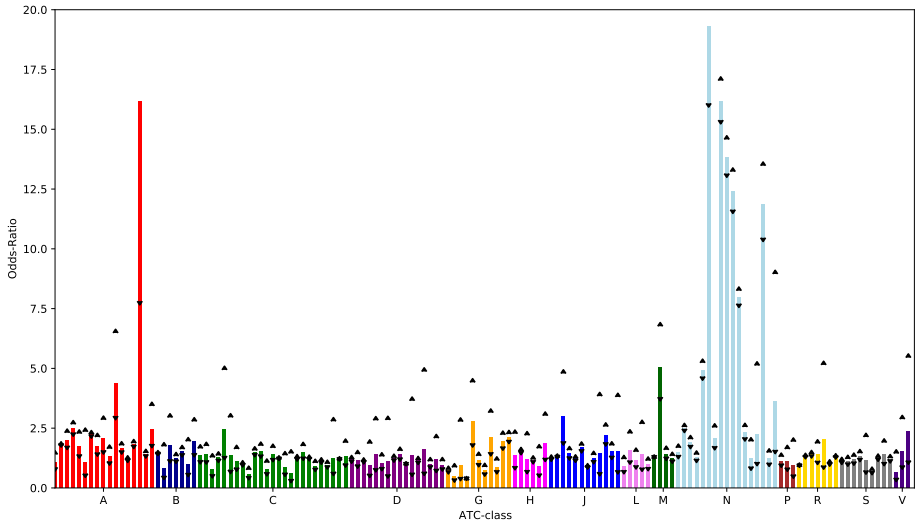


Figure 5.1: Univariate Odds-Ratios for the various medication classes, A: Alimentary Tract and Metabolism, B: Blood and Blood Forming Organs, C: Cardiovascular System, D: Dermatologicals, G: Genito Urinary System and Sex Hormones, H: Systemic Hormonal Preparations, J: Anti-infectives for Systemic Use, L: Antineoplastic and Immunomodulating Agents, M: Musculoskeletal system, N: Nervous System, P: Antiparasitic Products, R: Respiratory System, S: Sensory Organs, V: Various.

B1/B6/B12, and digestives including enzymes), and one targets the muscular system (centrally acting muscle relaxants).

Testing all medication classes with ORs > 3.00 in a single multivariate model, all ORs were attenuated, yet all but two medication classes remained significantly associated with an increased risk of suicide. One medication class that was associated with an increased risk of suicide in the univariate analyses ("drugs used in addictive disorders") was associated with a reduced risk of suicide in the multivariate model. The last remaining medication class ("other nervous system drugs") was no longer significantly associated with suicide risk (Section 5.3.1).

After adjusting for age, sex and mental healthcare use, risk of suicide was elevated among users of 102/142 medication classes compared to non-users. Among these there were no classes with an odds-ratio > 10.00, and 12 classes with odds-ratios > 3.00. All but one of these

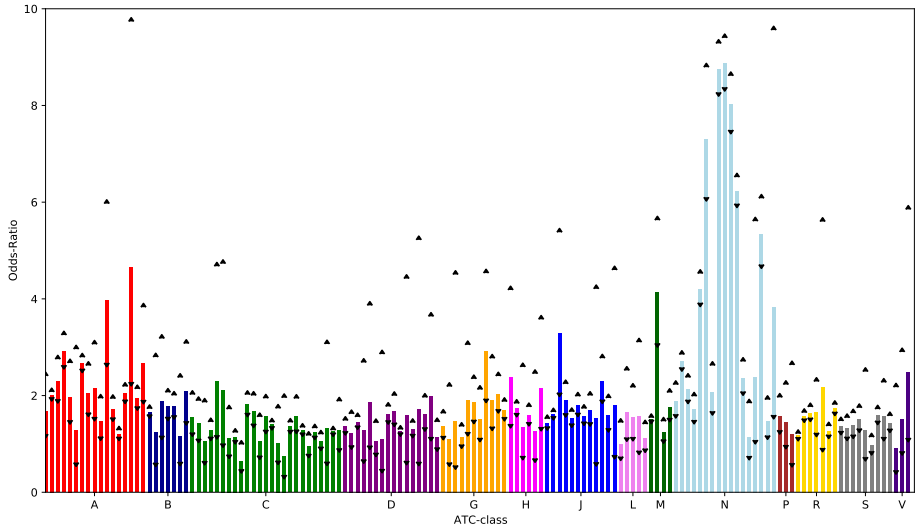


Figure 5.2: Conditional Odds-Ratios for the various medication classes, classes are the same as in Figure 5.1.

medication classes (J01D, beta-lactam anti-bacterials other than penicillins) had yielded a significant increased risk of suicide in the unadjusted analyses. As before, after testing all medication classes with ORs > 3.00 in a single adjusted multivariate model, all but two medication classes (drugs used in addictive disorders, and other nervous drugs) remained significantly associated with an increased risk of suicide. One medication class that was associated with an increased risk of suicide in the univariate analyses ('drugs used in affective disorders'), was associated with a reduced risk of suicide in the multivariate model (Section 5.3.1).

5.3.2 Lower risk of suicide

Risk of suicide was, significantly reduced among users of four medication classes compared to non-users, namely hormonal contraceptives, low-ceiling diuretics, decongestants and anti-allergics, and non-steroidal anti-infectives and anti-septics (Section 5.3.2). In multivariate analyses, considering only the medication classes with significant reductions in the univariate case we see that all four medication classes retained significant reductions.

Table 5.1: Medications with an odds-ratio over 3.00 in the univariate analyses, UV OR = Univariate odds-ratio, MV OR = Multivariate odds-ratio, CI = confidence interval.

		Unadjusted		Adjusted for sex,age, MH	
Class	Description	UV OR [CI]	MV OR [CI]	UV OR [CI]	MV OR [CI]
N04A	Anti-cholinergic agents	19.29 [15.93,23.36]	1.75 [1.62,1.89]	7.31 [6.03,8.87]	1.45 [1.34,1.57]
N05A	Anti-psychotics	16.18 [15.23,17.19]	6.25 [6.07,6.44]	8.76 [8.19,9.36]	3.23 [3.13,3.33]
A11D	Vitamin B1, B6, B12	16.15 [7.66,34.06]	2.25 [1.67,3.04]	4.65 [2.20,9.81]	1.46 [1.08,1.96]
N05B	Anxiolytics	13.83 [12.99,14.73]	4.14 [4.03,4.25]	8.87 [8.30,9.48]	3.01 [2.93,3.10]
N05C	Hypnotics and sedatives	12.40 [11.48,13.38]	3.01 [2.85,3.19]	8.03 [7.42,8.69]	1.98 [1.87,2.10]
N07B	Drugs used in addictive disorders	11.85 [10.31,13.63]	0.61 [0.56,0.67]	5.34 [4.67,6.11]	0.68 [0.62,0.74]
N06A	Antidepressants	7.79 [7.55,8.40]	1.74 [1.20,2.51]	6.23 [5.89,6.60]	1.90 [1.31,2.74]
M03B	Centrally acting muscle relaxants	5.02 [3.64,6.92]	1.52 [1.33,1.73]	4.14 [3.00,5.71]	1.47 [1.29,1.67]
N03A	Anti-epileptics	4.93 [4.51,5.39]	1.71 [1.65,1.77]	4.20 [3.84,4.60]	1.57 [1.52,1.63]
A09A	Digestive medications, including enzymes	4.35 [2.85,6.63]	2.11[1.79,2.50]	3.96 [2.60,6.05]	1.97 [1.67,2.34]
N07X	Other nervous system drugs	3.62 [1.44,9.10]	0.99 [0.97,1.01]	3.83 [1.52,9.64]	0.99 [0.97,1.02]
J01D	Beta-lactam antibacterials, other than penicillins	Below threshold	Not included	3.29 [1.98,5.46]	2.06 [1.68,2.52]

Table 5.2: Medications with an significant reduction in suicide risk in the univariate analyses, UV OR = Univariate odds-ratio, MV OR = Multivariate odds-ratio, CI = confidence interval.

		Unadjusted		Adjusted for sex,age, MH	
Class	Description	UV OR [CI]	MV OR [CI]	UV OR [CI]	MV OR [CI]
G03A	Hormonal contraceptives for systemic use	0.39 [0.32,0.48]	0.28 [0.25,0.31]	Not significant	Not applicable
C03B	Low ceiling diuretics, excluding thiazides	0.56 [0.34,0.91]	0.38 [0.30,0.49]	Not significant	Not applicable
S01G	Decongestants and antiallergics	0.69 [0.55,0.87]	0.76 [0.73,0.78]	Not significant	Not applicable
G01A	Anti-infectives and antiseptics excluding combinations with corticosteroids	0.73 [0.59,0.91]	0.56 [0.50,0.62]	Not significant	Not applicable

However, after adjusting for the effects of sex, age, and mental healthcare usage, none of the medication classes were associated with a significant reduction in suicide rates.

5.3.3 Sensitivity analyses

Among medication classes associated with suicide risk in multivariate analyses adjusted for sex, age, and mental healthcare ($p < 0.000352$ and $OR > 3.00$, as listed in Section 5.3.1), all those targeting the nervous system showed a significant interaction with sex. In all cases, females showed stronger associations compared to males, however in

males too the medication classes were associated with elevated risk of suicide. Another eight medication classes not identified in the main analyses showed significant interactions with sex, such that the elevated risk of suicide was stronger in females compared to males (drugs for peptic ulcer and gastro-oesophagel reflux disease (A02B), propulsives (A03F), drugs for constipation (A06A), vitamin A and D, including combinations of the two (A11C), anti-inflammatory and anti-rheumatic products (M01A), opioids (N02A), drugs for obstructive airway diseases (R03A), anti-histamines for systemic use (R06A)). No associations were found to be stronger in males compared to in females.

Among those medication classes associated with suicide risk (Section 5.3.1), all those targeting the nervous system showed a significant interaction with age. In all cases, younger individuals (either 10-30 years old or 30-60 years old) showed stronger associations compared to older individuals (60+ years old), however, in older individuals the medication classes were associated with elevated risk of suicide as well. Another six medication classes not identified in the main analyses showed significant interactions with age such that the elevated risk of suicide was stronger in younger compared to older individuals, namely vitamin A and D, including combinations of the two (A11C), potassium (A12B), i.v. solutions (B05B), high-ceiling diuretics (C03C), beta blocking agents (C07A) and drugs affecting bone structure and mineralization (M05B). Finally, three medication classes were identified that showed a stronger association in older compared to younger individuals, namely plain corticosteroids (D07A), other combinations of corticosteroids (D07X) and psychostimulants, agents used for attention-deficit/hyperactivity disorder and nootropics (N06B).

Among medication classes associated with suicide risk, five interacted with mental healthcare use such that associations were stronger among those receiving mental healthcare. All of those targeted the nervous system (anti-epileptics (N03A), anti-psychotics (N05A), anxiolytics (N05B), hypnotics and sedatives (N05C), and anti-depressants (N06A)), and in all cases use of these medications was associated with a significant increased risk of suicide among those not receiving mental healthcare as well.

A large number of other medication classes (68/137) also interacted with mental healthcare use, such that associations between medication use and suicide were stronger among those receiving mental healthcare. A total of 34 medication classes were identified to be associated with a higher risk of suicide among those receiving mental healthcare but not among those not receiving mental healthcare. No medication classes were identified that were associated with a higher risk of suicide among those not receiving mental healthcare.

All other results of the sensitivity analyses are shown in the supplementary material.

5.4 Discussion

In the current study we applied regular and conditional logistic regression analyses to national health insurance data to investigate associations between medication prescriptions and suicide risk. As expected, we found a stronger association with suicide among those receiving medications acting on the central nervous system that are typically prescribed to treat psychiatric conditions, including antidepressants, anxiolytics and antipsychotic agents. However, there was also an increased risk among one class of medications that acts on the central nervous system but is primarily prescribed for somatic conditions, namely centrally acting muscle relaxants. In addition, we found that suicide risk was higher among those being prescribed medications targeting the alimentary and digestive tract, namely digestive medications and vitamin B supplements. Finally, we found associations with suicide among users of beta-lactam antibacterials. Associations between medication use and suicide were generally stronger among younger individuals, females, and those receiving mental healthcare. For the latter group we also found that a wide range of somatic medications had a substantial association with suicide risk, which was not identified among those not receiving mental healthcare.

The aim of our study was to identify groups at high risk of suicide based on pharmacological treatment, with particular focus on those medication types prescribed primarily for somatic rather

than psychiatric conditions. We identified two medication classes targeting the alimentary tract that were associated with an elevated risk of suicide. First, suicide risk was higher among patients being prescribed digestive medications. In the multivariate adjusted model, being prescribed digestive medications yielded an elevated suicide risk comparable to being prescribed antidepressants. Several mechanisms might explain this association. First, the association between digestives and suicide might reflect the increased prevalence of other health conditions among patients with psychiatric problems, including abdominal pain (e.g., Lexne et al. [95]) and irritable bowel syndrome [56]. Similarly, psychiatric patients are more prone to poor lifestyle habits, increasing the likelihood of digestive problems in this group [84, 124, 137, 151]. Third, prescription of digestive medications might point to other medication types causing obstipation. Especially anti-inflammatory drugs, opioids and antidepressants are known to have such side-effects. Finally, depressive disorders, as well as several other psychiatric disorders, have been shown to be associated with changes in microbiome composition [98, 108]. It has been speculated that emotional affect may influence gastrointestinal bacteria and vice versa via the gut-brain axis [136], suggesting a direct causal link between suicidality and gastrointestinal health. For insight in causal mechanisms underlying the observed association, we recommend additional research applying alternative study designs, such as for example, monitoring the psychiatric symptoms of high risk groups at regular intervals (daily/weekly) to see whether symptoms appear/worsen after taking the medication, or conversely the medication follows appearing/worsening symptoms.

In addition, we found an elevated suicide risk among those being prescribed vitamin B preparations. Vitamin B is prescribed to treat vitamin B deficiencies, in most cases resulting from persistent deficient dietary patterns, excessive alcohol intake and/or (less commonly) from intestinal anomalies. Interestingly, in our population, vitamin B was prescribed almost exclusively to those receiving mental healthcare. We speculate that the association between vitamin B medications and suicide risk might reflect an increased suicide risk among patient groups characterised by chronicity and poor self-care, including those with eating disorders, substance use disorders and

schizophrenia.

Two other findings are of interest. First, we found that pharmacological treatment plays a role especially among people receiving mental healthcare. Interestingly, an elevated risk of suicide was found for those being prescribed a wide range of medication types, i.e., no one medication type stood out. It is well documented that somatic comorbidities are a serious risk factor for suicide [39, 50]. Recipients of mental healthcare may be more vulnerable to the effects of poor somatic health due to poorer coping mechanisms. Alternatively, poor somatic health may compound to or exacerbate existing psychiatric symptoms. Future studies, especially longitudinal ones, might be able to address this issue. We also found that medication prescription of metabolic and nervous system medications is associated with suicide risk especially among females. It is unclear why this is the case.

The aim of the current study was to identify patient groups at high risk of suicide to guide selective preventive efforts. Our findings suggest that, among those being prescribed psychiatric medication, medical professionals should pay additional attention to younger patients and women. Prescription of vitamin B and digestive medications is of relevance as well, especially in the context of (severe) psychiatric disorders, but also outside of this context. Additionally, care should be taken to monitor psychological symptoms when prescribing centrally acting muscle relaxants and/or beta-lactam anti-bacterials. Finally, mental healthcare professionals should be extra alert treating patients with one or more somatic comorbidities, regardless of their origin.

We wish to highlight four relevant limitations to our study. First, this study is observational in nature, which means no conclusions can be drawn about causality. Second, as medications are limited to the ATC4 level, associations that are highly specific might be masked. Third, our study is based on prescription information rather than (self-)reporting of actual intake. While this limitation does not affect our ability to detect patient groups at higher risk, it should be kept in mind when speculating about potential causal effects of medication. Finally, medications not covered by health insurance

were not included. Although the vast majority of medications are covered in the Netherlands, especially the ones commonly prescribed, we cannot exclude the possibility that associations might have been missed. Our study has strengths as well. Most importantly, our data covers the entire population of the Netherlands, granting statistical power to investigate even rare events such as suicide . Secondly, we can account for the different base rates among the different sexes, age groups, and those receiving mental healthcare versus those not receiving mental healthcare.

To conclude, we found that a large number of medication classes were associated with a higher risk of suicide which could not be explained by sex, age, or mental healthcare use. Of special note were two clusters: one is the nervous systems cluster, and the second was the alimentary tract and metabolism cluster. Additionally, the classes of centrally acting muscle relaxants and beta-lactam antibacterials are noteworthy. It is important for physicians to be aware of this when prescribing said medications, and to ensure proper patient follow-up after prescription.

5.A Appendix

5.A.1 Data description

Description of data, for each ATC 4 code of an included medication class, the name is given, the amount of person years in which it was described and the number of person years in which it was prescribed that resulted in suicide.

ATC4 Code	Name	Size patient group	Number of suicides within patient group
A01A	Stomalotogical preparations	690653	82
A02B	Drugs for peptic ulcer and gastro-oesophageal reflux disease	23833644	4410
A03A	Drugs for functional gastrointestinal disorders	1277718	286
A03F	Propulsives	2781564	760
A04A	Anti-emetics and anti-nauseants	589755	116
A05A	Bile Therapy	142613	17
A06A	Drugs for constipation	12691591	2949
A07A	Intestinal anti-infectives	895234	176
A07D	Anti-propulsives	390002	91
A07E	Intestinal anti-inflammatory agents	871116	129
A09A	Digestives, including enzymes	146545	72
A10A	Insulins and analogues	2905338	539
A10B	Blood glucose lowering drugs, excluding insulins	7337169	988
A11C	Vitamin A and D, including combinations of the two	6381164	1282

ATC4 Code	Name	Size patient group	Number of suicides within patient group
A11D	Vitamin B1, plain and in combination with vitamin B6 and B12	12606	23
A12A	Calcium	5118830	814
A12B	Potassium	324216	90
B01A	Anti-thrombotic agents	18416132	2896
B02A	Anti-hemorrhagics	194591	18
B02B	Vitamin K and other hemostatics	211284	43
B03A	Iron preparations	2794879	390
B03B	Vitamin B12 and folic acid	3081814	525
B03X	Other anti-anemic preparations	203131	23
B05B	I.V. Solutions	348507	77
C01A	Cardiac glycosides	983509	150
C01B	Anti-arrhythmics, class I and III	787067	123
C01C	Cardiac stimulants excluding cardiac glycosides	386864	34
C01D	Vasodilators used in cardiac diseases	2978523	426
C01E	Other cardiac preparations	86429	24
C02A	Anti-adrenergic agents, centrally acting	122639	19
C02C	Anti-adrenergic agents, pheripherally acting	459677	58
C03A	Low-ceiling diuretics, thiazides	6700208	759
C03B	Low-ceiling diuretics, excluding thiazides	836656	53
C03C	High-ceiling diuretics	4063398	682
C03D	Aldosterone antagonists and other potassium-sparing agents	1579722	275
C03E	Diuretics and potassium-sparing agents in combination	755580	67
C05A	Agents for treatment of hemorrhoids and anal fissures for topical use	1236318	198
C07A	Beta blocking agents	17170364	2361
C07B	Beta blocking agents and thiazides	400832	39
C07C	Beta blocking agents and other diuretics	193323	13
C08C	Selective calcium channel blockers with mainly vascular effects	7965165	1117
C08D	Selective calcium channel blockers with direct cardiac effects	1223445	205
C09A	Ace inhibitors, plain	10515197	1465
C09B	Ace inhibitors, combinations	1620691	167
C09C	Angiotensin II receptor blockers (ARBs), plain	7078264	892
C09D	Angiotensin II receptor blockers (ARBs), combinations	3216348	349
C09X	Other agents acting on the renin-angiotensin system	121365	17
C10A	Lipid modifying agents, plain	20123844	2700
C10B	Lipid modifying agents, combinations	463908	70
D01A	Anti-fungals for topical use	5255701	701
D01B	Anti-fungals for systemic use	962916	124
D02A	Emollients and protectives	8344156	1056
D02B	Protectives against uv-radiation	207116	22
D04A	Anti-pruritics, including anti-histamines, anesthetics, etc,	143411	23
D05A	Anti-psoriatics for topical use	856800	100
D05B	Anti-psoriatics for systemic use	103166	13
D06A	Anti-biotics for topical use	5187489	720
D06B	Chemotherapeutics for topical use	1873843	292
D07A	Corticosteroids, plain	17318533	1979
D07C	Corticosteroids, combinations with anti-biotics	78202	12
D07X	Corticosteroids, other combinations	4716751	606
D08A	Anti-septics and disinfectants	54430	10
D10A	Anti-acne preparations for topical use	2024515	230
D10B	Anti-acne preparations for systemic use	242680	33
D11A	Other dermatological preparations	1377737	148
G01A	Anti-infectives and anti-septics, excluding combinations with corticosteroids	3196319	266
G02B	Contraceptives for topical use	443250	25
G02C	Other gynecologicals	93284	10
G03A	Hormonal contraceptives for systemic use	6893936	311
G03B	Androgens	167667	53
G03C	Estrogens	1443668	188
G03D	Progestogens	1173316	95
G03F	Progestogens and estrogens in combination	264719	63
G03H	Anti-androgens	825405	82
G04B	Urologicals	1418329	309
G04C	Drugs used in benign prostatic hypertrophy	3349703	785
H01B	Posterior pituitary lobe hormones	247530	38
H02A	Coritcosteroids for systemic use, plain	8167875	1384
H02B	Corticosteroids for systemic use, combinations	201200	27
H03A	Thyroid preparations	4820380	632
H03B	Anti-thyroid preparations	255203	26
H04A	Glycogenolytic hormones	216833	46

ATC4 Code	Name	Size patient group	Number of suicides within patient group
J01A	Tetracyclines	8389103	1164
J01C	Beta-lactam anti-bacterials, penicillins	19093748	2759
J01D	Other beta-lactam anti-bacterials	148457	50
J01E	Sulfonamides and trimethoprim	2070915	332
J01F	Macrolides, lincosamides, and streptogramins	5810226	796
J01M	Quinolone anti-bacterials	3809290	714
J01X	Other anti-bacterials	7773382	791
J02A	Anti-mycotics for systemic use	2133756	300
J04A	Drugs for treatment of tuberculosis	74429	12
J05A	Direct acting anti-virals	1094281	269
J07A	Bacterial vaccines	1284994	220
J07B	Viral vaccines	81155	14
L01B	Anti-metabolites	727305	74
L02A	Hormones and related agents	363408	64
L02B	Hormone antagonists and related agents	714466	93
L03A	Immunostimulants	163468	26
L04A	Immunosuppressants	1360475	149
M01A	Anti-inflammatory and anti-rheumatic products, non-steroids	27830189	3899
M03B	Muscle relaxants, centrally acting agents	221051	125
M04A	Anti-gout preparations	1862857	298
M05B	Drugs affecting bone structure and mineralization	2680636	375
N01B	Anesthetics, local	1888861	319
N02A	Opioids	9598980	2512
N02B	Other analgesics and anti-pyretics	3015246	639
N02C	Anti-migraine preparations	2588929	375
N03A	Anti-epileptics	3388705	1756
N04A	Anti-cholinergic agents	166088	356
N04B	Dopaminergic agents	814891	190
N05A	Anti-psychotics	3139995	4511
N05B	Anxiolytics	3244388	4097
N05C	Hypnotics and sedatives	2012476	2494
N06A	Anti-depressants	10806058	6799
N06B	Psychostimulants, agents used for adhd and nootropics	1983655	510
N06D	Anti-dementia drugs	332603	47
N07A	Parasympathomimetics	67728	17
N07B	Drugs used in addictive disorders	521917	678
N07C	Anti-vertigo preparations	1054886	145
N07X	Other nervous system drugs	36648	15
P01A	Agents against amoebasis and other protozoal diseases	1481654	188
P01B	Anti-malarials	460163	58
P03A	Ectoparasiticides, including scabicides	182169	19
R01A	Decongestants and other nasal preparations for topical use	14247515	1536
R03A	Adrenergics, inhalants	12728906	1871
R03B	Other drugs for obstructive airway diseases, inhalants	6780475	1060
R03D	Other systemic drugs for obstructive airway diseases	622245	99
R05C	Expectorants, excluding combinations with cough suppressants	60662	14
R05D	Cough suppressants, excluding combinations with expectorants	6679251	783
R06A	Anti-histamines for systemic use	12957172	1884
S01A	Anti-infectives	6033974	763
S01B	Anti-inflammatory agents	2476166	312
S01C	Anti-inflammatory agents and anti-infectives in combination	2119634	282
S01E	Anti-cglaucoma preparations and miotics	2413630	362
S01F	Mydriatics and cycloplegics	207608	27
S01G	Decongestants and anti-allergics	3171278	251
S01X	Other opthalmologicals	6417361	905
S02A	Anti-infectives	524104	82
S02C	Corticosteroids and anti-infectives in combination	4562836	621
V01A	Allergens	216217	16
V03A	All other therapeutic products	160254	28
V07A	All other non-therapeutic products	63761	17

5.A.2 Results univariate logistic regression

Results of the univariate logistic regression models. For each model only the statistics (Beta, Standard Error of Beta, p-value, t-test, and Odds Ratio) for the parameter corresponding to the medication class are reported. The p-values are reported up to ϵ = 2.2E-16 which is the machine precision.

ATC name	Beta estimate	Standard Error	p-value	t-test	Odds Ratio estimate
A01A	0.05	0.11	0.67	0.42	1.05
A02B	0.60	0.02	$< \epsilon$	34.91	1.82
A03A	0.69	0.06	$< \epsilon$	11.55	1.99
A03F	0.90	0.04	$< \epsilon$	24.42	2.47
A04A	0.55	0.09	2.67E-9	5.95	1.74
A05A	0.05	0.24	0.83	0.21	1.05
A06A	0.80	0.02	$< \epsilon$	40.30	2.24
A07A	0.56	0.08	2.3E-13	7.33	1.74
A07D	0.72	0.11	5.3E-12	6.90	2.06
A07E	0.27	0.09	2.29E-3	3.05	1.31
A09A	1.47	0.12	$< \epsilon$	12.45	4.35
A10A	0.50	0.04	$< \epsilon$	11.54	1.66
A10B	0.18	0.03	2.9E-8	5.55	1.20
A11C	0.60	0.03	$< \epsilon$	20.87	1.83
A11D	2.78	0.21	$< \epsilon$	13.32	16.15
A12A	0.35	0.04	$< \epsilon$	9.81	1.42
A12B	0.90	0.11	$< \epsilon$	8.51	2.46
B01A	0.38	0.02	$< \epsilon$	18.64	1.46
B02A	-0.20	0.24	0.39	-0.86	0.82
B02B	0.59	0.15	1.2E-4	3.84	1.80
B03A	0.21	0.05	3.4E-5	4.15	1.24
B03B	0.42	0.04	$< \epsilon$	9.43	1.52
B03X	-0.00	0.21	1.00	-0.00	1.00
B05B	0.67	0.11	4.5E-9	5.87	1.95
C01A	0.30	0.08	2.6E-4	3.65	1.35
C01B	0.32	0.09	3.5E-4	3.57	1.38
C01C	-0.25	0.17	0.14	-1.48	0.78
C01D	0.24	0.05	1.2E-6	4.85	1.27
C01E	0.90	0.20	1.1E-5	4.39	2.45
C02A	0.31	0.23	0.17	1.36	1.37
C02C	0.11	0.13	0.41	0.82	1.11
C03A	-0.00	0.04	1.00	-0.00	1.00
C03B	-0.58	0.14	2.2E-5	-4.24	0.56
C03C	0.40	0.04	$< \epsilon$	10.38	1.50
C03D	0.43	0.06	8.4E-13	7.15	1.54
C03E	-0.25	0.12	4.4E-2	-2.01	0.78
C05A	0.35	0.07	1.0E-6	4.89	1.42
C07A	0.22	0.02	$< \epsilon$	9.91	1.24
C07B	-0.15	0.16	0.34	-0.95	0.86
C07C	-0.52	0.28	6.0E-2	-1.88	0.59
C08C	0.22	0.03	2.9E-13	7.30	1.25
C08D	0.39	0.07	1.9E-8	5.62	1.48
C09A	0.22	0.03	2.2E-16	8.16	1.25
C09B	-0.10	0.08	0.22	-1.23	0.91
C09C	0.11	0.03	1.2E-3	3.24	1.12
C09D	-0.04	0.05	0.42E-1	-0.81	0.96
C09X	0.21	0.24	0.38E-1	0.88	1.24
C10A	0.19	0.02	$< \epsilon$	9.34	1.21
C10B	0.29	0.12	1.6E-2	2.40	1.33
D01A	0.17	0.04	1.2E-5	4.39	1.18
D01B	0.13	0.09	0.15	1.43	1.14
D02A	0.12	0.03	2.3E-4	3.69	1.12
D02B	-0.06	0.21	0.76	-0.30	0.94
D04A	0.35	0.21	9.5E-2	1.67	1.42
D05A	0.03	0.10	0.77	0.30	1.03
D05B	0.11	0.28	0.70	0.38	1.11
D06A	0.21	0.04	3.2E-8	5.53	1.23
D06B	0.32	0.06	4.4E-8	5.48	1.38
D07A	0.01	0.02	0.69	0.40	1.01
D07C	0.30	0.29	0.29	1.05	1.35
D07X	0.13	0.04	1.7E-3	3.14	1.14

ATC name	Beta estimate	Standard Error	p-value	t-test	Odds Ratio estimate
D08A	0.48	0.32	0.13	1.53	1.62
D10A	0.00	0.07	0.97E-1	0.04	1.00
D10B	0.18	0.17	0.29	1.05	1.20
D11A	-0.05	0.08	0.52	-0.65	0.95
G01A	-0.31	0.06	3.8E-7	-5.08	0.73
G02B	-0.70	0.20	4.8E-4	-3.49	0.50
G02C	-0.06	0.32	0.86	-0.18	0.95
G03A	-0.95	0.06	$< \epsilon$	-16.54	0.39
G03B	1.03	0.14	7.9E-14	7.47	2.80
G03C	0.14	0.07	5.5E-2	1.92	1.15
G03D	-0.34	0.10	1.0E-3	-3.29	0.71
G03F	0.74	0.13	3.7E-9	5.90	2.10
G03H	-0.13	0.11	0.23	-1.19	0.88
G04B	0.66	0.06	$< \epsilon$	11.54	1.94
G04C	0.75	0.04	$< \epsilon$	20.54	2.11
H01B	0.30	0.16	6.1E-2	1.87	1.36
H02A	0.43	0.03	$< \epsilon$	15.33	1.53
H02B	0.17	0.19	0.38	0.88	1.18
H03A	0.15	0.04	1.9E-4	3.72	1.16
H03B	-0.11	0.20	0.58	-0.54	0.90
H04A	0.63	0.15	2.1E-5	4.26	1.87
J01A	0.21	0.03	1.4E-12	7.09	1.24
J01C	0.28	0.02	$< \epsilon$	13.53	1.32
J01D	1.09	0.14	1.3E-14	7.71	2.98
J01E	0.35	0.06	2.0E-10	6.36	1.42
J01F	0.20	0.04	5.0E-8	5.45	1.22
J01M	0.52	0.04	$< \epsilon$	13.59	1.68
J01X	-0.11	0.04	2.0E-3	-3.09	0.89
J02A	0.22	0.06	1.7E-4	3.76	1.24
J04A	0.35	0.29	0.22	1.22	1.42
J05A	0.78	0.06	$< \epsilon$	12.74	2.19
J07A	0.42	0.07	8.0E-10	6.15	1.52
J07B	0.42	0.27	0.12	1.57	1.52
L01B	-0.11	0.12	0.35	-0.93	0.90
L02A	0.44	0.13	4.1E-4	3.53	1.56
L02B	0.14	0.10	0.18	1.34	1.15
L03A	0.34	0.20	8.4E-2	1.73	1.40
L04A	-0.03	0.08	0.68	-0.42	0.97
M01A	0.26	0.02	$< \epsilon$	14.46	1.30
M03B	1.61	0.09	$< \epsilon$	17.97	5.02
M04A	0.35	0.06	2.1E-9	5.99	1.42
M05B	0.21	0.05	3.9E-5	4.11	1.24
N01B	0.40	0.06	7.6E-13	7.17	1.50
N02A	0.92	0.02	$< \epsilon$	42.92	2.50
N02B	0.64	0.04	$< \epsilon$	15.96	1.90
N02C	0.25	0.05	1.7E-6	4.79	1.28
N03A	1.59	0.03	$< \epsilon$	63.75	4.93
N04A	2.96	0.05	$< \epsilon$	55.27	19.29
N04B	0.73	0.07	$< \epsilon$	9.97	2.07
N05A	2.78	0.02	$< \epsilon$	164.03	16.18
N05B	2.63	0.02	$< \epsilon$	149.53	13.83
N05C	2.52	0.02	$< \epsilon$	117.41	12.40
N06A	2.08	0.01	$< \epsilon$	138.35	7.97
N06B	0.83	0.04	$< \epsilon$	18.59	2.30
N06D	0.22	0.15	0.13	1.52	1.25
N07A	0.80	0.24	1.0E-3	3.28	2.22
N07B	2.47	0.04	$< \epsilon$	63.22	11.85
N07C	0.19	0.08	2.0E-2	2.33	1.21
N07X	1.29	0.26	6.5E-7	4.97	3.62
P01A	0.11	0.07	1.2E-1	1.56	1.12
P01B	0.11	0.13	0.42	0.81	1.11
P03A	-0.08	0.23	0.72	-0.36	0.92
R01A	-0.05	0.03	4.2E-2	-2.03	0.95
R03A	0.28	0.02	$< \epsilon$	11.69	1.33
R03B	0.34	0.03	$< \epsilon$	10.69	1.40
R03D	0.34	0.10	7.1E-4	3.38	1.41
R05C	0.71	0.27	7.8E-3	2.66	2.04
R05D	0.04	0.04	0.33	0.97	1.04
R06A	0.27	0.02	$< \epsilon$	11.25	1.31
S01A	0.11	0.04	2.0E-3	3.09	1.12
S01B	0.11	0.06	5.9E-2	1.89	1.11
S01C	0.16	0.06	6.6E-3	2.71	1.18
S01E	0.29	0.05	7.7E-8	5.37	1.33
S01F	0.14	0.19	0.47	0.72	1.15
S01G	-0.36	0.06	9.8E-9	-5.74	0.69
S01X	0.23	0.03	2.0E-11	6.71	1.26

ATC name	Beta estimate	Standard Error	p-value	t-test	Odds Ratio estimate
S02A	0.32	0.11	3.4E-3	2.93	1.38
S02C	0.19	0.04	3.7E-6	4.63	1.21
V01A	-0.43	0.25	8.8E-2	-1.71	0.65
V03A	0.43	0.19	2.2E-2	2.29	1.54
V07A	0.86	0.24	4.2E-4	3.53	2.35

5.A.3 Results conditional logistic regression models

Results of the conditional logistic regression models. For each model only the statistics (Beta, Standard Error of Beta, p-value, t-test, and Odds Ratio) for the parameter corresponding to the medication class are reported. The p-values are reported up to ϵ = 2.2E-16 which is the machine precision.

ATC name	Beta estimate	Standard Error	p-value	t-test	OR estimate
A01A	0.51	0.11	3.5E-6	4.64	1.67
A02B	0.70	0.02	$< \epsilon$	38.52	2.02
A03A	0.83	0.06	$< \epsilon$	13.86	2.29
A03F	1.07	0.04	$< \epsilon$	28.70	2.91
A04A	0.68	0.09	3.4E-13	7.28	1.97
A05A	0.24	0.24	0.31	1.00	1.28
A06A	0.98	0.02	$< \epsilon$	47.47	2.67
A07A	0.72	0.08	$< \epsilon$	9.52	2.06
A07D	0.77	0.11	3.0E-13	7.29	2.15
A07E	0.39	0.09	1.2E-05	4.38	1.47
A09A	1.38	0.12	$< \epsilon$	11.65	3.96
A10A	0.54	0.04	$< \epsilon$	12.29	1.72
A10B	0.19	0.03	2.2E-08	5.60	1.21
A11C	0.71	0.03	$< \epsilon$	24.31	2.04
A11D	1.54	0.21	2.0E-13	7.35	4.65
A12A	0.66	0.04	$< \epsilon$	17.85	1.94
A12B	0.98	0.11	$< \epsilon$	9.29	2.67
B01A	0.51	0.02	$< \epsilon$	22.08	1.66
B02A	0.21	0.24	0.37	0.90	1.24
B02B	0.63	0.15	3.4E-5	4.14	1.89
B03A	0.58	0.05	$< \epsilon$	11.17	1.78
B03B	0.57	0.04	$< \epsilon$	12.87	1.78
B03X	0.15	0.21	0.47	0.72	1.16
B05B	0.74	0.11	9.8E-11	6.47	2.10
C01A	0.44	0.08	9.6E-08	5.33	1.56
C01B	0.35	0.09	1.0E-4	3.89	1.42
C01C	0.05	0.17	0.78	0.28	1.05
C01D	0.25	0.05	7.8E-7	4.94	1.28
C01E	0.83	0.20	5.1E-5	4.05	2.29
C02A	0.75	0.23	1.1E-3	3.26	2.11
C02C	0.11	0.13	0.38	0.87	1.12
C03A	0.14	0.04	3.0E-4	3.62	1.15
C03B	-0.43	0.14	2.0E-3	-3.10	0.65
C03C	0.59	0.04	$< \epsilon$	14.47	1.81
C03D	0.51	0.06	$< \epsilon$	8.32	1.67
C03E	0.05	0.12	0.66	0.44	1.06
C05A	0.45	0.07	3.4E-10	6.28	1.57
C07A	0.34	0.02	$< \epsilon$	14.29	1.40
C07B	0.02	0.16	0.89	0.14	1.02
C07C	-0.28	0.28	0.31	-1.02	0.75
C08C	0.31	0.03	$< \epsilon$	9.53	1.36
C08D	0.45	0.07	2.1E-10	6.35	1.57
C09A	0.25	0.03	$< \epsilon$	8.67	1.28
C09B	-0.05	0.08	0.48	-0.70	0.95
C09C	0.21	0.04	1.7E-09	6.03	1.24
C09D	0.05	0.05	0.36	0.91	1.05
C09X	0.28	0.24	0.25	1.14	1.32

ATC name	Beta estimate	Standard Error	p-value	t-test	OR estimate
C10A	0.23	0.02	$< \epsilon$	9.89	1.25
C10B	0.24	0.12	4.1E-2	2.03	1.28
D01A	0.31	0.04	1.3E-15	7.99	1.36
D01B	0.21	0.09	2.1E-2	2.31	1.23
D02A	0.38	0.03	$< \epsilon$	11.81	1.46
D02B	0.25	0.21	0.24	1.19	1.29
D04A	0.63	0.21	2.7E-3	3.00	1.87
D05A	0.06	0.10	0.58	0.56	1.06
D05B	0.09	0.28	0.76	0.31	1.09
D06A	0.48	0.04	$< \epsilon$	12.60	1.62
D06B	0.52	0.06	$< \epsilon$	8.79	1.68
D07A	0.23	0.02	$< \epsilon$	9.76	1.26
D07C	0.47	0.29	0.10	1.63	1.60
D07X	0.27	0.04	7.6E-11	6.51	1.31
D08A	0.54	0.32	9.0E-2	1.70	1.71
D10A	0.47	0.07	1.3E-12	7.09	1.61
D10B	0.69	0.17	8.3E-05	3.94	1.99
D11A	0.13	0.08	0.11	1.60	1.14
G01A	0.31	0.06	9.5E-07	4.90	1.36
G02B	0.10	0.20	0.62	0.49	1.10
G02C	0.39	0.32	0.22	1.24	1.48
G03A	0.13	0.06	3.5E-2	2.10	1.14
G03B	0.65	0.14	2.5E-06	4.70	1.91
G03C	0.62	0.07	2.2E-16	8.28	1.86
G03D	0.42	0.10	5.8E-05	4.02	1.52
G03F	1.07	0.13	$< \epsilon$	8.44	2.93
G03H	0.65	0.11	7.9E-9	5.77	1.91
G04B	0.70	0.06	$< \epsilon$	12.13	2.02
G04C	0.53	0.04	$< \epsilon$	13.57	1.70
H01B	0.87	0.16	9.2E-8	5.34	2.38
H02A	0.55	0.03	$< \epsilon$	19.26	1.73
H02B	0.29	0.19	0.13	1.52	1.34
H03A	0.46	0.04	$< \epsilon$	11.28	1.59
H03B	0.23	0.20	0.25	1.15	1.25
H04A	0.77	0.15	2.0E-7	5.20	2.16
J01A	0.36	0.03	$< \epsilon$	11.81	1.43
J01C	0.48	0.02	$< \epsilon$	23.00	1.61
J01D	1.19	0.14	$< \epsilon$	8.40	3.29
J01E	0.64	0.06	$< \epsilon$	11.52	1.90
J01F	0.43	0.04	$< \epsilon$	11.72	1.53
J01M	0.59	0.04	$< \epsilon$	15.20	1.80
J01X	0.46	0.04	$< \epsilon$	12.20	1.58
J02A	0.52	0.06	$< \epsilon$	8.96	1.69
J04A	0.42	0.29	0.14	1.46	1.53
J05A	0.83	0.06	$< \epsilon$	13.48	2.29
J07A	0.46	0.07	8.3E-12	6.83	1.59
J07B	0.59	0.27	2.8E-2	2.19	1.80
L01B	0.00	0.12	0.99	0.01	1.00
L02A	0.50	0.13	6.3E-5	4.00	1.66
L02B	0.44	0.10	3.0E-5	4.17	1.55
L03A	0.46	0.20	2.0E-2	2.32	1.58
L04A	0.10	0.08	0.21	1.25	1.11
M01A	0.42	0.02	$< \epsilon$	22.81	1.52
M03B	1.42	0.09	$< \epsilon$	15.81	4.14
M04A	0.22	0.06	1.4E-4	3.80	1.25
M05B	0.57	0.05	$< \epsilon$	10.63	1.77
N01B	0.63	0.06	$< \epsilon$	11.18	1.88
N02A	1.00	0.02	$< \epsilon$	45.82	2.71
N02B	0.75	0.04	$< \epsilon$	18.55	2.12
N02C	0.54	0.05	$< \epsilon$	10.19	1.71
N03A	1.44	0.03	$< \epsilon$	56.92	4.20
N04A	1.99	0.05	$< \epsilon$	36.77	7.31
N04B	0.73	0.07	$< \epsilon$	9.97	2.08
N05A	2.17	0.02	$< \epsilon$	116.49	8.76
N05B	2.18	0.02	$< \epsilon$	117.70	8.87
N05C	2.08	0.02	$< \epsilon$	93.82	8.03
N06A	1.83	0.02	$< \epsilon$	115.65	6.23
N06B	0.86	0.05	$< \epsilon$	18.79	2.36
N06D	0.13	0.15	0.38	0.87	1.14
N07A	0.87	0.24	3.4E-4	3.58	2.39
N07B	1.68	0.04	$< \epsilon$	42.05	5.34
N07C	0.39	0.08	3.07E-6	4.67	1.48
N07X	1.34	0.26	2.07E-7	5.19	3.83
P01A	0.45	0.07	9.71E-10	6.11	1.57
P01B	0.36	0.13	5.6E-3	2.77	1.44
P03A	0.18	0.23	0.44	0.77	1.19

ATC name	Beta estimate	Standard Error	p-value	t-test	OR estimate
R01A	0.16	0.03	3.2E-9	5.92	1.17
R03A	0.46	0.02	$< \epsilon$	18.60	1.58
R03B	0.50	0.03	$< \epsilon$	15.47	1.64
R03D	0.50	0.10	6.4E-7	4.98	1.65
R05C	0.78	0.27	3.5E-3	2.92	2.18
R05D	0.24	0.04	8.9E-11	6.48	1.27
R06A	0.55	0.02	$< \epsilon$	22.50	1.73
S01A	0.31	0.04	2.2E-16	8.27	1.36
S01B	0.27	0.06	2.0E-6	4.76	1.32
S01C	0.32	0.06	8.7E-8	5.35	1.38
S01E	0.41	0.05	4.4E-14	7.55	1.50
S01F	0.26	0.19	0.18	1.33	1.29
S01G	-0.03	0.06	0.64	-0.47	0.97
S01X	0.46	0.03	$< \epsilon$	13.25	1.59
S02A	0.46	0.11	3.4E-5	4.14	1.58
S02C	0.36	0.04	$< \epsilon$	8.77	1.43
V01A	-0.08	0.25	0.74	-0.33	0.92
V03A	0.42	0.19	2.8E-2	2.19	1.51
V07A	0.91	0.24	1.7E-4	3.76	2.49

5.A.4 Results conditional logistic regression for sensitivity check sex

Results of the conditional logistic regression models. For each model only the statistics (Beta, Standard Error) for the parameter corresponding to the medication class and the interaction term of the medication class and 'being female' are reported, reference is being male.

ATC Name	Beta Estimate	Standard Error	Beta Estimate x Female	Standard Error x Female
A01A	0.45	0.14	0.16	0.23
A02B	0.60	0.02	0.29	0.04
A03A	0.77	0.09	0.10	0.12
A03F	0.87	0.06	0.35	0.08
A04A	0.37	0.14	0.60	0.19
A05A	0.22	0.33	0.04	0.49
A06A	0.89	0.03	0.23	0.04
A07A	0.50	0.12	0.43	0.15
A07D	0.75	0.13	0.04	0.22
A07E	0.44	0.11	-0.17	0.19
A09A	1.15	0.16	0.65	0.24
A10A	0.51	0.05	0.10	0.10
A10B	0.20	0.04	-0.03	0.08
A11C	0.61	0.04	0.23	0.06
A11D	1.33	0.26	0.78	0.44
A12A	0.60	0.06	0.11	0.08
A12B	0.80	0.16	0.37	0.21
B01A	0.49	0.03	0.06	0.05
B02A	0.81	0.41	-0.80	0.50
B02B	0.50	0.19	0.40	0.32
B03A	0.73	0.08	-0.28	0.10
B03B	0.47	0.06	0.21	0.09
B03X	-0.24	0.32	0.84	0.42
B05B	0.60	0.15	0.35	0.23
C01A	0.58	0.09	-0.54	0.21
C01B	0.29	0.11	0.24	0.20
C01C	-0.49	0.30	0.97	0.37
C01D	0.26	0.06	-0.05	0.12
C01E	0.86	0.24	-0.11	0.47
C02A	0.86	0.35	-0.18	0.47
C02C	0.03	0.16	0.29	0.28
C03A	0.13	0.05	0.02	0.08
C03B	-0.47	0.18	0.11	0.28

ATC Name	Beta Estimate	Standard Error	Beta Estimate x Female	Standard Error x Female
C03C	0.58	0.05	0.04	0.08
C03D	0.56	0.07	-0.15	0.14
C03E	0.13	0.17	-0.14	0.25
C05A	0.39	0.10	0.15	0.14
C07A	0.34	0.03	-0.00	0.05
C07B	0.17	0.19	-0.43	0.35
C07C	-0.15	0.35	-0.31	0.57
C08C	0.25	0.04	0.18	0.07
C08D	0.43	0.09	0.05	0.15
C09A	0.26	0.03	-0.04	0.07
C09B	-0.00	0.09	-0.20	0.18
C09C	0.19	0.04	0.06	0.07
C09D	0.07	0.07	-0.06	0.12
C09X	0.22	0.30	0.17	0.51
C10A	0.19	0.03	0.12	0.05
C10B	-0.03	0.16	0.85	0.24
D01A	0.32	0.05	-0.03	0.08
D01B	0.08	0.11	0.38	0.19
D02A	0.31	0.04	0.15	0.06
D02B	0.32	0.32	-0.12	0.43
D04A	0.77	0.27	-0.34	0.43
D05A	-0.00	0.12	0.20	0.22
D05B	-0.39	0.41	1.23	0.56
D06A	0.40	0.05	0.21	0.08
D06B	0.51	0.08	0.03	0.12
D07A	0.22	0.03	0.04	0.05
D07C	0.33	0.38	0.39	0.59
D07X	0.22	0.05	0.13	0.09
D08A	0.71	0.35	-0.68	0.79
D10A	0.42	0.09	0.12	0.13
D10B	0.79	0.21	-0.34	0.39
D11A	0.11	0.11	0.06	0.17
G01A	-0.12	0.71	0.43	0.71
G02B	-12.75	4213.95	12.85	4213.95
G02C	0.99	0.38	-1.31	0.69
G03A	-15.36	4392.37	15.50	4392.37
G03B	0.64	0.14	0.28	1.01
G03C	1.76	0.35	-1.18	0.36
G03D	-15.73	5734.52	16.15	5734.52
G03F	-12.52	4785.90	13.60	4785.90
G03H	0.91	0.26	-0.31	0.29
G04B	0.68	0.07	0.06	0.12
G04C	0.51	0.04	0.70	0.22
H01B	0.96	0.20	-0.24	0.34
H02A	0.47	0.04	0.19	0.06
H02B	-0.25	0.33	0.98	0.41
H03A	0.40	0.07	0.10	0.09
H03B	0.14	0.33	0.14	0.41
H04A	0.60	0.19	0.48	0.30
J01A	0.29	0.04	0.19	0.06
J01C	0.45	0.03	0.08	0.04
J01D	1.27	0.16	-0.28	0.32
J01E	0.52	0.09	0.21	0.11
J01F	0.41	0.05	0.05	0.07
J01M	0.54	0.05	0.13	0.08
J01X	0.53	0.09	-0.09	0.10
J02A	0.46	0.09	0.12	0.12
J04A	0.47	0.33	-0.17	0.67
J05A	0.92	0.07	-0.31	0.14
J07A	0.42	0.08	0.15	0.15
J07B	0.62	0.32	-0.10	0.59
L01B	-0.05	0.15	0.15	0.24
L02A	0.55	0.14	-0.20	0.31
L02B	0.68	0.16	-0.40	0.21
L03A	0.51	0.27	-0.12	0.39
L04A	0.12	0.10	-0.04	0.17
M01A	0.36	0.02	0.17	0.04
M03B	1.27	0.12	0.42	0.18
M04A	0.20	0.06	0.18	0.17
M05B	0.51	0.09	0.09	0.11
N01B	0.55	0.08	0.19	0.11
N02A	0.88	0.03	0.29	0.04
N02B	0.63	0.06	0.26	0.08
N02C	0.49	0.09	0.07	0.11
N03A	1.22	0.03	0.50	0.05
N04A	1.78	0.07	0.54	0.11

ATC Name	Beta Estimate	Standard Error	Beta Estimate x Female	Standard Error x Female
N04B	0.62	0.10	0.29	0.15
N05A	2.00	0.02	0.47	0.04
N05B	2.03	0.02	0.39	0.04
N05C	1.86	0.03	0.50	0.04
N06A	1.71	0.02	0.32	0.03
N06B	0.83	0.05	0.09	0.10
N06D	0.12	0.18	0.02	0.31
N07A	0.59	0.35	0.62	0.49
N07B	1.51	0.05	0.55	0.08
N07C	0.42	0.12	-0.06	0.17
N07X	0.69	0.45	1.26	0.55
P01A	0.48	0.12	-0.04	0.15
P01B	0.41	0.19	-0.08	0.26
P03A	0.23	0.27	-0.19	0.52
R01A	0.15	0.03	0.03	0.06
R03A	0.37	0.03	0.21	0.05
R03B	0.42	0.04	0.19	0.07
R03D	0.32	0.15	0.38	0.20
R05C	0.96	0.29	-0.87	0.76
R05D	0.15	0.05	0.20	0.07
R06A	0.44	0.03	0.25	0.05
S01A	0.25	0.05	0.16	0.08
S01B	0.23	0.07	0.11	0.12
S01C	0.24	0.08	0.23	0.12
S01E	0.46	0.06	-0.16	0.12
S01F	0.20	0.24	0.17	0.41
S01G	-0.10	0.09	0.15	0.13
S01X	0.34	0.05	0.24	0.07
S02A	0.26	0.15	0.49	0.22
S02C	0.31	0.05	0.13	0.09
V01A	0.13	0.28	-0.81	0.64
V03A	0.43	0.21	-0.07	0.46
V07A	0.43	0.38	1.06	0.49

5.A.5 Results conditional logistic regression for sensitivity check age

Results of the conditional logistic regression models. For each model only the statistics (Beta, Standard Error) for the parameter corresponding to the medication class and the interaction terms of the medication class with the age groups are reported, reference is 60+.

ATC Name	Beta Estimate	Standard Error	Beta Estimate x 10-30 years old	Standard Error x 10-30 years old	Beta Estimate x 30-60 years old	Standard Error x 30-60 years old
A01A	0.48	0.18	-0.68	0.40	0.28	0.24
A02B	0.63	0.03	0.27	0.09	0.12	0.04
A03A	0.92	0.10	-0.01	0.20	-0.17	0.13
A03F	0.96	0.06	-0.06	0.15	0.23	0.08
A04A	0.53	0.15	0.75	0.32	0.19	0.20
A05A	0.39	0.33	-0.11	1.05	-0.31	0.50
A06A	0.96	0.03	0.05	0.08	0.03	0.04
A07A	0.64	0.12	-0.59	0.43	0.20	0.16
A07D	0.75	0.14	-17.07	2934.21	0.13	0.21
A07E	0.38	0.14	-0.46	0.47	0.05	0.18
A09A	1.27	0.17	-0.02	0.73	0.22	0.24
A10A	0.42	0.06	0.07	0.26	0.29	0.09
A10B	0.17	0.04	1.17	0.36	0.03	0.07
A11C	0.62	0.04	0.48	0.11	0.11	0.06
A11D	1.96	0.38	0.61	0.80	-0.65	0.46
A12A	0.61	0.05	0.12	0.26	0.14	0.08
A12B	0.50	0.17	1.18	0.53	0.99	0.22

ATC Name	Beta Estimate	Standard Error	Beta Estimate x 10-30 years old	Standard Error x 10-30 years old	Beta Estimate x 30-60 years old	Standard Error x 30-60 years old
B01A	0.47	0.03	0.41	0.17	0.07	0.05
B02A	0.65	0.50	0.44	0.71	-0.75	0.59
B02B	0.44	0.19	0.81	1.02	0.70	0.33
B03A	0.57	0.08	0.32	0.17	-0.05	0.11
B03B	0.48	0.07	0.03	0.24	0.20	0.09
B03X	-0.09	0.26	-16.23	8262.68	1.01	0.44
B05B	-0.02	0.22	1.40	0.44	1.31	0.26
C01A	0.41	0.09	-15.32	6129.31	0.33	0.25
C01B	0.31	0.11	0.91	1.01	0.15	0.21
C01C	-0.06	0.33	0.54	0.47	-0.02	0.42
C01D	0.19	0.06	1.72	1.00	0.21	0.11
C01E	0.67	0.28	-15.61	5785.70	0.41	0.41
C02A	0.48	0.45	0.32	0.84	0.41	0.53
C02C	0.02	0.16	1.44	1.01	0.26	0.28
C03A	0.10	0.05	-0.11	1.00	0.09	0.08
C03B	-0.42	0.17	-12.31	1596.26	-0.01	0.30
C03C	0.48	0.05	0.35	0.71	0.46	0.09
C03D	0.44	0.07	-16.95	5693.82	0.30	0.14
C03E	-0.12	0.15	-13.78	4185.92	0.64	0.26
C05A	0.60	0.12	-0.67	0.33	-0.17	0.15
C07A	0.21	0.03	0.50	0.14	0.27	0.05
C07B	-0.00	0.20	-13.77	5652.59	0.06	0.33
C07C	-0.13	0.30	-10.34	2165.16	-0.73	0.77
C08C	0.27	0.04	-0.04	0.58	0.11	0.07
C08D	0.29	0.09	1.28	0.46	0.42	0.15
C09A	0.27	0.04	-0.23	0.50	-0.05	0.06
C09B	-0.02	0.09	-15.59	4935.02	-0.11	0.17
C09C	0.13	0.04	-0.15	0.71	0.22	0.07
C09D	0.04	0.07	-16.20	5807.15	0.03	0.12
C09X	0.26	0.30	-12.30	2604.17	0.07	0.51
C10A	0.17	0.03	-0.08	0.45	0.13	0.05
C10B	0.27	0.15	-15.54	5468.17	-0.05	0.25
D01A	0.34	0.06	-0.20	0.14	-0.03	0.08
D01B	0.39	0.17	-0.09	0.31	-0.29	0.20
D02A	0.43	0.05	-0.13	0.11	-0.08	0.07
D02B	0.30	0.33	-0.02	0.78	-0.09	0.45
D04A	0.57	0.33	0.27	0.67	0.05	0.45
D05A	0.34	0.14	-0.00	0.38	-0.58	0.21
D05B	0.17	0.45	-15.43	2883.95	-0.06	0.57
D06A	0.45	0.06	-0.09	0.12	0.09	0.08
D06B	0.58	0.10	0.12	0.18	-0.15	0.13
D07A	0.35	0.04	-0.14	0.08	-0.20	0.05
D07C	0.45	0.45	-0.15	1.10	0.08	0.61
D07X	0.45	0.06	-0.62	0.17	-0.23	0.09
D08A	-0.51	0.71	-6.78	78.20	1.78	0.79
D10A	0.71	0.17	-0.38	0.21	-0.18	0.20
D10B	1.04	0.71	-0.35	0.74	-0.42	0.78
D11A	0.16	0.14	-0.22	0.30	-0.01	0.18
G01A	0.45	0.18	-0.25	0.23	-0.13	0.20
G02B	-10.74	2371.59	11.02	2371.59	10.67	2371.59
G02C	0.44	0.71	0.32	1.00	-0.16	0.82
G03A	1.36	1.00	-1.38	1.00	-1.08	1.00
G03B	0.32	0.28	1.31	0.42	0.30	0.33
G03C	0.53	0.11	0.79	0.35	0.12	0.15
G03D	1.12	0.58	-0.73	0.62	-0.72	0.59
G03F	0.80	0.35	0.22	1.06	0.33	0.38
G03H	1.07	0.27	-0.65	0.33	-0.36	0.31
G04B	0.56	0.08	0.28	0.34	0.37	0.12
G04C	0.55	0.05	0.77	0.31	-0.13	0.09
H01B	0.82	0.32	-0.94	0.66	0.31	0.37
H02A	0.46	0.04	0.28	0.14	0.17	0.06
H02B	0.27	0.28	-15.79	4378.00	0.08	0.39
H03A	0.42	0.06	0.27	0.22	0.07	0.08
H03B	0.30	0.27	-0.23	1.04	-0.16	0.40
H04A	0.64	0.24	-0.89	0.75	0.38	0.31
J01A	0.37	0.05	-0.07	0.12	-0.00	0.06
J01C	0.42	0.04	-0.01	0.07	0.11	0.05
J01D	0.87	0.29	0.46	0.43	0.44	0.35
J01E	0.55	0.08	-0.47	0.28	0.27	0.11
J01F	0.34	0.07	-0.08	0.13	0.18	0.08
J01M	0.53	0.05	0.13	0.17	0.11	0.08
J01X	0.46	0.06	-0.10	0.13	0.03	0.08
J02A	0.75	0.12	-0.44	0.20	-0.26	0.14
J04A	0.16	0.58	0.67	0.91	0.30	0.69

ATC Name	Beta Estimate	Standard Error	Beta Estimate x 10-30 years old	Standard Error x 10-30 years old	Beta Estimate x 30-60 years old	Standard Error x 30-60 years old
J05A	0.63	0.12	0.04	0.25	0.30	0.14
J07A	0.46	0.12	-0.27	0.25	0.06	0.15
J07B	-0.95	1.00	1.47	1.23	1.96	1.04
L01B	0.03	0.14	-17.71	4820.60	-0.00	0.25
L02A	0.48	0.15	0.37	0.72	0.06	0.29
L02B	0.46	0.13	-16.29	6310.76	-0.07	0.23
L03A	0.06	0.45	1.42	0.73	0.42	0.51
L04A	0.02	0.13	-0.27	0.47	0.18	0.17
M01A	0.46	0.03	-0.10	0.07	-0.06	0.04
M03B	1.14	0.17	0.39	0.42	0.41	0.20
M04A	0.20	0.07	1.20	0.42	0.03	0.12
M05B	0.46	0.06	0.25	0.71	0.42	0.12
N01B	0.52	0.09	0.30	0.19	0.15	0.12
N02A	0.91	0.04	0.19	0.10	0.13	0.05
N02B	0.70	0.06	0.05	0.20	0.10	0.08
N02C	0.75	0.12	0.08	0.19	-0.31	0.13
N03A	1.20	0.04	0.80	0.10	0.31	0.06
N04A	1.62	0.15	1.19	0.20	0.32	0.16
N04B	0.66	0.09	-0.16	1.00	0.22	0.15
N05A	1.88	0.04	0.78	0.06	0.32	0.04
N05B	2.01	0.03	0.58	0.07	0.20	0.04
N05C	1.96	0.04	0.54	0.09	0.14	0.05
N06A	1.83	0.03	0.25	0.06	-0.04	0.03
N06B	1.21	0.18	-0.73	0.20	-0.19	0.19
N06D	0.09	0.16	-13.10	2239.32	0.39	0.47
N07A	0.61	0.33	1.50	1.05	0.60	0.50
N07B	1.66	0.10	0.76	0.17	-0.04	0.11
N07C	0.42	0.10	-0.91	1.01	-0.07	0.18
N07X	1.23	0.45	1.07	0.84	0.04	0.57
P01A	0.39	0.18	-0.07	0.26	0.10	0.20
P01B	0.46	0.18	0.87	0.53	-0.32	0.28
P03A	-0.61	1.00	1.27	1.04	0.43	1.07
R01A	0.25	0.05	-0.22	0.09	-0.11	0.06
R03A	0.38	0.04	-0.05	0.09	0.16	0.05
R03B	0.48	0.04	-0.32	0.16	0.07	0.07
R03D	0.51	0.16	0.10	0.37	-0.04	0.21
R05C	0.97	0.29	0.14	1.04	-1.39	1.04
R05D	0.28	0.06	0.01	0.15	-0.09	0.08
R06A	0.48	0.05	-0.07	0.08	0.12	0.06
S01A	0.42	0.06	-0.27	0.14	-0.16	0.08
S01B	0.26	0.07	-0.24	0.39	0.08	0.13
S01C	0.28	0.08	-0.45	0.39	0.17	0.13
S01E	0.42	0.06	-0.03	0.50	-0.06	0.13
S01F	-0.09	0.32	0.98	0.66	0.56	0.41
S01G	0.10	0.12	-0.18	0.20	-0.17	0.15
S01X	0.50	0.05	0.10	0.17	-0.12	0.07
S02A	0.25	0.20	0.16	0.43	0.34	0.25
S02C	0.39	0.07	-0.03	0.15	-0.05	0.09
V01A	-0.43	1.00	0.13	1.16	0.44	1.04
V03A	0.10	0.27	-18.29	14713.26	0.83	0.38
V07A	0.70	0.38	0.82	0.80	0.31	0.52

5.A.6 Results conditional logistic regression for sensitivity check mental healthcare

Results of the conditional logistic regression models. For each model only the statistics (Beta, Standard Error) for the parameter corresponding to the medication class and the interaction term of the medication class with 'receiving mental healthcare' are reported, reference is not receiving mental healthcare.

ATC Name	Beta Estimate	Standard Error	Beta Estimate x Mental Healthcare	Standard Error x Mental Healthcare
A01A	0.13	0.16	1.01	0.22
A02B	0.38	0.02	0.79	0.04
A03A	0.61	0.09	0.47	0.12
A03F	0.81	0.05	0.60	0.07
A04A	0.49	0.12	0.56	0.19
A05A	0.20	0.29	0.16	0.53
A06A	0.66	0.03	0.79	0.04
A07A	0.31	0.12	0.90	0.15
A07D	0.66	0.13	0.32	0.22
A07E	0.26	0.11	0.46	0.19
A09A	1.25	0.15	0.36	0.24
A10A	0.39	0.05	0.54	0.09
A10B	0.02	0.04	0.65	0.07
A11C	0.32	0.04	0.87	0.06
A11D	1.04	1.00	0.53	1.02
A12A	0.37	0.05	0.81	0.07
A12B	0.58	0.15	0.97	0.21
B01A	0.26	0.03	0.83	0.05
B02A	0.05	0.33	0.37	0.47
B02B	0.32	0.20	1.05	0.31
B03A	0.39	0.07	0.49	0.10
B03B	0.32	0.06	0.70	0.09
B03X	0.01	0.24	0.69	0.48
B05B	0.50	0.15	0.70	0.23
C01A	0.29	0.10	0.84	0.19
C01B	0.23	0.10	0.64	0.21
C01C	-0.18	0.23	0.62	0.35
C01D	0.08	0.06	0.73	0.11
C01E	0.57	0.27	0.79	0.41
C02A	0.46	0.33	0.66	0.46
C02C	-0.08	0.16	0.86	0.28
C03A	-0.08	0.05	0.91	0.08
C03B	-0.56	0.16	0.70	0.32
C03C	0.43	0.05	0.65	0.09
C03D	0.34	0.07	0.71	0.13
C03E	-0.14	0.15	0.87	0.27
C05A	0.29	0.10	0.41	0.14
C07A	0.03	0.03	0.98	0.05
C07B	-0.13	0.19	0.82	0.37
C07C	-0.32	0.30	0.28	0.77
C08C	0.08	0.04	0.93	0.07
C08D	0.30	0.08	0.60	0.15
C09A	0.02	0.03	0.90	0.06
C09B	-0.20	0.09	0.77	0.18
C09C	0.00	0.04	0.88	0.08
C09D	-0.07	0.06	0.65	0.13
C09X	0.02	0.30	1.04	0.51
C10A	-0.06	0.03	0.99	0.05
C10B	0.15	0.14	0.47	0.28
D01A	0.11	0.05	0.58	0.08
D01B	0.09	0.11	0.36	0.19
D02A	0.12	0.04	0.74	0.06
D02B	0.09	0.28	0.46	0.43
D04A	0.44	0.28	0.48	0.42
D05A	-0.13	0.13	0.63	0.21
D05B	-0.14	0.35	0.75	0.57
D06A	0.25	0.05	0.65	0.08
D06B	0.31	0.08	0.59	0.12
D07A	0.03	0.03	0.63	0.05
D07C	-0.13	0.45	1.49	0.59
D07X	0.05	0.05	0.63	0.08
D08A	0.47	0.38	0.24	0.69
D10A	0.33	0.09	0.36	0.13
D10B	0.62	0.22	0.19	0.36
D11A	-0.20	0.12	0.88	0.17
G01A	0.08	0.10	0.42	0.13
G02B	-0.03	0.30	0.24	0.40
G02C	-0.35	0.58	1.37	0.69
G03A	-0.22	0.09	0.71	0.13
G03B	0.29	0.20	0.86	0.28
G03C	0.37	0.10	0.61	0.15
G03D	0.12	0.16	0.59	0.21
G03F	0.89	0.18	0.38	0.25
G03H	0.69	0.15	-0.10	0.23
G04B	0.49	0.07	0.61	0.12

ATC Name	Beta Estimate	Standard Error	Beta Estimate x Mental Healthcare	Standard Error x Mental Healthcare
G04C	0.34	0.05	0.83	0.08
H01B	0.76	0.21	0.28	0.33
H02A	0.31	0.04	0.73	0.06
H02B	0.03	0.25	0.83	0.39
H03A	0.16	0.06	0.78	0.08
H03B	-0.08	0.27	0.84	0.39
H04A	0.43	0.20	0.92	0.30
J01A	0.11	0.04	0.72	0.06
J01C	0.19	0.03	0.81	0.04
J01D	0.96	0.19	0.60	0.28
J01E	0.39	0.07	0.70	0.11
J01F	0.27	0.05	0.45	0.07
J01M	0.39	0.05	0.65	0.08
J01X	0.15	0.05	0.76	0.07
J02A	0.45	0.08	0.19	0.12
J04A	-0.33	0.50	1.57	0.61
J05A	0.56	0.09	0.68	0.12
J07A	0.17	0.09	0.84	0.14
J07B	0.47	0.33	0.38	0.56
L01B	-0.15	0.14	0.70	0.26
L02A	0.41	0.15	0.44	0.29
L02B	0.37	0.12	0.25	0.24
L03A	0.51	0.24	-0.18	0.43
L04A	-0.02	0.10	0.49	0.18
M01A	0.14	0.02	0.76	0.04
M03B	1.22	0.13	0.41	0.18
M04A	0.09	0.07	0.72	0.14
M05B	0.33	0.07	0.86	0.11
N01B	0.35	0.08	0.74	0.11
N02A	0.78	0.03	0.55	0.04
N02B	0.51	0.06	0.61	0.08
N02C	0.25	0.08	0.62	0.11
N03A	0.99	0.04	0.82	0.05
N04A	1.74	0.17	0.28	0.18
N04B	0.61	0.09	0.37	0.15
N05A	1.93	0.04	0.34	0.04
N05B	1.81	0.03	0.56	0.04
N05C	1.87	0.04	0.31	0.05
N06A	1.63	0.02	0.39	0.03
N06B	0.84	0.09	0.02	0.10
N06D	0.41	0.17	-0.88	0.35
N07A	0.70	0.30	0.56	0.51
N07B	1.67	0.08	0.01	0.09
N07C	0.21	0.11	0.60	0.17
N07X	0.93	0.38	1.00	0.52
P01A	0.27	0.10	0.40	0.15
P01B	0.30	0.15	0.26	0.29
P03A	-0.29	0.35	1.02	0.47
R01A	-0.05	0.03	0.60	0.05
R03A	0.21	0.03	0.70	0.05
R03B	0.25	0.04	0.74	0.07
R03D	0.32	0.13	0.49	0.20
R05C	0.68	0.32	0.40	0.59
R05D	0.02	0.05	0.67	0.08
R06A	0.02	0.04	1.17	0.05
S01A	0.14	0.05	0.58	0.08
S01B	0.08	0.07	0.85	0.12
S01C	0.10	0.07	0.88	0.13
S01E	0.26	0.06	0.77	0.12
S01F	-0.01	0.25	0.84	0.39
S01G	-0.25	0.08	0.62	0.13
S01X	0.25	0.04	0.67	0.07
S02A	0.11	0.15	0.92	0.22
S02C	0.14	0.05	0.67	0.08
V01A	-0.10	0.30	0.05	0.54
V03A	0.46	0.20	-0.27	0.54
V07A	1.11	0.26	-1.05	0.75

Part II

Dependency and Feature Importance

6

The Berkelmans-Pries dependency function: a generic measure of dependence between random variables

6.1 Introduction

In as early as 1958, Kruskal stated that 'There are infinitely many possible measures of association, and it sometimes seems that almost as many have been proposed at one time or another' [89]. Many years later, even more dependency measures have been suggested. Yet, rather surprisingly, there still does not exist consensus on a general dependency function. Often the statement 'Y is dependent on X' means that Y is not independent of X. However, there are different levels of dependency. For example, random variable (RV) Y can be fully determined by RV X (i.e., $Y(\omega) = f(X(\omega))$ for all $\omega \in \Omega$ (the outcome space) and for a measurable function f), or only partially.

But how should we quantify how much Y is dependent on X?

Based on [16]: G. Berkelmans, J. Pries, R.D. van der Mei, S. Bhulai. The BP dependency function: a generic measure of dependence between random variables. Accepted at Journal of Applied Probability.

Intuitively, and assuming that the dependency measure is normalised to the interval [0,1], one would say that if Y is fully determined by X then the dependency of Y w.r.t. X is as strong as possible, and so the dependency measure should be 1. On the other side of the spectrum, if X and Y are independent, then the dependency measure should be 0; and vice versa, it is desirable that dependence 0 *implies* that X and Y are stochastically independent. Note that the commonly used *Pearson correlation coefficient* does not meet these requirements. In fact, many examples exists where Y is fully determined by X while the correlation is zero.

Taking a step back, why is it actually useful to examine dependencies in a dataset? Measuring dependencies between the variables can lead to critical insights, which will lead to improved data analysis. First of all, it can reveal important explanatory relationships. How do certain variables interact? If catching a specific disease is highly dependent on the feature value of variable X, research should be done to investigate if this information can be exploited to reduce the number of patients with this disease. For example, if hospitalisation time is dependent on a healthy lifestyle, measures can be taken to try to improve the overall fitness of a population. Dependencies can therefore function as an actionable steering rod. It is however important to keep in mind that dependency does not always mean causality. Dependency relations can also occur due to mere coincidence or as a byproduct of another process.

Dependencies can also be used for dimensionality reduction. If Y is highly dependent on X, not much information is lost when only X is used in the data-set. In this way, redundant variables or variables that provide little additional information, can be removed to reduce the dimensionality of the data-set. With fewer dimensions, models can be trained more efficiently.

In these situations a dependency function can be very useful. However, finding the proper dependency function can be hard, as many attempts have already been made. In fact, most of us have a 'gut feeling' for what a dependency function should entail. To make this feeling more mathematically sound, Rényi [126] proposed a list of ideal properties for a dependency function. A long list of follow-up

papers (see the references in Table 6.1 below) use this list as the basis for a wish list, making only minor changes to it, adding or removing some properties.

In view of the above, the contribution of this paper is threefold:

- We determine a new list of ideal properties for a dependency function;
- We present a new dependency function and show that it fulfills all requirements;
- We provide Python code to determine the dependency function for the discrete and continuous case [28].

The remainder of this paper is organized as follows. In Section 6.2, we summarize which ideal properties have been stated in previous literature. By critically assessing these properties, we derive a new list of ideal properties for a dependency function (see Table 6.2), which lays the foundation for a new search for a general-purpose dependency function. In Section 6.3, the properties are checked for existing methods, and we conclude that there does not yet exist a dependency function that has all desired properties. Faced by this, in Section 6.4 we define a new dependency function and show in Section 6.5 that this function meets all the desired properties. We then propose a possible extension to the notion of conditional dependency in Section 8.1. Finally, Section 6.6 outlines the general findings and addresses possible future research opportunities.

6.2 Desired properties of a dependency function

What properties should an ideal dependency function have? In this section, we summarize previously suggested properties. Often, these characteristics are posed without much argumentation. Therefore, we analyze and discuss which properties are actually ideal and which properties are to be believed not relevant, or even wrong.

In Table 6.1 below, a summary is given of (twenty-two) 'ideal properties' found in previous literature, grouped into five different categories. These properties are denoted by I.1-22. From these

properties we derive a new set of desirable properties denoted by II.1-8, see Table 6.2. Next, we discuss the properties suggested in previous literature and how the new list is derived from them.

Asymmetry (Desired property II.1):
At first glance, it seems obvious that a dependency function should adhere to property I.13 and be symmetric. However, this is a common misconception for the dependency function. Y can be fully dependent on X, but this does not mean that X is fully dependent on Y. Lancaster [93] indirectly touched upon this same point by defining *mutual complete dependence*. First it is stated that Y is *completely dependent* on X if $Y = f(X)$. X and Y are called *mutually completely dependent* if X is completely dependent on Y and vice versa. Thus, this indirectly shows that dependence should not necessarily be symmetric, otherwise the extra definition would be redundant. In [93] the following great asymmetric example was given.

Example 6.2.1. Let $X \sim \mathcal{U}(0,1)$ be uniformly distributed and let $Y = -1$ if $X \leq \frac{1}{2}$ and $Y = 1$ if $X > \frac{1}{2}$.

Then, Y is fully dependent on X, but not vice versa. To drive the point home even more, we give another asymmetric example.

Example 6.2.2. X is uniformly randomly drawn out of $\{1,2,3,4\}$ and $Y := X \mod 2$.

Y is fully dependent on X, because given X the value of Y is deterministically known. On the other hand, X is not completely known given Y. Note that $Y = 1$ still leaves the possibility for $X = 1$ or $X = 3$. Thus, when assessing the dependency between variable X and variable Y, Y is fully dependent on X, whereas X is not fully dependent on Y. In other words, $\mathrm{Dep}(X,Y) \neq \mathrm{Dep}(Y,X)$.

In conclusion, *an ideal dependency function should not always be symmetric*. To emphasise this point even further, we change the notation of the dependency function. Instead of $\mathrm{Dep}(X,Y)$, we will denote $\mathrm{Dep}(Y|X)$ for how much Y is dependent on X. Based by this, property I.13 is changed into II.1.

Range (Desired property II.2):

An ideal dependency function should be scaled to the interval $[0, 1]$. Otherwise, it can be very hard to draw meaningful conclusions from a dependency score without a known maximum or minimum. What would a score of 4.23 mean without any information about the possible range? Therefore, property I.1 is retained. A special note on the range for the well-known *Pearson correlation coefficient* [120], which is $[-1, 1]$: The negative or positive sign denotes the direction of the linear correlation. When examining more complex relationships, it is unclear what 'direction' entails. We believe that a dependency function should measure by *how much* variable Y is dependent on X, and not necessarily in which way. In summary, we require: $0 \leq \mathrm{Dep}(Y|X) \leq 1$.

Independence equals a dependency of 0 (Desired property II.3):
If Y is independent of X, it should hold that the dependency achieves the lowest possible value, namely zero. Otherwise, it is vague what a dependency score lower than the dependency between two independent variables means. A major issue of the commonly used *Pearson correlation coefficient*, is that zero correlation does not imply independence. This makes it complicated to derive conclusions from a correlation score. Furthermore, note that if Y is independent of X, it should automatically hold that X is also independent of Y. In this case, X and Y are independent, because otherwise some dependency relation should exist. Thus, we require: $\mathrm{Dep}(Y|X) = 0 \iff X$ and Y are independent.

Desired property II.4 (Functional dependence equals a dependency of 1):
If Y is strictly dependent on X (and thus fully determined by X), the highest possible value should be attained. It is otherwise unclear what a higher dependency would mean. However, it is too restrictive to demand that the dependency is only 1 if Y is strictly dependent on X. Rényi [126] stated 'It seems at the first sight natural to postulate that $\delta(\xi, \eta) = 1$ only if there is a strict dependence of the mentioned type between ξ and η, but this condition is rather restrictive, and it is better to leave it out'. Take, for example, $Y \sim \mathcal{U}(-1, 1)$ and $X := Y^2$. Knowing X reduces the infinite set of possible values for Y to only two $\left(\pm\sqrt{X}\right)$, whereas it would reduce

to one if Y was fully determined by X. It would be very restrictive to enforce $\mathrm{Dep}\,(Y|X) < 1$, as there is only an infinitesimal difference compared to the strictly dependent case. Summarising, we require: $Y = f(X) \rightarrow \mathrm{Dep}\,(Y|X) = 1$.

Unambiguity (Desired property II.5):
Kruskal [89] once stated 'It is important to recognise that the question 'Which single measure of association should I use?,' is often unimportant. There may be no reason why two or more measures should not be used; the point I stress is that, whichever ones are used, they should have clear-cut population interpretations.' It is very important that a dependency score leaves no room for ambiguity. The results should stroke with our natural expectation. Therefore, we introduce a new requirement based on a simple example: suppose we have a number of independent RVs and observe one of these at random. The dependency of each random variable on the observed variable should be equal to the probability it is picked. More formally, let $Y_1, Y_2, \ldots, Y_N, S$ be independent variables with S a selection variable s.t. $\mathbb{P}(S = i) = p_i$ and $\sum_{i=1}^{N} p_i = 1$. When X is defined as $X = \sum_{i=1}^{N} \mathbb{1}_{S=i} \cdot Y_i$, it should hold that $\mathrm{Dep}\,(Y_i|X) = p_i$ for all $i \in \{1, \ldots, N\}$. Simply said, the dependency function should give desired results in specific situations, where we can argue what the outcome should be. This is one of these cases.

Generally applicable (Desired property II.6):
Our aim is to find a general dependency function, which we denote by $Dep(X|Y)$. This function must be able to handle all kinds of variables: *continuous*, *discrete*, and *categorical* (even nominal). These types of variables occur frequently in a data-set. A general dependency function should be able to measure the dependency of a categorical variable Y on a continuous variable X. Stricter than I.9-12, we want a single dependency function that is applicable to any combination of these variables.

There is one exception to this generality. In the case that Y is almost surely constant it is completely independent as well as completely determined by X. Arguing what the value of a dependency function should be in this case is a bit similar to arguing the value of $\frac{0}{0}$. Therefore, we argue that in this case it should be either undefined

or return some value that represents the fact that Y is almost surely constant (for example -1 since this cannot be normally attained).

Invariance under isomorphisms (Desired property II.7):
Properties I.14-20 discuss when the dependency function should be invariant. Most are only meant for variables with an ordering, as 'strictly increasing', 'translation' and 'scaling' are otherwise ill-defined. As the dependency function should be able to handle nominal variables, we assume that the dependency is invariant under isomorphisms, see II.7. Note that this is a stronger assumption than I.14-20. Compare Example 6.2.2 with the following example.

Example 6.2.3. Let X' be uniformly randomly drawn out of $\{\circ, \triangle, \square, \Diamond\}$ and $Y' = \clubsuit$ if $X' \in \{\circ, \square\}$ and $Y' = \spadesuit$ if $X' \in \{\triangle, \Diamond\}$.

It should hold that $\mathrm{Dep}(Y|X) = \mathrm{Dep}(Y'|X')$ and $\mathrm{Dep}(X|Y) = \mathrm{Dep}(X'|Y')$, as the relationship between the variables is the same (only altered using isomorphisms). So, for any isomorphisms f and g we require $\mathrm{Dep}(g(Y)|f(X)) = \mathrm{Dep}(Y|X)$.

Non-increasing under functions of X (Desired property II.8):
Additionally, $\mathrm{Dep}(Y|X)$ should not increase if a measurable function f is applied to X since any dependence on $f(X)$ corresponds to a dependence on X (but not necessarily the other way around). The information gained from knowing X can only be reduced, never increased by applying a function.

However, though it might be natural to expect the same for functions applied to Y, consider once again Example 6.2.2 (but with X and Y switched around) and the following 2 functions: $f_1(Y) := Y \mod 2$ and $f_2(Y) := \left\lceil \frac{Y}{2} \right\rceil$. Then $f_1(Y)$ is completely predicted by X and should therefore have a dependency of 1 while $f_2(Y)$ is independent of X and should therefore have a dependency of 0. So the dependency should be free to increase or decrease for functions applied to Y. To conclude, for any measurable function f we require: $\mathrm{Dep}(Y|f(X)) \leq \mathrm{Dep}(Y|X)$.

Exclusion of Pearson correlation coefficient as a special

case:
According to properties I.21-22, when X and Y are normally distributed the dependency function should coincide with or be a function of the *Pearson correlation coefficient*. However, these properties lack a good argumentation for why this would be ideal. It is not obvious why this would be a necessary condition. Even more, there are many known problems and pitfalls with the correlation coefficient [55, 79], so it seems undesirable to force an ideal dependency function to reduce to a function of the correlation coefficient, when the variables are normally distributed. This is why we leave these properties out.

6.3 Do existing dependency measures satisfy the desired properties?

In this section, we assess whether existing dependency functions have the properties listed above. In doing so, we limit this section to the most commonly used dependency measures. Table 6.3 shows which properties each investigated measure adheres to.

Although the desired properties listed in Table 6.2 seem not too restrictive, many dependency measures fail to have many of these properties. One of the most commonly used dependency measures, the *Pearson correlation coefficient*, does not even satisfy any one of the desirable properties. Furthermore, almost all measures are not asymmetric. The one measure that comes closes to fulfilling all requirements, is the *uncertainty coefficient* [120].

This is a normalised asymmetric variant of the *mutual information* [120], where the discrete variant is defined as

$$C_{XY} = \frac{I(X,Y)}{H(Y)} = \frac{\sum_{x,y} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)\cdot p_Y(y)}\right)}{-\sum_y p_Y(y)\log(p_Y(y))},$$

where $H(Y)$ is the entropy of Y and $I(X,Y)$ is the mutual information of X and Y.

Table 6.1: A summary of desirable properties for a dependency function stated in previous literature.

Property group	Property	Article(s)
Range	I.1. $0 \leq \text{Dep}(X,Y) \leq 1$	[4, 55, 69, 71, 80, 126, 127, 148, 150]
	I.2. $\text{Dep}(X,Y) = 0 \Leftarrow X$ and Y are independent	[69, 80, 127]
	I.3. $\text{Dep}(X,Y) = 0 \Rightarrow X$ and Y are independent	[150]
	I.4. $\text{Dep}(X,Y) = 0 \Leftrightarrow X$ and Y are independent	[4, 55, 71, 111, 126, 148]
	I.5. $\text{Dep}(X,Y) = 1 \Leftrightarrow Y = LX$ with probability 1, where L is a similarity transformation	[111]
	I.6. $\text{Dep}(X,Y) = 1 \Leftarrow X$ and Y are strictly dependent	[4, 69, 126, 127]
	I.7. $\text{Dep}(X,Y) = 1 \Leftrightarrow X$ and Y are comonotonic or countermonotonic	[55]
	I.8. $\text{Dep}(X,Y) = 1 \Leftrightarrow X$ and Y are strictly dependent	[71]
General	I.9. $\text{Dep}(X,Y)$ is defined for any X, Y where both are not constant	[71, 111, 126]
	I.10. Well-defined for both continuous and discrete variables	[69]
	I.11. Defined for both categorical and continuous variables; and for ordinal categorical variables for which there may be underlying continuous variables	[80]
	I.12. There is a close relationship between the measure for the continuous variables and the measure for the discretization of the variables	[80]
Symmetric	I.13. $\text{Dep}(X,Y) = \text{Dep}(Y,X)$	[4, 55, 126, 127, 148]
Applying function to argument	I.14. $\text{Dep}(f(X), g(Y)) = \text{Dep}(X,Y)$ with f, g strictly monotonic functions	[4]
	I.15. $\text{Dep}(f(X), Y) = \text{Dep}(X,Y)$ with $f : \mathbb{R} \to \mathbb{R}$ strictly monotonic on the range of X	[55]
	I.16. $\text{Dep}(f(X), f(Y)) = \text{Dep}(X,Y)$ with f continuous and strictly increasing	[69, 148]
	I.17. $\text{Dep}(f(X), g(Y)) = \text{Dep}(X,Y)$ if $f(\cdot), g(\cdot)$ map the real axis in a one-to-one way onto itself	[80, 126]
	I.18. $\text{Dep}(X,Y)$ is invariant with respect to all similarity transformations	[111]
	I.19. $\text{Dep}(X,Y)$ is invariant with respect to translation and scaling	[148]
	I.20. $\text{Dep}(X,Y)$ is scale invariant	[150]
Behaviour normal distribution	I.21. $\text{Dep}(X,Y)$ is a function of the Pearson correlation if the joint distribution of X and Y is normal	[4, 69, 150]
	I.22. $\text{Dep}(X,Y) = \lvert\rho(X,Y)\rvert$ if the joint distribution of X and Y is normal, where ρ is the Pearson correlation	[80, 126]

Table 6.2: New list of desirable properties for a dependency function.

Property group	Property
Asymmetric	II.1. There exist RVs X, Y such that $\mathrm{Dep}(Y\|X) \neq \mathrm{Dep}(X\|Y)$.
Intuitive	II.2. $0 \leq \mathrm{Dep}(Y\|X) \leq 1$ for all RVs X and Y. II.3. $\mathrm{Dep}(Y\|X) = 0 \Leftrightarrow X$ and Y are independent. II.4. $\mathrm{Dep}(Y\|X) = 1 \Leftarrow Y$ is strictly dependent on X. II.5. If $Y_1, Y_2, \ldots, Y_N, S$ independent with $\mathbb{P}(S \in [N]) = 1$, $\mathbb{P}(S = i) = p_i$ and $X = Y_S$ then $\mathrm{Dep}(Y_i\|X) = p_i$ must hold.
General	II.6. Applicable for any combination of continuous, discrete and categorical RVs X, Y, where Y is not a.s. constant.
Functions	II.7. $\mathrm{Dep}(g(Y)\|f(X)) = \mathrm{Dep}(Y\|X)$ for any isomorphisms f, g. II.8. $\mathrm{Dep}(Y\|f(X)) \leq \mathrm{Dep}(Y\|X)$ for any measurable function f.

In this chapter we use the following notation: $p_X(x) = \mathbb{P}(X = x)$, $p_Y(y) = \mathbb{P}(Y = y)$, and $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$. In addition, for a set H we define $p_X(H) = \mathbb{P}(X \in H)$ (and similarly for p_Y and $p_{X,Y}$).

However, the *uncertainty coefficient* does not satisfy properties II.5 and II.6. For example, if $Y \sim \mathcal{U}(0, 1)$ is uniformly drawn, the entropy of Y becomes:

$$\begin{aligned} H(Y) &= -\int_0^1 f_Y(y) \ln\left(f_Y(y)\right) dy \\ &= -\int_0^1 1 \cdot \ln(1)\, dy \\ &= 0. \end{aligned}$$

Thus, for any X, the uncertainty coefficient is now undefined (division by zero). Therefore, the uncertainty coefficient is not as generally applicable as property II.6 requires.

Two other measures that satisfy many (but not all) properties are *mutual dependence* [4] and *maximal correlation* [62]. Mutual dependence is defined as the Hellinger distance [73] d_h between the joint distribution and the product of the marginal distributions, defined as follows (cf. [4]):

$$d(X, Y) \triangleq d_h(f_{XY}(x, y), f_X(x) \cdot f_Y(y)). \tag{6.1}$$

Maximal correlation is defined as (cf. [126]):

$$S(X, Y) = \sup_{f,g} R(f(X), g(Y)), \tag{6.2}$$

where R is the Pearson correlation coefficient, and where f, g are Borel-measurable functions, such that $R(f(X), g(Y))$ is defined [126].

Clearly, Equations (6.1) and (6.2) are symmetric. Neither the joint distribution nor the product of the marginal distributions change by switching X and Y. Furthermore, the Pearson correlation coefficient is symmetric, making the maximal correlation also symmetric. Therefore, both measures do not have property II.1.

There are two more measures (one of which is a variation of the other) which satisfy many (but not all) properties, and additionally closely resemble the measure we intend to propose. Namely, the *strong mixing coefficient* [30]

$$\alpha(X,Y) = \sup_{A\in\mathcal{E}_X, B\in\mathcal{E}_Y} \{|\mu_{X,Y}(A\times B) - \mu_X(A)\mu_Y(B)|\},$$

and its relaxation, the *β-mixing coefficient* [30]

$$\beta(X,Y) = \sup\left\{\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}|(\mu_{X,Y}(A_i\times B_j) - \mu_X(A_i)\mu_Y(B_j))|\right\},$$

where the supremum ranges over all finite partitions $(A_1, A_2, \ldots, A_I)$ and $(B_1, B_2, \ldots, B_J)$ of E_X and E_Y with $A_i \in \mathcal{E}_X$ and $B_j \in \mathcal{E}_Y$. However, these measures fail the properties II.1, II.4, and II.5.

6.4 The Berkelmans-Pries dependency function

After devising a new list of ideal properties (see Table 6.2) and showing that these properties are not fulfilled by existing dependency functions (see Table 6.3), we will now introduce a new dependency function that will meet all requirements. Throughout the rest of the thesis, we will refer to this function as the *Berkelmans-Pries (BP)* dependency function.

The key question surely is: What is dependency? Although this question deserves an elaborate philosophical study, we believe that

Table 6.3: Properties of previous dependencies functions (✗= property not satisfied, ✓= property satisfied, for Property 6: ∞ means 'holds in principle but can be infinite').

Measure	Asymmetric	Intuitive				General	Functions	
	II.1	II.2	II.3	II.4	II.5	II.6	II.7	II.8
Pearson correlation coefficient [120]	✗	✗	✗	✗	✗	✗	✗	✗
Spearman's rank correlation coefficient [120]	✗	✗	✗	✗	✗	✗	✗	✗
Kendall rank correlation coefficient [120]	✗	✗	✗	✗	✗	✗	✗	✗
Mutual information [120]	✗	✗	✓	✗	✗	∞	✓	✓
Uncertainty coefficient [120]	✓	✓	✓	✓	✗	✗	✓	✓
Total correlation [162]	✗	✗	✓	✗	✗	∞	✓	✓
Mutual dependence [4]	✗	✓	✓	✗	✗	✓	✓	✓
Δ_{L_1} [35]	✗	✗	✓	✗	✗	✓	✓	✓
Δ_{SD} [35]	✗	✗	✓	✗	✗	✗	✗	✗
Δ_{ST} [35]	✗	✗	✗	✗	✗	✗	✗	✗
Monotone correlation [86]	✗	✓	✓	✗	✗	✗	✗	✗
Maximal correlation [62]	✗	✓	✓	✓	✗	✓	✓	✓
Distance correlation [150]	✗	✓	✓	✗	✗	✗	✗	✗
Maximum canonical correlation (first) [75]	✗	✓	✗	✗	✗	✗	✗	✗
Strong mixing coefficient [30]	✗	✓	✓	✗	✗	✓	✓	✓
β-mixing coefficient [30]	✗	✓	✓	✗	✗	✓	✓	✓

measuring the dependency of Y on X, is essentially measuring how much the distribution of Y changes on average based on the knowledge of X, divided by the maximum possible change. This is the key insight, which the BP dependency function is based on. To measure this, we first have to determine the difference between the distribution of Y *with* and *without* conditioning on the value of X times the probability that X takes on this value in Section 6.4.1. Secondly, we have to measure what the maximum possible change in probability mass is, which is used to properly scale the dependency function and make it asymmetric (see Section 6.4.2).

6.4.1 Definition expected absolute change in distribution

We start by measuring the *expected absolute change in distribution* (UD), which is the difference between the distribution of Y *with* and *without* conditioning on the value of X times the probability that X takes on this value. For discrete RVs, we obtain the following definition.

Definition 6.4.1 (Discrete UD)**.** For any discrete RVs X and Y,

$$\mathrm{UD}\left(X,Y\right) := \sum_{x} p_X(x) \cdot \sum_{y} \left| p_{Y|X=x}(y) - p_Y(y) \right|.$$

More explicit formulations of UD for specific combinations of RVs are given in Section 6.A.1. For example, when X and Y remain discrete and take values in E_X and E_Y, respectively, equivalently it can be defined as:

$$\mathrm{UD}\left(X,Y\right) := 2 \sup_{A \subset E_X \times E_Y} \left\{ \sum_{(x,y)\in A} \left(p_{X,Y}(x,y) - p_X(x) \cdot p_Y(y) \right) \right\}.$$

Similarly, for continuous RVs, we obtain the following definition for UD.

Definition 6.4.2 (Continuous UD)**.** For any continuous RVs X and Y,

$$\mathrm{UD}\left(X,Y\right) := \int_{\mathbb{R}} \int_{\mathbb{R}} |f_{X,Y}(x,y) - f_X(x) f_Y(y)| dy dx.$$

Note that this is the same as Δ_{L_1} [35].

In the general case, UD is defined in the following manner.

Definition 6.4.3 (General UD)**.** For $X : (\Omega, \mathcal{F}, \mu) \to (E_X, \mathcal{E}(X))$ and $Y : (\Omega, \mathcal{F}, \mu) \to (E_Y, \mathcal{E}(Y))$, UD is defined as

$$\mathrm{UD}\left(X,Y\right) := 2 \sup_{A \in \mathcal{E}(X) \bigotimes \mathcal{E}(Y)} \left\{ \mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A) \right\},$$

where $\mathcal{E}(X) \bigotimes \mathcal{E}(Y)$ is the σ-algebra generated by the sets $C \times D$ with $C \in \mathcal{E}(X)$ and $D \in \mathcal{E}(Y)$. Furthermore, $\mu_{(X,Y)}$ denotes the joint probability measure on $\mathcal{E}(X) \bigotimes \mathcal{E}(Y)$ and $\mu_X \times \mu_Y$ is the product measure.

6.4.2 Maximum UD given Y

Next, we have to determine the maximum of UD for a fixed Y in order to scale the dependency function to $[0, 1]$. To this end, we prove that for a given Y:

$$X \text{ fully determines } Y \Rightarrow \mathrm{UD}\left(X,Y\right) \geq \mathrm{UD}\left(X',Y\right),$$

for any RV X'.

The full proof for the general case is given in Section 6.A.2, which uses the upper bound determined in Section 6.A.2. However, we will show the discrete case here to give some intuition about the proof. Let $C_y = \{x | p_{X,Y}(x,y) \geq p_X(x) \cdot p_Y(y)\}$, then

$$\begin{aligned}
\mathrm{UD}(X,Y) &= 2\sum_y \left(p_{X,Y}(C_y \times \{y\}) - p_X(C_y)\cdot p_Y(y)\right) \\
&\leq 2\sum_y \left(\min\left\{p_X(C_y), p_Y(y)\right\} - p_X(C_y)\cdot p_Y(y)\right) \\
&= 2\sum_y \left(\min\left\{p_X(C_y)(1-p_Y(y)), (1-p_X(C_y))p_Y(y)\right\}\right) \\
&\leq 2\sum_y \left(p_Y(y)(1-p_Y(y))\right) \\
&= 2\sum_y \left(p_Y(y) - p_Y(y)^2\right) \\
&= 2\left(1 - \sum_y p_Y(y)^2\right),
\end{aligned}$$

with equality iff both inequalities are equalities. This occurs iff $p_{X,Y}\left(C_y \times \{y\}\right) = p_X(C_y) = p_Y(y)$ for all y. So we have equality when for all y the set C_y has the property that $x \in C_y$ iff $Y = y$. Or equivalently $Y = f(X)$ for some function f. Thus,

$$UD(X,Y) \leq 2\left(1 - \sum_y p_Y(y)^2\right),$$

with equality iff $Y = f(X)$ for some function f.

Note that this holds for every X that fully determines Y. In particular, for $X := Y$ it now follows that

$$\mathrm{UD}(Y,Y) = 2 \cdot (1 - \sum_y p_Y(y)^2) \geq \mathrm{UD}(X',Y),$$

for any RV X'.

6.4.3 The definition of the Berkelmans-Pries dependency function

Finally, we can define the BP dependency function to measure how much Y is dependent on X. We call a random variable Y :

$(\Omega, \mathcal{F}, \mu) \to (E_Y, \mathcal{E}_Y)$ non-trivial if E_Y is not an atom. We call it trivial if E_Y is an atom (in most practical cases this is the same as a random variable being a.s. constant).

Definition 6.4.4 (BP dependency function)**.** For any RVs X and Y the *Berkelmans-Pries dependency function* is defined as

$$\mathrm{Dep}\,(Y|X) := \begin{cases} \frac{\mathrm{UD}(X,Y)}{\mathrm{UD}(Y,Y)} & \text{if } Y \text{ is non-trivial,} \\ \text{undefined} & \text{if } Y \text{ is trivial.} \end{cases}$$

This is the difference between the distribution of Y *with* and *without* conditioning on the value of X times the probability that X takes on this value divided by the largest possible difference for an arbitrary X'. Note that $\mathrm{UD}\,(Y,Y) = 0$ if and only if Y is trivial (see Section 6.A.2), which leads to division by zero. However, we previously argued in Section 6.2 that if Y is almost surely constant, it is completely independent as well as completely determined by X. It should therefore be undefined.

6.5 Properties of the Berkelmans-Pries dependency function

Next, we show that our new BP dependency function satisfies all requirements from Table 6.2. To this end, we use properties of UD (see Section 6.A.2) to derive properties II.1-8.

Property II.1 (Asymmetry): It holds for Example 6.2.1 that $\mathrm{UD}\,(X,Y) = 1$, $\mathrm{UD}\,(X,X) = 2$, and $\mathrm{UD}\,(Y,Y) = 1$. Thus,

$$\mathrm{Dep}\,(Y|X) = \frac{\mathrm{UD}\,(X,Y)}{\mathrm{UD}\,(Y,Y)} = 1,$$
$$\mathrm{Dep}\,(X|Y) = \frac{\mathrm{UD}\,(X,Y)}{\mathrm{UD}\,(X,X)} = \frac{1}{2}.$$

Therefore, we see that $\mathrm{Dep}\,(Y|X) \neq \mathrm{Dep}\,(X|Y)$ for this example, thus making the BP dependency function asymmetric.

Property II.2 (Range): In Section 6.A.2, we show that for every X, Y it holds that $\mathrm{UD}\,(X,Y) \geq 0$. Furthermore, in Section 6.A.2 we

prove that $\mathrm{UD}(X,Y) \leq 2\left(1-\sum_{y\in d_Y}\mu_Y(\{y\})^2\right)$ for all RVs X. In Section 6.A.2 we show for almost all cases that this bound is tight for $\mathrm{UD}(Y,Y)$. Thus, it must hold that $0 \leq \mathrm{UD}(X,Y) \leq \mathrm{UD}(Y,Y)$ and it then immediately follows that $0 \leq \mathrm{Dep}(Y|X) \leq 1$.

Property II.3 (Independence equals a dependency of 0): In Section 6.A.2, we prove that

$$\mathrm{UD}(X,Y) = 0 \Leftrightarrow X \text{ and } Y \text{ are independent.}$$

Furthermore, note that $\mathrm{Dep}(Y|X) = 0$ if and only if $\mathrm{UD}(X,Y) = 0$. Thus,

$$\mathrm{Dep}(Y|X) = 0 \Leftrightarrow X \text{ and } Y \text{ are independent.}$$

Property II.4 (Functional dependence equals a dependency of 1): In Section 6.A.2, we show that if X fully determines Y and X' is any RV we have that $\mathrm{UD}(X,Y) \geq \mathrm{UD}(X',Y)$. This holds in particular for $X := Y$. Thus, if X fully determines Y it follows that $\mathrm{UD}(X,Y) = \mathrm{UD}(Y,Y)$, so

$$\mathrm{Dep}(Y|X) = \frac{\mathrm{UD}(X,Y)}{\mathrm{UD}(Y,Y)} = 1.$$

In conclusion: if there exists a measurable function f such that $Y = f(X)$, then $\mathrm{Dep}(Y|X) = 1$

Property II.5 (Unambiguity): We show the result for discrete RVs below. For the proof of the general case see Section 6.A.2. Let E be the range of the independent $Y_1, Y_2, \ldots, Y_N$. By definition, it holds that $\mathbb{P}(X = x) = \sum_j \mathbb{P}(Y_j = x) \cdot \mathbb{P}(S = j)$, so for all

$i \in \{1, \dots N\}$

$$
\begin{aligned}
\mathrm{UD}\,(X, Y_i) &= 2 \sup_{A \subset E \times E} \sum_{(x,y) \in A} \left(p_{X,Y_i}(x,y) - p_X(x) p_{Y_i}(y)\right) \\
&= 2 \sup_{A \subset E \times E} \sum_{(x,y) \in A} \left(\sum_j p_{Y_j, Y_i, S}(x, y, j) - p_X(x) p_{Y_i}(y) \right) \\
&= 2 \sup_{A \subset E \times E} \left\{ \sum_{(x,y) \in A} \left(\sum_{j \neq i} \mathbb{P}(Y_j = x) \mathbb{P}(Y_i = y) \mathbb{P}(S = j) \right.\right. \\
&\qquad\qquad + \mathbb{P}(Y_i = x, Y_i = y) \mathbb{P}(S = i) \\
&\qquad\qquad \left.\left. - \sum_j \mathbb{P}(Y_j = x) \mathbb{P}(S = j) \mathbb{P}(Y_i = y) \right) \right\} \\
&= 2 \sup_{A \subset E \times E} \left\{ \sum_{(x,y) \in A} \left(p_i \mathbb{P}(Y_i = x, Y_i = y) \right.\right. \\
&\qquad\qquad \left.\left. - p_i \mathbb{P}(Y_i = x) \mathbb{P}(Y_i = y) \right) \right\} \\
&= p_i \cdot \mathrm{UD}\,(Y_i, Y_i)\,.
\end{aligned}
$$

This leads to

$$
\mathrm{Dep}\,(Y_i | X) = \frac{\mathrm{UD}\,(X, Y_i)}{\mathrm{UD}\,(Y_i, Y_i)} = \frac{p_i \cdot \mathrm{UD}\,(Y_i, Y_i)}{\mathrm{UD}\,(Y_i, Y_i)} = p_i.
$$

Therefore, we can conclude that property II.5 holds.

Property II.6 (Generally applicable): The BP dependency function can be applied for any combination of continuous, discrete and categorical variables. It can handle arbitrary many RVs as input by combining them. Thus, the BP dependency function is generally applicable.

Property II.7 (Invariance under isomorphisms): In Section 6.A.2, we prove that applying a measurable function to X or Y does not increase UD. Thus, it must hold for all isomorphisms f, g that

$$
\begin{aligned}
\mathrm{UD}\,(X, Y) &= \mathrm{UD}\left(f^{-1}(f(X)), g^{-1}(g(Y))\right) \\
&\leq \mathrm{UD}\,(f(X), g(Y)) \\
&\leq \mathrm{UD}\,(X, Y)\,.
\end{aligned}
$$

Therefore, all inequalities are actually equalities. In other words,

$$\mathrm{UD}\left(f(X), g(Y)\right) = \mathrm{UD}\left(X, Y\right).$$

It now immediately follows for the BP dependency function that

$$\begin{aligned}\mathrm{Dep}\left(g(Y)|f(X)\right) &= \frac{\mathrm{UD}\left(f(X), g(Y)\right)}{\mathrm{UD}\left(g(Y), g(Y)\right)} \\ &= \frac{\mathrm{UD}\left(X, Y\right)}{\mathrm{UD}\left(Y, Y\right)} \\ &= \mathrm{Dep}\left(Y|X\right),\end{aligned}$$

thus Property II.7 is satisfied.

Desired property II.8 (Non-increasing under functions of X): In Section 6.A.2, we prove that transforming X or Y using a measurable function does not increase UD. In other words, for any measurable function f, it holds that

$$\mathrm{UD}\left(f(X), Y\right) \leq \mathrm{UD}\left(X, Y\right).$$

Consequently, Property II.8 holds for the BP dependency function, as

$$\begin{aligned}\mathrm{Dep}\left(Y|f(X)\right) &= \frac{\mathrm{UD}\left(f(X), Y\right)}{\mathrm{UD}\left(Y, Y\right)} \\ &\leq \frac{\mathrm{UD}\left(X, Y\right)}{\mathrm{UD}\left(Y, Y\right)} \\ &= \mathrm{Dep}\left(Y|X\right).\end{aligned}$$

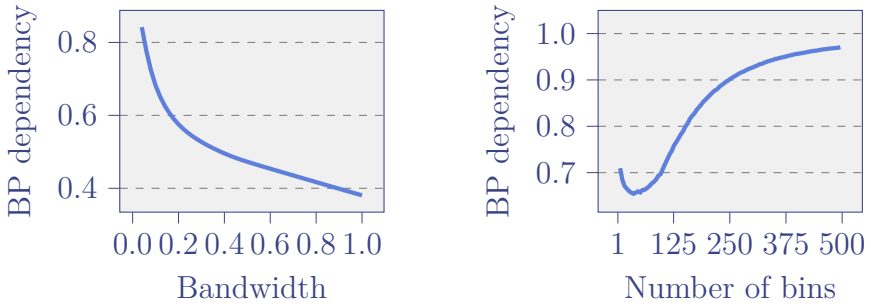
6.6 Discussion and further research

Motivated by the need to measure and quantify the level dependence between random variables, we have proposed a general-purpose dependency function. The function meets an extensive list of important and desired properties, and can be viewed as a powerful alternative to the classical Pearson correlation coefficient, which is often used by data analysts today.

Whilst it is recommended to use our new dependency function, it is important to understand the limitations and potential pitfalls of the new dependency function. Below we elaborate on these aspects.

The underlying probability density function of a RV is often unknown in practice; instead, a set of outcomes is observed. These samples can then be used (in a simple manner) to approximate any discrete distribution. However, this is generally not the case for continuous variables. There are mainly two categories for dealing with continuous variables: either (1) the observed samples are combined using kernel functions into a continuous function (*kernel density estimation* [68]), or (2) the continuous variable is reduced to a discrete variable using *data binning*. The new dependency measure can be applied thereafter.

A main issue is that the dependency measure is dependent of parameter choices of either *kernel density estimation* or *data binning*. To illustrate this, we conduct the following experiment: Let $X \sim \mathcal{U}(0, 1)$ and define $Y = X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.1)$. Next, we draw 5,000 samples of X and ϵ and determine each corresponding Y. For *kernel density estimation*, we use Gaussian kernels with constant bandwidth. The result of varying the bandwidth on the dependency score can be seen in Figure 6.1a. With *data binning*, both X and Y are binned using bins with fixed size. Increasing or decreasing the number of bins changes the size of the bins. The impact of changing the number of bins on the dependency score, can be seen in Figure 6.1b.



(a) Kernel density estimation **(b) Data binning**

Figure 6.1: Influence of chosen bandwidth (a) / number of bins (b) on the dependency score $\text{Dep}(Y|X)$ with 5,000 samples of $X \sim \mathcal{U}(0,1)$ and $Y = X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.1)$.

The main observation from Figures 6.1a and 6.1b is that the selection of the parameters is important. In the case of the *kernel density estimation*, we see the traditional *trade-off* between *over-fitting* when the bandwidth is too small and *under-fitting* when the bandwidth is too large. On the other hand, with *data binning*, we see different behaviour: Having too few bins seems to overestimate the dependency score and as bins increase the estimator of the dependency score decreases up to a certain point, where-after it starts increasing again. The bottom of the curve seems to be marginally higher than the true dependency score of 0.621.

This observation raises a range of interesting questions for future research. For example, are the dependency scores estimated by binning consistently higher than the true dependency? Is there a correction that can be applied to get an unbiased estimator? Is the minimum of this curve an asymptotically consistent estimator? Which binning algorithms give the closest approximation of the true dependency?

An interesting observation, with respect to kernel density estimation, is that it appears that at a bandwidth of 0.1 the estimator of the dependency score is close to the true dependency score of approximately 0.621. However, this parameter choice could only be made if the underlying probability process was known *a priori*.

Yet, there is another challenge with kernel density estimation, when X consists of many variables or feature values. Each time Y is conditioned on a different value of X, either the density needs to be estimated again or the estimation of the joint distribution needs to be integrated. Both can rapidly become very time-consuming. When using data binning, it suffices to bin the data once. Furthermore, no integration is required making it much faster. Therefore, our current recommendation would be to bin the data and not use kernel density estimation.

Another exciting research avenue would be to fundamentally explore the set of functions that satisfy all desired dependency properties. Is the BP dependency function the only measure that satisfies all conditions? If there exist two solutions, can we derive a new solution by smartly combining them? Without property II.5 any order-preserving bijection of $[0, 1]$ with itself would preserve all properties when applied to a solution. However, property II.5 does restrict the solution space. It remains an open problem if this is restrictive enough to result in a unique solution: the BP dependency function.

6.A Appendix

The following general notation is used throughout this appendix. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_X, \mathcal{E}_X)$ and $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_Y, \mathcal{E}_Y)$ be RVs. Secondly, let $\mu_X(A) = \mathbb{P}(X^{-1}(A))$, $\mu_Y(A) = \mathbb{P}(Y^{-1}(A))$ be measures induced by X and Y on $(E_X, \mathcal{E}_X)$ and $(E_Y, \mathcal{E}_Y)$, respectively. Furthermore, $\mu_{X,Y}(A) = \mathbb{P}(\{\omega \in \Omega | (X(\omega), Y(\omega)\}) \in A)$ is the joint measure and $\mu_X \times \mu_Y$ the product measure on $(E_X \times E_Y, \mathcal{E}_X \bigotimes \mathcal{E}_Y)$ generated by $(\mu_X \times \mu_Y)(A \times B) = \mu_X(A)\mu_Y(B)$.

6.A.1 Formulations of UD

In this appendix, we give multiple formulations of the *expected absolute change in distribution* (UD). Depending on the type of RVs, these formulations can be used.

General case:

For any X, Y the UD is defined as

$$\begin{aligned} \mathrm{UD}\,(X,Y) :&= \sup_{A\in\mathcal{E}(X)\bigotimes\mathcal{E}(Y)} \left\{\mu_{(X,Y)}(A) - (\mu_X\times\mu_Y)(A)\right\} \\ &+ \sup_{B\in\mathcal{E}(X)\bigotimes\mathcal{E}(Y)} \left\{(\mu_X\times\mu_Y)(B) - \mu_{(X,Y)}(B)\right\} \\ &= 2\sup_{A\in\mathcal{E}(X)\bigotimes\mathcal{E}(Y)} \left\{\mu_{(X,Y)}(A) - (\mu_X\times\mu_Y)(A)\right\}. \end{aligned} \tag{6.3}$$

Discrete RVs only:

When X, Y are discrete RVs, Equation (6.3) simplifies into

$$\mathrm{UD}\,(X,Y) := \sum_{x,y} \left|p_{X,Y}(x,y) - p_X(x)\cdot p_Y(y)\right|,$$

or equivalently

$$\mathrm{UD}\,(X,Y) := \sum_x p_X(x)\cdot\sum_y \left|p_{Y|X=x}(y) - p_Y(y)\right|.$$

Similarly, when X and Y take values in E_X and E_Y, respectively, Equation (6.3) becomes

$$\begin{aligned} \mathrm{UD}\,(X,Y) :&= \sup_{A\subset E_X\times E_Y} \left\{\sum_{(x,y)\in A} (p_{X,Y}(x,y) - p_X(x)p_Y(y))\right\} \\ &+ \sup_{A\subset E_X\times E_Y} \left\{\sum_{(x,y)\in A} (p_X(x)p_Y(y) - p_{X,Y}(x,y))\right\} \\ &= 2\sup_{A\subset E_X\times E_Y} \left\{\sum_{(x,y)\in A} (p_{X,Y}(x,y) - p_X(x)p_Y(y))\right\}. \end{aligned}$$

Continuous RVs only:

When X, Y are continuous RVs, Equation (6.3) becomes

$$\mathrm{UD}\,(X,Y) := \int_{\mathbb{R}}\int_{\mathbb{R}} |f_{X,Y}(x,y) - f_X(x)f_Y(y)|dydx,$$

when the joint distribution exists, or equivalently

$$\mathrm{UD}\,(X,Y) := \int_{\mathbb{R}} f_X(x) \int_{\mathbb{R}} |f_{Y|X=x}(y) - f_Y(y)| dy dx.$$

Another formulation (more measure theoretical) would be:

$$\mathrm{UD}\,(X,Y) := 2 \cdot \sup_{A \in \mathcal{B}(\mathbb{R}^2)} \left\{ \int_A (f_{X,Y}(x,y) - f_X(x) f_Y(y)) dy dx \right\}.$$

Mix of discrete and continuous:

When X is discrete and Y is continuous, Equation (6.3) reduces to

$$\mathrm{UD}\,(X,Y) := \sum_x p_X(x) \int_y |f_{Y|X=x}(y) - f_Y(y)| dy.$$

provided it is well-defined.

Vice versa, if X is continuous and Y is discrete, Equation (6.3) becomes

$$\mathrm{UD}\,(X,Y) := \int_x f_X(x) \sum_y |p_{Y|X=x}(y) - p_Y(y)| dx.$$

provided it is well-defined.

6.A.2 Properties of UD

In this appendix, we prove properties of UD that are used in Section 6.5 to show that the BP dependency function satisfies all properties in Table 6.2.

Symmetry of UD:

For the proofs below it is useful to show that $\mathrm{UD}\,(X,Y)$ is symmetric ()i.e., $\mathrm{UD}\,(X,Y) = \mathrm{UD}\,(Y,X)$ for every X, Y).

It directly follows from the definition as

$$\begin{aligned}
\mathrm{UD}\,(X,Y) &= 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \left\{ \mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A) \right\} \\
&= 2 \sup_{A \in \mathcal{E}_Y \bigotimes \mathcal{E}_X} \left\{ \mu_{(Y,X)}(A) - (\mu_Y \times \mu_X)(A) \right\} \\
&= \mathrm{UD}\,(Y,X)\,.
\end{aligned}$$

Independence and UD $= 0$:

Since we are considering a measure of dependence it is useful to know what the conditions for independence are. Below we will show that we have independence of X and Y if and only if $\mathrm{UD}\,(X, Y) = 0$.

Note that

$$\begin{aligned}\mathrm{UD}\,(X,Y) = &\sup_{A\in\mathcal{E}_X\bigotimes\mathcal{E}_Y}\Big\{\mu_{(X,Y)}(A) - (\mu_X\times\mu_Y)(A)\Big\}\\ &+\sup_{B\in\mathcal{E}_X\bigotimes\mathcal{E}_Y}\Big\{(\mu_X\times\mu_Y)(B) - \mu_{(X,Y)}(B)\Big\}\\ &\geq \Big(\mu_{(X,Y)}(E_X\times E_Y) - (\mu_X\times\mu_Y)(E_X\times E_Y)\Big)\\ &+\Big((\mu_X\times\mu_Y)(E_X\times E_Y) - \mu_{(X,Y)}(E_X\times E_Y)\Big)\\ &= 0,\end{aligned}$$

with equality if and only if $\mu_{(X,Y)} = \mu_X \times \mu_Y$ on $\mathcal{E}_X \bigotimes \mathcal{E}_Y$, so if and only if X and Y are independent. So in conclusion properties i) and ii) below are equivalent:

(i) X and Y are independent random variables,

(ii) $\mathrm{UD}\,(X, Y) = 0$.

Upper bound for a given Y:

To scale the dependency function it is useful to know what the range of $\mathrm{UD}\,(X, Y)$ is for a given random variable Y. We already know it is lower bounded by 0 (see Section 6.A.2). However, we have not yet established an upper bound. What follows down below is a derivation of the upper bound.

A μ_Y-atom A is a set such that $\mu_Y(A) > 0$ and for any measurable $B \subset A$ we have $\mu_Y(B) \in \{0, \mu_Y(A)\}$. Consider the following equivalence relation $\sim$ on the μ_Y-atoms characterized by $S \sim T$ if and only if $\mu_Y(S \triangle T) = 0$. Then let I be a set containing exactly one representative from each equivalence class. Note that I is countable, so we can enumerate the elements $A_1, A_2, A_3 \ldots$. Additionally, for any $A, B \in I$ we have that $\mu_Y(A \cap B) = 0$.

Next, we define $B_i := A_i \setminus \bigcup_{j=1}^{i-1} A_j$ to obtain a set of disjoint μ_Y-atoms. In what follows we assume I to be infinite (but remember that I is countable), but the proof works exactly the same for finite I when you replace ∞ with $|I|$.

Let $E_Y^* := E_Y \setminus \bigcup_{j=1}^{\infty} B_j$, so that the B_j's and the E_Y^* form a partition of E_Y. Furthermore, let $b_j := \mu_Y(B_j)$ be the probabilities of being in the individual atoms in I (and therefore the sizes corresponding to the equivalence classes of atoms). It now holds for any RV X that:

$$\begin{aligned}
\mathrm{UD}\,(X,Y) &= 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \{\mu_{X,Y}(A) - (\mu_X \times \mu_Y)(A)\} \\
&\leq 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \{\mu_{X,Y}\left(A \cap (E_X \times E_Y^*)\right) \\
&\quad -(\mu_X \times \mu_Y)\left(A \cap (E_X \times E_Y^*)\right)\} \\
&\quad + 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \left\{ \sum_{j=1}^{\infty} \left(\mu_{X,Y}\left(A \cap (E_X \times B_j)\right)\right.\right. \\
&\quad \left.\left. -(\mu_X \times \mu_Y)\left(A \cap (E_X \times B_j)\right)\right)\right\}.
\end{aligned} \tag{6.4}$$

Now note that the first term is at most $\mu_Y(E_Y^*) = 1 - \sum_{i=1}^{\infty} b_i$. To bound the second term, we examine each individual term of the summation. First we note that the set of finite unions of 'rectangles' (Cartesian products of elements in $\mathcal{E}_X$ and $\mathcal{E}_Y$)

$$R := \left\{ C \in \mathcal{E}_X \bigotimes \mathcal{E}_Y \mid \exists k \in \mathbb{N} \text{ s.t. } C = \bigcup_{i=1}^{k} (A_i \times B_i), \text{ with } \forall i : A_i \in \mathcal{E}_X \wedge B_i \in \mathcal{E}_Y \right\}$$

is an algebra. Therefore, for any $D \in \mathcal{E}_X \bigotimes \mathcal{E}_Y$ and $\epsilon > 0$, there exists a $D_\epsilon \in R$ such that $\nu(D_\epsilon \triangle D) < \epsilon$, where $\nu := \mu_{X,Y} + (\mu_X \times \mu_Y)$. Specifically for $A \cap (E_X \times B_j)$ and $\epsilon > 0$, there exists a $B_{j,\epsilon} \in R$ such that $\nu(B_{j,\epsilon} \triangle A \cap (E_X \times B_j)) < \epsilon$ and $B_{j,\epsilon} \subset E_X \times B_j$ holds, since intersecting with this set only decreases the expression whilst remaining in R.

Thus, we have that

$$\begin{aligned}
&|\mu_{X,Y}(A \cap (E_X \times B_j)) - \mu_{X,Y}(B_{j,\epsilon})| \\
&+|(\mu_X \times \mu_Y)(A \cap (E_X \times B_j)) - (\mu_X \times \mu_Y)(B_{j,\epsilon})| < \epsilon.
\end{aligned}$$

Therefore, it must hold that

$$\begin{aligned}&\mu_{X,Y}(A \cap (E_X \times B_j)) - (\mu_X \times \mu_Y)(A \cap (E_X \times B_j)) \\ \leq &\mu_{X,Y}(B_{j,\epsilon}) - (\mu_X \times \mu_Y)(B_{j,\epsilon}) + \epsilon.\end{aligned}$$

Since $B_{j,\epsilon}$ is a finite union of 'rectangles', we can also write it as a finite union of k disjoint 'rectangles such that $B_{j,\epsilon} = \bigcup_{i=1}^{k} S_i \times T_i$ with $S_i \in \mathcal{E}_X$ and $T_i \in \mathcal{E}_Y$ for all i. It now follows that

$$\begin{aligned}&\mu_{X,Y}(B_{j,\epsilon}) - (\mu_X \times \mu_Y)(B_{j,\epsilon}) + \epsilon \\ &= \epsilon + \sum_{i=1}^{k} \mu_{X,Y}(S_i \times T_i) - (\mu_X \times \mu_Y)(S_i \times T_i).\end{aligned}$$

For all i it holds that $T_i \subset B_j$ which means that either $\mu_Y(T_i) = 0$ or $\mu_Y(T_i) = b_j$, since B_j is an atom of size b_j. This allows us to separate the sum

$$\begin{aligned}&\epsilon + \sum_{i=1}^{k} \mu_{X,Y}(S_i \times T_i) - (\mu_X \times \mu_Y)(S_i \times T_i) \\ &= \epsilon + \sum_{i:\mu_Y(T_i)=0} (\mu_{X,Y}(S_i \times T_i) - (\mu_X(S_i) \times \mu_Y(T_i))) \\ &+ \sum_{i:\mu_Y(T_i)=b_j} (\mu_{X,Y}(S_i \times T_i) - (\mu_X(S_i) \times \mu_Y(T_i))) \\ &= \star.\end{aligned}$$

The first sum is equal to zero, since $\mu_{X,Y}(S_i \times T_i) \leq \mu_Y(T_i) = 0$. The second sum is upper bounded by $\mu_{X,Y}(S_i \times T_i) \leq \mu_{X,Y}(S_i \times B_j)$. By defining $S' = \bigcup_{i:\mu_Y(T_i)=b_j} S_i$, we obtain

$$\begin{aligned}\star &\leq \epsilon + 0 + \sum_{i:\mu_Y(T_i)=b_j} (\mu_{X,Y}(S_i \times B_j) - b_j \cdot \mu_X(S_i)) \\ &= \epsilon + \mu_{X,Y}(S' \times B_j) - b_j \cdot \mu_x(S') \\ &\leq \epsilon + \min\{(1 - b_j) \cdot \mu_X(S'), b_j \cdot (1 - \mu_X(S'))\} \\ &\leq \epsilon + b_j - b_j^2.\end{aligned}$$

But, since this is true for any $\epsilon > 0$, it holds that

$$\mu_{X,Y}(A \cap (E_X \times B_j)) - (\mu_X \times \mu_Y)(A \cap (E_X \times B_j)) \leq b_j - b_j^2.$$

Plugging this back into Equation (6.4) gives

$$\begin{aligned}
\mathrm{UD}\,(X,Y) \leq 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} & \{\mu_{X,Y}(A \cap (E_X \times E_Y^*)) \\
& -(\mu_X \times \mu_Y)(A \cap (E_X \times E_Y^*))\} \\
& + 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \left\{ \sum_{j=1}^{\infty} (\mu_{X,Y}(A \cap (E_X \times B_j)) \right. \\
& \left. -(\mu_X \times \mu_Y)(A \cap (E_X \times B_j))) \right\} \\
& \leq 2 \left(1 - \sum_{i=1}^{\infty} b_i\right) + 2 \cdot \sum_{j=1}^{\infty} (b_j - b_j^2) \\
& = 2 \left(1 - \sum_{i=1}^{\infty} b_i^2\right).
\end{aligned}$$

Note that in the continuous case the summation is equal to 0, so the upper bound simply becomes 2. In the discrete case, where E_Y is the set in which Y takes its values, the expression becomes

$$\mathrm{UD}\,(X,Y) \leq 2 \left(1 - \sum_{i \in E_Y} \mathbb{P}(Y = i)^2\right).$$

Functional dependence attains maximum UD:

Since we established an upper bound in Section 6.A.2, the next step is to check whether this bound is actually attainable. What follows is a proof that this bound is achieved for any random variable X for which it holds that $Y = f(X)$ for some measurable function f.

Let $Y = f(X)$ for some measurable function f, then $\mu_X(f^{-1}(C)) = \mu_Y(C)$ for all $C \in \mathcal{E}_Y$. Let the μ_Y-atoms B_j and E_Y^* be the same as in Section 6.A.2. Since E_Y^* contains no atoms, for every $\epsilon > 0$ there exists a partition $T_1, \ldots, T_k$ for some $k \in \mathbb{N}$ such that $\mu_Y(T_i) < \epsilon$ for all i. Then, consider the set $K = (\bigcup_i (f^{-1}(T_i) \times T_i)) \cup \bigcup_j (f^{-1}(B_j) \times$

B_j). It now follows that

$$\begin{aligned}
\mathrm{UD}\,(X,Y) &= 2 \sup_{A\in\mathcal{E}_Y \bigotimes \mathcal{E}_Y} \mu_{X,Y}(A) - (\mu_X \times \mu_Y)(A) \\
&\geq 2\mu_{X,Y}(K) - (\mu_X \times \mu_Y)(K) \\
&= 2\left(\sum_i (\mu_{X,Y}(f^{-1}(T_i) \times T_i) - \mu_X(f^{-1}(T_i))\mu_Y(T_i)) \right. \\
&\left. + \sum_j (\mu_{X,Y}(f^{-1}(B_j) \times B_j) - \mu_X(f^{-1}(B_j))\mu_Y(B_j))\right) \\
&\geq 2\left(\sum_i (\mu_Y(T_i) - \epsilon * \mu_Y(T_i)) + \sum_j (b_j - b_j^2)\right) \\
&= 2\left((1 - \sum_j b_j) - \epsilon(1 - \sum_j b_j) + \sum_j (b_j - b_j^2)\right).
\end{aligned}$$

But, since this holds for any $\epsilon > 0$ we have

$$\mathrm{UD}\,(X,Y) \geq 2\left(1 - \sum_j b_j^2\right).$$

As this is also the upper bound from Section 6.A.2, equality must hold. Thus, we can conclude that $\mathrm{UD}\,(X,Y)$ is maximal for Y if $Y = f(X)$ (so in particular if $X = Y$). As a result, for any RVs X_1, X_2, Y with $Y = f(X_1)$ for some measurable function f, we have $\mathrm{UD}\,(X_1,Y) \geq \mathrm{UD}\,(X_2,Y)$. Note that a corollary of this proof is that $\mathrm{UD}\,(Y,Y) = 0$ if and only if there exists a μ_Y-atom B_i with $\mu_Y(B_i) = 1$, or in other words there are no events that occur with a probability strictly between 0 and 1. So if and only if Y is trivial.

Unambiguity:

In Section 6.5, we show for discrete RVs that property II.5 holds. In this section, we prove the general case. Let $Y_1, \ldots, Y_N$ and S be independent RVs where S takes values in $1, \ldots, N$ with $\mathbb{P}(S = i) = p_i$. Finally, define $X := Y_S$. Then we will show that $\mathrm{Dep}\,(Y_i|X) = p_i$.

Let $\mathcal{E}$ be the σ-algebra on which the independent Y_i are defined. Then we have $\mu_{X,Y_i,S}(A \times \{j\}) = \mu_{Y_j,Y_i}(A)\mu_S(\{j\}) = p_j \mu_{Y_j,Y_i}(A)$ for all j. Additionally, we have $\mu_X(A) = \sum_j p_j \mu_{Y_j}(A)$. Lastly, due to independence for $i \neq j$ we have $\mu_{Y_j,Y_i} = \mu_{Y_j} \times \mu_{Y_i}$. Combining this, gives

$$\begin{aligned}
\mathrm{UD}\,(X, Y_i) &= 2 \sup_{A \in \mathcal{E} \times \mathcal{E}} \{\mu_{X,Y_i}(A) - (\mu_X \times \mu_{Y_i})(A)\} \\
&= 2 \sup_{A \in \mathcal{E} \times \mathcal{E}} \left\{ \sum_j \mu_{X,Y_i,S}(A \times \{j\}) - \sum_j p_j (\mu_{Y_j} \times \mu_{Y_i})(A) \right\} \\
&= 2 \sup_{A \in \mathcal{E} \times \mathcal{E}} \left\{ \sum_j p_j \left(\mu_{Y_j,Y_i}(A) - (\mu_{Y_j} \times \mu_{Y_i})(A) \right) \right\} \\
&= 2 \sup_{A \in \mathcal{E} \times \mathcal{E}} \{p_i \left(\mu_{Y_i,Y_i}(A) - (\mu_{Y_i} \times \mu_{Y_i})(A) \right)\} \\
&= p_i \cdot \mathrm{UD}\,(Y_i, Y_i)\,.
\end{aligned}$$

Measurable functions never increase UD:

Next, we prove another useful property of UD: applying a measurable function to one of the variables does not increase the UD. Let $f : (E_X, \mathcal{E}_X) \to (E_{X'}, \mathcal{E}_{X'})$ be a measurable function. Then $h : E_X \times E_Y \to E_{X'} \times E_Y$ with $h(x, y) = (f(x), y)$ is measurable. Now it follows that

$$\begin{aligned}
\mathrm{UD}\,(f(X), Y) &= 2 \sup_{A \in \mathcal{E}_{X'} \bigotimes \mathcal{E}_Y} \left\{ \mu_{(f(X),Y)}(A) - (\mu_{f(X)} \times \mu_Y)(A) \right\} \\
&= 2 \sup_{A \in \mathcal{E}_{X'} \bigotimes \mathcal{E}_Y} \left\{ \mu_{(X,Y)}(h^{-1}(A)) - (\mu_X \times \mu_Y)(h^{-1}(A)) \right\},
\end{aligned}$$

with $h^{-1}(A) \in \mathcal{E}_X \bigotimes \mathcal{E}_Y$. Thus,

$$\begin{aligned}
\mathrm{UD}\,(f(X), Y) &\leq 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} (\mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A)) \\
&= \mathrm{UD}\,(X, Y)\,.
\end{aligned}$$

In Section 6.A.2, it is proven that UD is symmetric. Therefore, it also holds for $g : E_Y \to E_{Y'}$, that

$$\mathrm{UD}\,(X, g(Y)) \leq \mathrm{UD}\,(X, Y)\,.$$

6.A.3 Definitions previous methods and (references to) proofs/counter-examples properties

Pearson correlation coefficient

For X, Y random variables taking values in $\mathbb{R}$ and finite first and second moment the Pearson's correlation coefficient is defined as

$$\rho(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

Property II.1 fails by the symmetricity of the definition.

For $X = -Y$, it becomes equal to -1 which is not in the range $[0,1]$ so Property II.2 fails.

If $X \sim U(-1,1)$ and $Y = X^2$ then $\rho(X,Y) = 0$ but there is no independence, there is even full dependence of Y on X so we have that both Property II.3 and II.4 fail.

For $Y_1, Y_2, \ldots, Y_n, S$ independent with $\mathbb{P}(S = i) = p_i$ and $X = Y_S$ have that $\mathbb{E}(XY_i) - \mathbb{E}(X)\mathbb{E}(Y_i) = p_i(\mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2) = p_i\text{Var}(Y_i)$ so $\rho(X,Y_i) = p_i$ if and only if $\text{Var}(X) = \text{Var}(Y_i)$ which is not necessarily the case so Property II.5 fails.

Since it is only defined for random variables taking values in the reals (and not for example for random vectors) we have that Property II.6 fails.

Finally for $Y = X^2$ and $X \sim U(0,1)$ we have that $\mathbb{E}(X) = \frac{1}{2}, \mathbb{E}(X^2) = \mathbb{E}(Y) = \frac{1}{3}, \mathbb{E}(XY) = \frac{1}{4}, \mathbb{E}(Y^2) = \frac{1}{5}$ so we have $\rho(X,Y) = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}\frac{4}{45}}} = \frac{\sqrt{15}}{4} < 1$. However applying the $\sqrt{}$-function to Y, we have $\rho(X,\sqrt{Y}) = \rho(X,X) = 1$ so Property II.7 and Property II.8 fail. So none of the properties are satisfied by the Pearson Correlation Coefficient.

Spearman rank correlation coefficient

For a sample of size n with values X_i, Y_i the Spearman's rank correlation coefficient is defined as

$$r_s(X_i, Y_i) = \rho(R(X_i), R(Y_i)),$$

where $R(X)$ and $R(Y)$ are the rank variables (note that for these to exist both the spaces where X and Y take values must be equipped with an ordering). For X, Y random variables one can define

$$r_s(X, Y) = \lim_{n\to\infty} r_s(X_n, Y_n).$$

Property II.1 fails by the symmetricity of the definition.

For $X = -Y$, it becomes equal to -1 which is not in the range $[0, 1]$ so Property II.2 fails.

If $X \sim U(-1, 1)$ and $Y = X^2$ then $r_s(X, Y) = 0$ but there is no independence, there is even full dependence of Y on X so we have that both Properties II.3 and II.4 fail.

Take Y_1, Y_2, S independent where Y_1 takes values 1, 3 with probability $\frac{1}{2}$ each, Y_2 takes values 2, 4 with probability $\frac{1}{2}$ each, and S takes values 1, 2 with probability $\frac{1}{2}$ each. Then let $X = Y_S$, then we have that (by conditioning on X and scaling the rank variables)

$$\begin{aligned} r_s(X, Y) &= \frac{\frac{1}{4}(\frac{1}{8}\cdot\frac{1}{4}) + \frac{1}{4}(\frac{5}{8}\cdot\frac{3}{4}) + \frac{1}{4}(\frac{3}{8}\cdot\frac{1}{2}) + \frac{1}{4}(\frac{7}{8}\cdot\frac{1}{2}) - \frac{1}{2}^2}{\sqrt{\frac{1}{2}\cdot(\frac{1}{8})^2 + \frac{1}{2}\cdot(\frac{3}{8})^2}\sqrt{(\frac{1}{4})^2}} \\ &= \frac{\frac{1}{32}}{\frac{\sqrt{5}}{32}} \\ &= \frac{1}{\sqrt{5}} \\ &\neq \frac{1}{2}, \end{aligned}$$

so Property II.5 fails.

Since it is only defined when rank variables are defined (or in other words if there is an ordering) we have that Property II.6 fails.

Suppose $X = -Y$ (not a.s. constant) then $r_s(X, Y) = -1$. But $r_s(X, -Y) = r_s(-X, Y) = 1$ so both Properties II.7 and II.8 fail.

Kendall rank correlation coefficient

For a sample of size n with values X_i, Y_i the Spearman's rank correlation coefficient is defined as

$$\tau(X_i, Y_i) = \frac{C - D}{C + D} = 1 - 2 \cdot \frac{D}{C + D},$$

where C is the number of concordant pairs i, j where X and Y agree on the ordering, so ones where both $X_i > X_j$ and $Y_i > Y_j$ or both $X_i < X_j$ and $Y_i < Y_j$. And D is the number of discordant pairs where X and Y disagree on the ordering, so both $X_i > X_j$ and $Y_i < Y_j$ or both $X_i < X_j$ and $Y_i > Y_j$. For X, Y random variables one can define

$$\begin{aligned}\tau(X, Y) &= \lim_{n \to \infty} \tau(X_i, Y_i) \\ &= 1 - 2\mathbb{P}(\text{discordant} \mid \text{either discordant or concordant}).\end{aligned}$$

Property II.1 fails by the symmetricity of the definition.

For $X = -Y$, it becomes equal to -1 which is not in the range $[0, 1]$ so Property II.2 fails.

For $X \sim U(\{-1, 0, 1\})$ and $Y = |X|$ we have that $\mathbb{P}(\text{discordant}|\text{concordant or discordant}) = \frac{1}{2}$ so $\tau(X, Y) = 0$ but do not have independence but full dependence. So both Properties II.3 and II.4 fail.

Take Y_1, Y_2, S independent where Y_1 takes values 1, 3 with probability $\frac{1}{2}$ each, Y_2 takes values 2, 4 with probability $\frac{1}{2}$ each, and S takes values 1, 2 with probability $\frac{1}{2}$ each. Then let $X = Y_S$, then we have that (by simply counting among all combinations of the 8 cases)

$$\tau(X, Y_1) = \frac{4}{7} \neq \frac{1}{2},$$

so Property II.5 fails.

Since $\tau(\cdot,\cdot)$ depends on an ordering existing we have that Property II.6 fails.

For $X = -Y$ (and Y not a.s. constant) and for $f(x) = -x$ have

$$\tau(f(X), Y) = 1 > -1 = \tau(X, Y),$$

so both Properties II.7 and II.8 fail.

Mutual information

In the main text we described the discrete case for mutual information. In general it is defined as

$$I(X;Y) = D_{KL}\left(\mathbb{P}_{X,Y} || \mathbb{P}_X \times \mathbb{P}_Y\right),$$

where for two probability measures P, Q on a measurable space $\mathcal{X}$ the Kullbach-Leibler divergence is defined as

$$D_{KL}(P||Q) = \int_{\mathcal{X}} \log(\frac{P(dx)}{Q(dx)}) P(dx).$$

Property II.1 fails by the symmetricity of the definition.

Since for X, Y multivariate Gaussian random variables with correlation ρ it can be shown that

$$I(X;Y) = -\frac{1}{2}\log(1-\rho^2),$$

we have that for $\rho = \sqrt{\frac{1}{e^4}}$ we get $I(X;Y) = 2 \notin [0,1]$ so Property II.2 fails, additionally Property II.4 fails when $\rho = 1$ where it is either ∞ or undefined (see also the upcoming paragraph concerning Property II.6).

Alternatively if we consider X a Bernoulli(p) random variable and $Y = X$ we have for almost all p that

$$I(X;Y) = p\log\left(\frac{p}{p^2}\right) + (1-p)\log\left(\frac{1-p}{(1-p)^2}\right) \neq 1,$$

so Property II.4 still fails even though $I(X;Y)$ is well-defined.

Since the Kullbach-Leibler divergence of P and Q is equal to 0 iff $P = Q$, we have that the mutual information is 0 iff $\mathbb{P}_{X,Y} = \mathbb{P}_X \times \mathbb{P}_Y$ so Property II.3 holds.

Let Y_1, Y_2, S be independent with $\mathbb{P}(Y_i = j) = \frac{1}{2} = \mathbb{P}(S = i)$ for $i, j \in \{1, 2\}$. Then

$$I(X; Y_1) = \frac{3}{8}\log\frac{3}{2} + \frac{1}{8}\log\frac{1}{2} + \frac{3}{8}\log\frac{3}{2} + \frac{1}{8}\log\frac{1}{2} = \frac{1}{4}\log\frac{27}{16} \neq \frac{1}{2},$$

so Property II.5 fails.

For $X = Y$ with $X \sim U(0, 1)$ we have that the joint distribution becomes singular whereas the product distribution is not. This results in the Kullbach-Leibler divergence not being defined and thus the mutual information not being defined. One could allow for the value of ∞ which is why we say in this case that whether or not Property II.6 is satisfied depends on ones perspective.

Since the KL divergence is invariant under bijections, so is mutual information. So therefore Property II.7 holds.

Property II.8 is a corollary of the Data-processing inequality.

Uncertainty coefficient

The uncertainty coefficient is a normalised version of mutual information, defined as follows:

$$U(Y|X) = \frac{I(X; Y)}{H(Y)},$$

where $H(Y)$ is the entropy of Y.

Since there exist X, Y with $H(X) \neq H(Y)$ and $I(X; Y) \neq 0$, these have $U(Y|X) \neq U(X|Y)$. Therefore Property II.1 holds.

Since $0 \leq I(X; Y) \leq H(Y)$ we have $U(Y|X) \in [0, 1]$ so Property II.2 holds.

Since $I(X; Y) = 0$ iff X, Y independent then the same holds for $U(Y|X)$. So Property II.3 holds.

If Y is completely determined by X and $U(Y|X)$ is well-defined, then this means that $P_{X,Y}$ is not singular so we have that Y must be discrete, but in that case $H(Y|X) = 0$ so $H(Y) = H(Y) - H(Y|X) = I(X;Y)$ and thus $U(Y|X) = 1$. So Property II.4 holds.

Using the same example as we used for mutual information we find $I(X;Y_1) = \frac{1}{4}\log(\frac{27}{16})$ and $H(Y_1) = -\log(\frac{1}{2}) \neq 2I(X;Y_1)$ so $U(Y_1|X) \neq \frac{1}{2}$ so Property II.5 fails.

Though $H(Y)$ is not always defined, one could generously extend the definition to include ∞ as we did for mutual information, however in that case, the example we had for mutual information would become

$$U(Y|X) = \frac{\infty}{\infty},$$

which is most definitely undefined. Therefore Property II.6 fails.

Since mutual information is invariant under bijections, and so is entropy we find that Property II.7 holds.

Finally since $H(Y)$ is invariant under changes to X and Property II.8 holds for mutual information we have that it holds for the uncertainty coefficient as well.

Total correlation

Total correlation can be seen as the multi-dimensional extension of mutual information. For $X_1, \dots X_n$ the total correlation is given as

$$C(X_1, \dots, X_n) = D_{KL}(\mathbb{P}_{X_1,\dots,X_n} || \mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}).$$

Property II.1 fails by symmetricity of the definition.

By considering $n = 2$ we see that we are once more at mutual information and therefore Properties II.2, II.4, and II.5 fail.

Since the Kullbach-Leibler divergence of P and Q is equal to 0 iff $P = Q$, we have that the total correlation is 0 iff $\mathbb{P}_{X_1,\dots,X_n} = \mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}$ so Property II.3 holds.

Since the KL-divergence is only defined when $\mathbb{P}_{X_1,\ldots,X_n} << \mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}$ we have the same situation as with mutual information, where we could allow for the value of ∞ in the case that there exists a set A with $\mathbb{P}_{X_1,\ldots,X_n}(A) > 0$ but $\mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}(A) = 0$.

Since the KL-divergence is invariant under bijections, so is total correlation. So therefore Property II.7 holds.

An earlier result [47] has shown that

$$C(X_1, X_2, \ldots, X_n) = C(X_1, \ldots X_{n-1}) + I((X_1, X_2, \ldots, X_{n-1}); X_n),$$

so since the first term is invariant under applying functions to X_n and the second term is non-increasing, we have that total correlation is non-increasing (and similarly for the other X_i. So Property II.8 holds.

Mutual dependence

Mutual dependence is defined as the Hellinger distance d_h between the joint and product distributions:

$$d(X, Y) = d_h(\mathbb{P}_{X,Y}, \mathbb{P}_X \cdot \mathbb{P}_Y),$$

where

$$d_h^2(P, Q) = \frac{1}{2} \int_x \left(\sqrt{\frac{P(dx)}{\lambda(dx)}} - \sqrt{\frac{Q(dx)}{\lambda(dx)}} \right)^2 \lambda(dx),$$

for some measure λ (independent of choice (as long as absolute continuity of P and Q is guaranteed) so can pick $\lambda = P + Q$ for example)

By symmetricity of the definition we have that Property II.1 fails.

Agarwal et al. [4] showed that Properties II.2, II.3, and II.4 hold. However they only showed this for the case that both X, Y are continuous. Consider the case that X is 1 with probability $\frac{1}{2}$ and 0

otherwise, and let $Y = X$. Then

$$\begin{aligned} d(X,Y) &= \frac{1}{2}\sum_{i=0}^{1}\sum_{j=0}^{1}\left(\sqrt{\mathbb{P}(X=i,Y=j)} - \sqrt{\mathbb{P}(X=i)\mathbb{P}(Y=j)}\right)^2 \\ &= \frac{1}{2}(2\cdot(\sqrt{\frac{1}{2}} - \sqrt{\frac{1}{4}})^2 + 2\cdot(0-\sqrt{\frac{1}{4}})^2) \\ &= \frac{3}{4} - \frac{1}{2}\sqrt{2} + \frac{1}{4} < 1, \end{aligned}$$

so Property II.4 fails in general.

The case above violates the special case of Property II.5 where $N = 1$ and $p_1 = 1$.

Since the Hellinger distance is always defined so is mutual dependence. So Property II.6 is satisfied.

Tjøstheim et al showed that Property II.7 is satisfied [153].

Clearly $d(X,Y)$ and $d^2(X,Y)$ have the same ordering. We can rewrite $d^2(X,Y)$ as

$$\begin{aligned} & 1 - \int_{E_X\times E_Y}\sqrt{\frac{d\mathbb{P}_{X,Y}}{d\lambda}}\sqrt{\frac{d(\mathbb{P}_X\mathbb{P}_Y)}{d\lambda}}d\lambda \\ &= 1 - \mathbb{E}_\lambda\left(\sqrt{\frac{d\mathbb{P}_{X,Y}}{d\lambda}}\sqrt{\frac{d(\mathbb{P}_X\mathbb{P}_Y)}{d\lambda}}\right) \\ &= 1 - \mathbb{E}_{\lambda|_{\sigma(h)}}\left(\mathbb{E}_\lambda\left(\sqrt{\frac{d\mathbb{P}_{X,Y}}{d\lambda}}\sqrt{\frac{d(\mathbb{P}_X\mathbb{P}_Y)}{d\lambda}}|\sigma(h)\right)\right), \end{aligned}$$

where $\sigma(h)$ is the smallest σ-algebra such that $h(X,Y) = (f(X),Y)$ is measurable. Now by Hölder we have

$$\begin{aligned} & \mathbb{E}_\lambda\left(\sqrt{\frac{d\mathbb{P}_{X,Y}}{d\lambda}}\sqrt{\frac{d(\mathbb{P}_X\mathbb{P}_Y)}{d\lambda}}|\sigma(h)\right) \\ &\leq \sqrt{\mathbb{E}_\lambda\left(\frac{d\mathbb{P}_{X,Y}}{d\lambda}|\sigma(h)\right)}\sqrt{\mathbb{E}_\lambda\left(\frac{d(\mathbb{P}_X\mathbb{P}_Y)}{d\lambda}|\sigma(h)\right)} \\ &= \sqrt{\frac{d(\mathbb{P}_{X,Y}|_{\sigma(h)})}{d(\lambda|_{\sigma(h)})}}\sqrt{\frac{d(\mathbb{P}_X\mathbb{P}_Y|_{\sigma(h)})}{d(\lambda|_{\sigma(h)})}}, \end{aligned}$$

so

$$\begin{aligned} d^2(X,Y) &= 1 - \int_{E_X \times E_Y} \sqrt{\frac{d\mathbb{P}_{X,Y}}{d\lambda}} \sqrt{\frac{d(\mathbb{P}_X \mathbb{P}_Y)}{d\lambda}} d\lambda \\ &\geq 1 - \int_{E_X \times E_Y} \sqrt{\frac{d\mathbb{P}_{X,Y}|_{\sigma(h)}}{d\lambda|_{\sigma(h)}}} \sqrt{\frac{d(\mathbb{P}_X \mathbb{P}_Y)|_{\sigma(h)}}{d\lambda|_{\sigma(h)}}} d\lambda|_{\sigma(h)} \\ &= 1 - \int_{E_{f(X)} \times E_Y} \sqrt{\frac{d\mathbb{P}_{f(X),Y}}{d\lambda'}} \sqrt{\frac{d(\mathbb{P}_{f(X)} \mathbb{P}_Y)}{d\lambda'}} d\lambda' \\ &= d^2(f(X),Y), \end{aligned}$$

6

so Property II.8 is satisfied

The Δ_{L_1} dependency function

For X, Y equipped with density functions f, g and joint density function h we have

$$\Delta_{L_1}(X,Y) = \int_{\mathcal{E}_X \bigotimes \mathcal{E}_Y} |h(x,y) - f(x)g(y)| dxdy.$$

It is symmetric by definition so Property II.1 fails.

For $X = Y \sim U(0,1)$ we have that $\Delta_{L_1} = 2 \notin [0,1]$. So Property II.2 fails and so do Properties II.4, and II.5.

We have that $\Delta_{L_1}(X,Y) = 0$ if and only if $h(x,y) - f(x)g(y) = 0$ a.e. so if and only if X and Y are independent. So Property II.3 holds.

Since Δ_{L_1} requires the existence of density functions it is not generally defined. However it can be extended to a general setting using Radon-Nikodym derivatives resulting in the UD. So Property II.6 holds.

Since Δ_{L_1} is the continuous version of the UD and the UD satisfies Properties II.7 and II.8 so does Δ_{L_1}.

The Δ_{SD} dependency function

For X, Y equipped with density functions f, g and joint density function h we have

$$\Delta_{SD}(X,Y) = \int_{\mathcal{E}_X \bigotimes \mathcal{E}_Y} (h(x,y) - f(x)g(y))^2 dxdy.$$

By definition it is symmetric so Property II.1 fails.

Since it is equal to 0 if and only if $h(x,y) = f(x)g(y)$ a.e. it is equal to 0 if and only if X and Y are independent so Property II.3 holds.

Let X, Y be jointly Gaussian with non-zero correlation (and also not equal to -1 or 1). Then $\Delta_{SD}(X,Y) > 0$. If we define $X_\alpha = \alpha \cdot X$ and similarly $Y_\beta = \beta \cdot Y$ then the following holds:

$$\begin{aligned}
\Delta_{SD}(X_\alpha, Y_\beta) &= \int_{\mathcal{E}_X \bigotimes \mathcal{E}_Y} (h_{\alpha,\beta}(\alpha x, \beta y) - f_\alpha(\alpha x) g_\beta(\beta y))^2 d(\alpha x) d(\beta y) \\
&= \int_{\mathcal{E}_X \bigotimes \mathcal{E}_Y} \left(\frac{1}{\alpha\beta} h(x,y) - \frac{1}{\alpha} f(x) \frac{1}{\beta} g(y) \right)^2 \alpha\beta dxdy \\
&= \frac{1}{\alpha\beta} \Delta_{SD}(X,Y),
\end{aligned}$$

since there are no restrictions on α or β this means multiple things: it can take values anywhere in $[0, \infty)$ and therefore Property II.2 fails, and for $\alpha = 1$ and $\beta < 1$ we have that Properties II.7 and II.8 are violated.

Since it requires the existence of joint distribution functions it is not generally defined. For example it is undefined for $X = Y$ with $X \sim U(0,1)$ due to the singular nature of $h(x,y)$. So Property II.6 fails. Note that there is also no easy extension by considering Radon-Nikodym derivatives since unlike the Δ_{L_1} case where the choice of reference measure does not change the result, the value here would vary wildly depending on the reference measure.

Finally if Y and X_i are as described in Property II.5 then αY and αX_i are also as described in Property II.5 so for it to hold we need

$$\Delta_{SD}(X_i, Y) = p_i = \Delta_{SD}(\alpha X_i, \alpha Y) = \frac{1}{\alpha^2} \Delta_{SD}(X_i, Y),$$

which does not hold so Property II.5 fails.

The Δ_{ST} dependency function

For X, Y equipped with density functions f, g and joint density function h we have

$$\Delta_{ST}(X,Y) = \int_{\mathcal{E}_X \bigotimes \mathcal{E}_Y} (h(x,y) - f(x)g(y))h(x,y)dxdy.$$

It is symmetric by definition so Property II.1 fails.

For X, Y taking values in $\{0,1\}$ with $\mathbb{P}(X=0, Y=0) = 0.355$, $\mathbb{P}(X=0, Y=1) = 0.245$, $\mathbb{P}(X=1, Y=0) = 0.245$, $\mathbb{P}(X=1, Y=1) = 0.355$ we have that

$$\begin{aligned}\Delta_{ST}(X,Y) &= 0.355 \cdot (0.355 - 0.6 \cdot 0.6) + 0.245 \cdot (0.245 - 0.6 \cdot 0.4) \\ &\quad + 0.245 \cdot (0.245 - 0.4 \cdot 0.6) + 0.155 \cdot (0.155 - 0.4 \cdot 0.4) \\ &= 0.51 \cdot (-0.005) + 0.49 \cdot (0.005) = -0.0001 \notin [0,1],\end{aligned}$$

so Property II.2 fails.

Taking the example above but with 0.35 instead of 0.355, 0.25 instead of 0.245 and 0.15 instead of 0.155 results in

$$\Delta_{ST}(X,Y) = 0.50 \cdot (-0.01) + 0.50 \cdot (0.01) = 0,$$

but do not have independence! So Property II.3 is violated.

If $X = Y$ with $\mathbb{P}(X=1) = \mathbb{P}(X=0) = \frac{1}{2}$ then

$$\Delta_{ST} = \frac{1}{2}\left(\frac{1}{2} - \frac{1}{4}\right) + \frac{1}{2}\left(\frac{1}{2} - \frac{1}{4}\right) = \frac{1}{4} \neq 1,$$

so Property II.4 is violated.

Our counterexample for Property II.4 also works for Property II.5 For the case $N = 1$ and $p_i = 1$. So Property II.5 fails.

Since it requires the existence of density functions it is not generally defined. Take for example once more the case that $X = Y$ with $X \sim U(0,1)$. Due to the singularity of $h(x,y)$ it is not defined.

So Property II.6 fails. An extension based on Radon-Nikodym derivatives will fail for the same reasons as for Δ_{SD}.

Let $X_\alpha = \alpha X$ and $Y_\beta = \beta Y$ for some continuous X, Y with $\Delta_{ST}(X, Y) \neq 0$, then through a similar derivation as was performed for Δ_{SD} we obtain once more

$$\Delta_{ST}(X_\alpha, Y_\beta) = \frac{1}{\alpha\beta}\Delta_{ST}(X, Y),$$

so Properties II.7 and II.8 are violated.

Monotone correlation

For X, Y random variables, have

$$\rho^*(X, Y) = \sup_{f,g} \rho(f(X), g(Y)),$$

where the supremum is taken over all monotonic functions f, g with $0 < \mathrm{Var}(f(X)) < \infty, 0 < \mathrm{Var}(g(Y)) < \infty$.

Note that this is the same as

$$\max(\sup_{f,g} \rho(f(X), g(Y)), -\inf_{f,g} \rho(f(X), g(Y))),$$

where f, g restricted to increasing functions.

It is symmetric by definition so Property II.1 is violated.

Since for any monotone f we have $f'(x) = -f(x)$ monotone, we have that $\rho^*(X, Y) \geq 0$ and since ρ is restricted to $[-1, 1]$ we have that $\rho^*(X, Y) \leq 1$ so it takes values in [0,1]. So Property II.2 is satisfied.

If X, Y independent then so are $f(X), g(Y)$ for all f, g so $\rho^*(X, Y) = 0$. If X, Y are not independent then there exist a, b such that $\mathbb{P}(X \in [a, \infty), Y \in [b, \infty)) \neq \mathbb{P}(X \in [a, \infty))\mathbb{P}(Y \in [b, \infty))$. Suppose $\mathbb{P}(X \in [a, \infty), Y \in [b, \infty)) > \mathbb{P}(X \in [a, \infty))\mathbb{P}(Y \in [b, \infty))$, then consider $f = \mathbb{1}_{[a,\infty)}$ and $g = \mathbb{1}_{[b,\infty)}$. Then $\mathrm{Cov}(f(X), g(Y)) = \mathbb{P}(X \in [a, \infty), Y \in [b, \infty)) - \mathbb{P}(X \in [a, \infty))\mathbb{P}(Y \in [b, \infty)) > 0$ so $\rho(f(X), g(Y)) > 0$ so $\rho^*(X, Y) > 0$. Suppose instead the inequality is the other way around. Then simply use $f = -\mathbb{1}_{[a,\infty)}$ instead for

the same result. So therefore X, Y are independent iff the monotone correlation is equal to 0 so Property II.3 is satisfied.

Suppose X, Y take values in $\{0, 1, 2, 3\}$ with

$\mathbb{P}(X = i, Y = j)$	$i = 0$	$i = 1$	$i = 2$	$i = 3$
$j = 0$	0	$\frac{1}{4}$	0	0
$j = 1$	0	0	0	$\frac{1}{4}$
$j = 2$	$\frac{1}{4}$	0	0	0
$j = 3$	0	0	$\frac{1}{4}$	0

then since correlation is invariant under translation and scaling by a positive real number we can restrict the supremum to increasing functions f, g that map 0 to 0 and 3 to 1. Then define $a_i = f(i)$, $b_j = g(j)$ for $i, j = 1, 2$. Then we obtain $\mathbb{E}(XY) = \frac{1}{4}(b_1 + a_2)$, $\mathbb{E}(X) = \frac{1}{4}(a_1 + a_2 + 1)$, $\mathbb{E}(Y) = \frac{1}{4}(b_1 + b_2 + 1)$, $\mathbb{E}(X^2) = \frac{1}{4}(a_1^2 + a_2^2 + 1)$, $\mathbb{E}(Y^2) = \frac{1}{4}(b_1^2 + b_2^2 + 1)$ so then we obtain

$$\begin{aligned}
\mathrm{Cov}(f(X), g(Y)) &= \frac{1}{16}(4b_1 + 4a_2 - (a_1 + a_2 + 1)(b_1 + b_2 + 1)) \\
\mathrm{Var}(X) &= \frac{1}{16}(3(a_1^2 + a_2^2 + 1) - 2a_1a_2 - 2a_1 - 2a_2) \\
\mathrm{Var}(Y) &= \frac{1}{16}(3(b_1^2 + b_2^2 + 1) - 2b_1b_2 - 2b_1 - 2b_2).
\end{aligned}$$

Now since we have a common $\frac{1}{16}$ which will cancel in the expression for ρ we shall now consider these terms times 16.

Now note that $16\mathrm{Var}(X)$ can be reduced to $(a_1^2 + (1 - a_1)^2) + (a_2^2 + (1 - a_2^2)) + (a_1 - a_2)^2 + 1$ which is lower bounded by $2 + (a_1 - a_2)^2$, and similarly $16\mathrm{Var}(Y)$ is lower bounded by $2 + (b_1 - b_2)^2$.

We will first establish an upper bound for $\rho(f(X), g(Y))$. Note that $16\mathrm{Cov}(f(X), g(Y))$ is always increasing in b_1, a_2 and decreasing in b_2, a_1. Also note that due to the nature of our definition $b_1 \leq b_2$. We will now consider two cases: $a_2 - a_1 \geq \frac{1}{2}$ and $a_2 - a_1 \leq \frac{1}{2}$.

Case 1: since the expression for $\mathrm{Cov}(f(X), g(Y))$ is continuous and defined on a closed bounded set it has a maximum. For this maximum it holds that a_1 cannot be decreased whilst still satisfying constraints and similarly a_2 cannot be increased, so therefore $a_1 = 0$,

$a_2 = 1$. Similarly we should not be able to increase b_1 or decrease b_2 so $b_1 = b_2$. Plugging it in we obtain that $16\text{Cov}(f(X), g(Y)) \leq 2$, but at the same time we have $16\text{Var}(X) \geq 2 + \frac{1}{4}$ and $16\text{Var}(Y) \geq 2$. So $\rho(f(X), g(Y)) \leq \frac{8}{9}$.

Case 2: since the expression for $\text{Cov}(f(X), g(Y))$ is continuous and defined on a closed bounded set it has a maximum. For this maximum it holds that a_1 cannot be decreased whilst still satisfying constraints and similarly a_2 cannot be increased, so therefore $a_2 = \frac{1}{2} + a_1$. Similarly we should not be able to increase b_1 or decrease b_2 so $b_1 = b_2$. Plugging this in we obtain

$$\begin{aligned} 16\text{Cov}(f(X), g(Y)) &= 4b_1 + 4\left(\frac{1}{2} + a_1\right) - \left(2a_1 + \frac{3}{2}\right)(2b_1 + 1) \\ &= b_1 + 2a_1 + \frac{1}{2} - 4a_1 b_1 \\ &= b_1(1 - 4a_1) + 2a_1 + \frac{1}{2}. \end{aligned}$$

Now we can take the maximum over b_1 we have that for $a_1 \leq \frac{1}{4}$ this is obtained when $b_1 = 1$ and it simplifies to $\frac{3}{2} - 2a_1 \leq \frac{3}{2}$, for $a_1 > \frac{1}{4}$ it is obtained when $b_1 = 0$ so it simplifies to $2a_1 + \frac{1}{2} \leq \frac{3}{2}$. So it is upper bounded by $\frac{3}{2}$, and so since at the same time $16\text{Var}(X) \geq 2$ and $16\text{Var}(Y) \geq 2$ we have that $\rho(f(X), g(Y)) \leq \frac{3}{4}$.

So in both cases we have $\rho(f(X), g(Y)) \leq \frac{8}{9}$ and similarly we can show that we have a lower bound of $-\frac{8}{9}$.

So the monotone correlation of X and Y is at most $\frac{8}{9}$ which is strictly less than 1. So Property II.4 fails.

Let Y_0, Y_1, S be independent with $Y_0 \sim U(\{0, 1\})$, $Y_1 \sim U(\{2, 3\}$, $S \sim U(\{0, 1\}$ and $X = Y_S$. Then for the monotonic function f which sends 0,1,2 to 0, and 3 to 20 we have $\mathbb{E}(f(X)f(Y_1)) = 400 \cdot \frac{1}{4} = 100$, $\mathbb{E}(f(X)) = 20 \cdot \frac{1}{4} = 5$, $\mathbb{E}(f(Y_1)) = 20 \cdot \frac{1}{2} = 10$, $\mathbb{E}(f(X)^2) = \frac{1}{4} \cdot 400 = 100$, and $\mathbb{E}(f(Y_1)^2) = \frac{1}{2} \cdot 400 = 200$ so

$$\rho(f(X), f(Y_1)) = \frac{100 - 10 \cdot 5}{\sqrt{100 - 5 \cdot 5}\sqrt{200 - 10 \cdot 10}} = \frac{1}{\sqrt{3}},$$

so the monotone correlation is at least $\frac{1}{\sqrt{3}}$ which is larger than $\frac{1}{2}$. So Property II.5 fails.

Since monotone functions are only defined when the original space the random variables X, Y take values in is equipped with an ordering, we have that Property II.6 fails.

Taking the counterexample for Property 4, and the bijection f which maps $f(0) = 2$, $f(1) = 0$, $f(2) = 3$, and $f(3) = 1$, then we have that $Y = f(X)$ and so the correlation and therefore the monotone correlation is equal to 1 which is strictly larger than the monotone correlation of X and Y. So Properties II.7, and II.8 fail.

Maximal correlation

For X, Y random variables, have

$$\rho' = \sup_{f,g} \rho(f(X), g(Y)),$$

where the supremum is taken over all Borel-measurable functions f, g with $0 < Var(f(X)) < \infty, 0 < Var(g(Y)) < \infty$.

It is symmetric by definition so therefore Property II.1 fails.

Since ρ' is at least ρ^* we have that it is lower bounded by 0. It is also upper bounded by the upper bound of correlation, namely 1. So it is restricted to $[0, 1]$. So Property II.2 is satisfied.

If X, Y are independent then so are $f(X), g(Y)$ for all f, g so $\rho' = 0$. If $\rho'(X, Y) = 0$ then so is $\rho^*(X, Y)$ so X, Y are independent. So Property II.3 is satisfied.

If $Y = f(X)$ for some Borel measurable function then $1 \geq \rho'(X, Y) \geq \rho(Y, f(X) = \rho(Y, Y) = 1$, so $\rho'(X, Y) = 1$. So Property II.4 is satisfied.

Take the counterexample for the monotone correlation case, then $\rho'(X, Y_1) \geq \rho^*(X, Y_1) \geq \frac{1}{\sqrt{3}} > \frac{1}{2}$ so Property II.5 fails.

Since any random variable X that has at least one event A with $0 < \mathbb{P}(A) < 1$ has at least one Borel-measurable function f with finite non-zero variance $f(X)$, we have that Property II.6 is satisfied.

For bijections f_b, g_b have

$$\begin{aligned}
\rho'(X,Y) &= \sup_{f,g} \rho(f(X), g(Y)) \\
&= \sup_{f\circ f_b^{-1}, g\circ g_b^{-1}} \rho(f\circ f_b^{-1}(f_b(X)), g\circ g_b^{-1}(g_b(Y)) \\
&\leq \sup_{f',g'} \rho(f'(f_b(X)), g'(g_b(Y))) \\
&= \sup_{f'\circ f_b, g'\circ g_b} \rho(f'\circ f_b(X), g'\circ g_b(Y)) \\
&\leq \sup_{f,g} \rho(f(X), g(Y)) \\
&= \rho'(X,Y).
\end{aligned}$$

So $\rho'(X,Y) = \rho'(f_b(X), g_b(Y))$ so Property II.7 is satisfied.

For a measurable function f' have that

$$\begin{aligned}
\rho'(f'(X), Y) &= \sup_{f,g} \rho(f(f'(X)), g(Y)) \\
&= \sup_{f\circ f', g} \rho(f\circ f'(X), g(Y)) \\
&\leq \sup_{f,g} \rho(f(X), g(Y)) \\
&= \rho'(X,Y),
\end{aligned}$$

so Property II.8 is satisfied.

Distance correlation

Let (X,Y), (X',Y'), and (X'',Y'') be i.i.d., then the distance covariance is specified as

$$\mathcal{V}^2(X,Y) = \mathrm{Cov}(||X-X'||, ||Y-Y'||) - 2\mathrm{Cov}(||X-X''||, ||Y-Y'||).$$

Then the distance correlation is specified by

$$\mathcal{R}^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)}},$$

as long as neither $\mathcal{V}^2(X,X)$ or $\mathcal{V}^2(Y,Y)$ is equal to 0. If either of them is equal to 0 then $\mathcal{R}^2(X,Y)$ is set to 0.

It is symmetric by definition so Property II.1 fails.

Szekely et al. [150] showed that it takes values in the range $[0, 1]$ so Property II.2 is satisfied.

They also showed that it is 0 if and only if X, Y independent so Property II.3 is satisfied.

Finally, they showed that if the distance correlation of X and Y is 1 it is necessarily the case that $X = a + bCY$ for some non-negative b and orthonormal matrix C. So since it is possible for Y to be determined by X in a non-linear fashion we have that Property II.4 fails.

For Y_0, Y_1, S i.i.d. $U(\{0, 1\})$ and $X = Y_S$ (so X is also distributed $U(\{0, 1\})$) then

$$\begin{aligned}\mathcal{V}^2(Y_1, Y_1) &= \mathcal{V}^2(X, X) \\ &= \mathrm{Cov}(||X - X'||, ||X - X'||) \\ &\quad - 2\mathrm{Cov}(||X - X''||, ||X - X'||) \\ &= \left(\frac{1}{2} - \frac{1}{2}^2\right) - 2\left(\frac{1}{4} - \frac{1}{2}^2\right) \\ &= \frac{1}{4}.\end{aligned}$$

Now through some easy bookkeeping it follows that for $(Y_1', X'), (Y_1'', X'')$ i.i.d. to (Y_1, X) we have that $|Y_1 - Y_1'|$ is distributed $U(\{0, 1\})$ and similarly for $|X - X'|$ and additionally we have that $|X - X''|$ and $|Y_1 - Y_1'|$ are independent and therefore

$$\begin{aligned}\mathcal{V}^2(X, Y_1) &= \mathrm{Cov}(||X - X'||, ||Y_1 - Y_1'||) \\ &\quad - 2\mathrm{Cov}(||X - X''||, ||Y_1 - Y_1'||) \\ &= \mathrm{Cov}(||X - X'||, ||Y_1 - Y_1'||) \\ &= \left(\frac{3}{8} - \frac{1}{2} \cdot \frac{1}{2}\right) \\ &= \frac{1}{8} \neq \frac{1}{2},\end{aligned}$$

so Property II.5 is violated.

Since the distance correlation requires the random variables to be real-valued vectors, it is not generally defined so Property II.6 fails.

Let S, X_0, Y_0 be idd $U(\{0,1\})$ and $X_t = tS + (1-S)X_0$, $Y_t = tS + (1-S)Y_0$ for $t \geq 2$.

Then the following can be shown through some bookkeeping:

$$\mathcal{V}^2(X_t, Y_t) = \frac{1}{4}t^2 - \frac{3}{8}t + \frac{9}{64},$$

and

$$\mathcal{V}^2(X_t, X_t) = \frac{1}{4}t^2 - \frac{3}{8}t + \frac{17}{64} = \mathcal{V}^2(Y_t, Y_t),$$

so therefore it holds that

$$\mathcal{R}^2(X_t, Y_t) = 1 - \frac{1}{8} \cdot \frac{1}{\frac{1}{4}t^2 - \frac{3}{8}t + \frac{17}{64}},$$

which is increasing in t (since $t \geq 2$). So any bijection that swaps t, s for $t, s \geq 2$ does not preserve the distance correlation. So Property II.7 fails.

If Property II.8 holds, then any bijection on X would have to preserve the distance correlation (since otherwise one could take one of the two directions to increase it violating Property II.8). However by symmetricity this would mean the same holds for Y, but then Property II.7 would hold which it does not. So Property II.8 fails.

Maximum canonical correlation (first)

For two vectors of real-valued random variables X and Y with finite second moments the first term of the maximum canonical correlation is defined as

$$CC(X, Y) = \sup_{a,b} \rho(\langle a, X\rangle, \langle b, Y\rangle),$$

where the supremum is taken over vectors a, b of the same dimension as X,Y respectively.

Note that for the special case that both dimensions are equal to 1, this is simply the absolute value of the Pearson correlation.

Since the definition is symmetric in X and Y we have that Property II.1 fails.

Since for any a we have that $CC(X,Y) \geq \max(\rho(a^TX, b^TY), \rho((-a)^TX, b^TY)) \geq 0$ and we have that $\rho \leq 1$ we have that $CC(X,Y) \in [0,1]$ so Property II.2 holds.

Since Property II.3 fails for the Pearson correlation it also fails for the maximum canonical correlation (since the absolute value is 0 iff the underlying value is 0).

Let $X \sim U([0,1])$, and $Y = X^2$, then X determines Y but has a Pearson correlation of 0 and therefore a maximum canonical correlation of 0. So Property II.4 fails.

We stay in the special case that both dimensions are equal to 1. For $Y_1, Y_2, \ldots, Y_n, S$ independent with $\mathbb{P}(S = i) = p_i$ and $X = Y_S$ have that $\mathbb{E}(XY_i) - \mathbb{E}(X)\mathbb{E}(Y_i) = p_i(\mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2) = p_i\text{Var}(Y_i)$ so $CC(X,Y_i) = |\rho(X,Y_i)| = p_i$ if and only if $\text{Var}(X) = \pm\text{Var}(Y_i)$ which is not necessarily the case, so Property II.5 fails.

Since the vectors are required to be real-valued it is not generally defined so Property II.6 fails.

Finally for $Y = X^2$ and $X \sim U(0,1)$ we have that $\mathbb{E}(X) = \frac{1}{2}, \mathbb{E}(X^2) = \mathbb{E}(Y) = \frac{1}{3}, \mathbb{E}(XY) = \frac{1}{4}, \mathbb{E}(Y^2) = \frac{1}{5}$ so we have $CC(X,Y) = |\rho(X,Y)| = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}\frac{4}{45}}} = \frac{\sqrt{15}}{4} < 1$. However applying the $\sqrt{}$-function to Y, we have $CC(X,Y) = |\rho(X,\sqrt{Y})| = \rho(X,X) = 1$ so Property II.7 and Property II.8 fail for the maximum canonical correlation.

Strong mixing coefficient

The strong mixing coefficient of two random variables X,Y is given by

$$\alpha(X,Y) = \sup_{A \in \mathcal{E}_X, B \in \mathcal{E}_Y} \{|\mu_{X,Y}(A \times B) - \mu_X(A)\mu_Y(B)|\}.$$

The strong mixing coefficient is symmetric by definition so Property II.1 fails.

Since $-1 \leq \mu_{X,Y}(A\times B)-\mu_X(A)\mu_Y(B) \leq 1$ we have that $|\mu_{X,Y}(A\times B) - \mu_X(A)\mu_Y(B)| \in [0,1]$ so $\alpha(X,Y) \in [0,1]$

We have that X, Y are independent if and only if $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for all $A \in \mathcal{E}_X, B \in \mathcal{E}_Y$ so if and only if $\alpha(X,Y) = 0$. So Property II.3 holds.

For $X = Y \sim U(\{0,1\})$ it is easy to see that $\alpha(X,Y) = \frac{1}{4} \neq 1$ so Property II.4 fails.

The above example also functions as a counterexample for Property II.5.

Since there are no assumptions made regarding random variables X, Y we have that Property II.6 holds.

Let $f : (E_X, \mathcal{E}_X) \to (E_{X'}, \mathcal{E}_{X'})$ be a measurable function. Then $h : E_X \times E_Y \to E_{X'} \times E_Y$ with $h(x,y) = (f(x), y)$ is measurable. Now it follows that

$$\begin{aligned}
\alpha(f(X),Y) &= \sup_{A\in\mathcal{E}_{X'},B\in\mathcal{E}_Y} \left\{|\mu_{(f(X),Y)}(A\times B) - \mu_{f(X)}(A)\mu_Y(B)|\right\} \\
&= \sup_{A\in\mathcal{E}_{X'},B\in\mathcal{E}_Y} \Big\{|\mu_{(X,Y)}(h^{-1}(A\times B)) \\
&\qquad -(\mu_X\times\mu_Y)(h^{-1}(A\times B))|\Big\} \\
&= \sup_{A\in\mathcal{E}_{X'},B\in\mathcal{E}_Y} \Big\{|\mu_{(X,Y)}(f^{-1}(A)\times B) \\
&\qquad -(\mu_X\times\mu_Y)(f^{-1}(A)\times B)|\Big\} \\
&= \sup_{A\in\mathcal{E}_{X'},B\in\mathcal{E}_Y} \Big\{|\mu_{(X,Y)}(f^{-1}(A)\times B) \\
&\qquad -(\mu_X(f^{-1}(A))\mu_Y(B)|\Big\} \\
&\leq \sup_{A\in\mathcal{E}_X,B\in\mathcal{E}_Y} \left\{|\mu_{(X,Y)}(A\times B) - (\mu_X(A)\mu_Y(B)|\right\} \\
&= \alpha(X,Y),
\end{aligned}$$

so Property II.8 holds, and by symmetricity also fro any measurable g we have $\alpha(X, g(Y)) \leq \alpha(X,Y)$ so Property II.7 follows.

β-mixing coefficient

The β-mixing coefficient is similar to the strong mixing coefficient given above, however with a relaxation, instead of taking the supremum over rectangles, instead it takes the supremum over partitionings into rectangles:

$$\beta(X,Y) = \sup\left\{\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}|\mu_{X,Y}(A_i \times B_j) - \mu_X(A_i)\mu_Y(B_j)|\right\},$$

where the A_i are measurable sets that partition E_X and similarly the B_j measurable sets that partition E_Y.

It can be shown (through a similar approximation argument as used in Section 6.A.2 to establish the upper bound of $\mathrm{UD}(X,Y)$ given Y) that $\beta(X,Y) = \frac{1}{2}\mathrm{UD}(X,Y)$. Therefore we can conclude Property II.1, II.4, and II.5 fail, while Properties II.2, II.3, II.6, II.7, and II.8 hold.

7 The Berkelmans-Pries Feature Importance method: a generic measure of informativeness of features

7.1 Introduction

How important are you? This is a question that researchers (especially data scientists) have wondered for many years. Researchers need to understand how important a random variable (RV) X is for determining Y. Which features are important for predicting the weather? Can indicators be found as symptoms for a specific disease? Can redundant variables be discarded to increase performance? These kinds of questions are relevant in almost any research area. Especially nowadays, as the rise of machine learning models generates the need to demystify prediction models. Altmann et al. [8] state that 'In life sciences, interpretability of machine learning models is as important as their prediction accuracy.' Although this might not hold for all research areas, interpretability is very useful.

Based on [121]: J. Pries, G. Berkelmans, S. Bhulai, R.D. van der Mei. 'The Berkelmans-Pries Feature Importance method: a generic measure of informativeness of features'. Submitted for publication.

Knowing how predictions are made and why, is crucial for adapting these methods in everyday life.

Determining *Feature Importance* (FI) is the art of discovering the importance of each feature X_i when predicting Y. The following two cases are particularly useful. (I) Finding variables that are not important: *redundant* variables can be discovered using FI methods. Irrelevant features could degrade the performance of a prediction model due to high dimensionality and irrelevant information [87]. Eliminating redundant features could therefore increase both the speed and the accuracy of a prediction model. (II) Finding variables that are important: *important* features could reveal underlying structures that give valuable insights. Observing that variable X is important for predicting Y could steer research efforts into the right direction. Although it is critical to keep in mind that high FI does not mean causation. However, FI values do, for example, 'enable an anaesthesiologist to better formulate a diagnosis by knowing which attributes of the patient and procedure contributed to the current risk predicted' [102]. In this way, a FI method can have very meaningful impact.

Over the years, many FI methods have been suggested, which results in a wide range of FI values for the same dataset. For example, stochastic methods do not even repeatedly predict the same FI values. This makes interpretation difficult. Examine e.g., a result of Fryer et al. [61], where one measure assigns a FI of 3.19 to a variable, whereas another method gives the same variable a FI value of 0.265. This raises a lot of questions: 'Which FI method is correct?', 'Is this variable deemed important?', and more generally 'What information does this give us?'. To assess the performance of a FI method, the ground truth should be known, which is often not the case [2, 74, 154, 171]. Therefore, when FI methods were developed, the focus has not yet lied on predicting the *exact* correct FI values. Additionally, many FI methods do not have desirable properties. For example, two features that contain the same amount of information should get the same FI. We later show that this is often not the case.

To improve interpretability, we introduce a new FI method called *Berkelmans-Pries* FI method, which is based on *Shapley values* [139]

and the *Berkelmans-Pries* dependency function [16]. Multiple existing methods already use Shapley values, which has been shown to give many nice properties. However, by *additionally* using the *Berkelmans-Pries* dependency function, even more useful properties are obtained. Notably, we prove that this approach accurately predicts the FI in some cases where the ground truth FI can be derived in an exact manner. By combining *Shapley values* and the *Berkelmans-Pries* dependency function a powerful FI method is created. This chapter is a significant step forward for the field of FI, because of the following reasons:

- We introduce a new FI method;
- We prove multiple useful properties of this method;
- We provide some cases where the ground truth FI can be derived in an exact manner;
- We prove for these cases that our FI method accurately predicts the correct FI;
- We obtain the largest collection of existing FI methods;
- We test if these methods adhere to the same properties, which shows that no method comes close to fulfilling all the useful properties;
- We provide Python code to determine the FI values [29].

7.2 Berkelmans-Pries FI

Recall that Kruskal [89] stated that 'There are infinitely many possible measures of association, and it sometimes seems that almost as many have been proposed at one time or another.' Although this quote was about dependency functions, it could just as well have been about FI methods. Over the years, many FI methods have been suggested, but it remains unclear which method should be used when and why [74]. In this section, we propose yet another new FI method called the *Berkelmans-Pries FI method* (BP-FI). Although it is certainly subjective what it is that someone wants from a FI method, we show in Section 7.3 that BP-FI has many

useful and intuitive properties. The BP-FI method is based on two key elements: (1) *Shapley values* and (2) the *Berkelmans-Pries dependency function*. We will discuss these components first to clarify how the BP-FI method works.

7.2.1 Shapley value approach

The *Shapley value* is a unique game-theoretical way to assign *value* to each *player* participating in a multiplayer *game* based on four axioms [139]. This concept is widely used in FI methods, as it can be naturally adapted to determine how *important* (value) each *feature* (player) is for *predicting a target variable* (game). Let N_{v} be the number of features, then the Shapley value of feature i is defined by

$$\phi_i(v) = \sum_{S \subseteq \{1,\ldots,N_{\mathrm{v}}\} \setminus \{i\}} \frac{|S|! \cdot (N_{\mathrm{v}} - |S| - 1)!}{N_{\mathrm{v}}!} \cdot (v(S \cup \{i\}) - v(S)), \quad (7.1)$$

where $v(S)$ can be interpreted as the 'worth' of the coalition S [139]. The principle behind this formulation can also be explained in words: For every possible sequence of features up to feature i, the added value of feature i is the difference between the worth before it was included (i.e., $v(S)$) and after (i.e., $v(S \cup \{i\})$). Averaging these added values over all possible sequences of features gives the final Shapley value for feature i.

SHAP There are multiple existing FI methods that use Shapley values [51, 61, 104], which immediately ensures some useful properties. The most famous of these methods is SHAP [104]. This method is widely used for *local* explanations (see Section 7.4.1). To measure the local FI for a specific sample x and a prediction model f, the *conditional expectation* is used as characteristic function (i.e., v in Equation (7.1)). Let $x = (x_1, x_2, \ldots, x_{N_{\mathrm{v}}})$, where x_i is the feature value of feature i, then SHAP FI values can be determined using:

$$v_x(S) := \mathbb{E}\left[f(z) | z_i = x_i \text{ for all } i \in S\right]. \quad (7.2)$$

Observe that the characteristic function v_x is defined locally for each x. To get *global* FI values, an average can be taken over all local FI

values. Our novel FI method uses a different characteristic function, namely the *Berkelmans-Pries* dependency function. This leads to many additional useful properties. Furthermore, the focus of this chapter is not on *local* explanations, but *global* FI values.

7.2.2 Berkelmans-Pries dependency function

A new dependency measure, called the *Berkelmans-Pries* dependency function, was introduced in Chapter 6, which is used in the formulation of the BP-FI method. It is shown that the BP dependency function satisfies a list of desirable properties, whereas existing dependency measures did not. It has a measure-theoretical formulation, but this reduces to a simpler and more intuitive version when all variables are discrete [16]. We want to highlight this formulation to give some intuition behind the BP dependency function. It is given by

$$\mathrm{Dep}\left(Y|X\right) := \begin{cases} \frac{\mathrm{UD}(X,Y)}{\mathrm{UD}(Y,Y)} & \text{if } Y \text{ is not a.s. constant,} \\ \text{undefined} & \text{if } Y \text{ is a.s. constant,} \end{cases} \tag{7.3}$$

where (in the discrete case) it holds that

$$\mathrm{UD}\left(X,Y\right) := \sum_x p_X(x) \cdot \sum_y \left| p_{Y|X=x}(y) - p_Y(y) \right|. \tag{7.4}$$

The BP dependency measure can be interpreted in the following manner. The numerator is the expected absolute difference between the distribution of Y and the distribution of Y given X. If Y is highly dependent on X, the distribution changes as knowing X gives information about Y, whereas if Y is independent of X, there is no difference between these two distributions. The denominator is the maximal possible change in distribution of Y for any variable, which is used to standardise the dependency function. Note that the BP dependency function is *asymmetric*: $\mathrm{Dep}\left(Y|X\right)$ is the dependency of Y on X, not vice versa. Due to the many desirable properties, the BP dependency function is used for the BP-FI.

7.2.3 Berkelmans-Pries FI method

Abbreviated notation improves readability of what comes next, which is why we define

$$w(S, N_{\mathrm{v}}) := \frac{|S|! \cdot (N_{\mathrm{v}} - |S| - 1)!}{N_{\mathrm{v}}!}, \tag{7.5}$$

$$D(X, Y, S) := \mathrm{Dep}\,(Y|S \cup \{X\}) - \mathrm{Dep}\,(Y|S)\,. \tag{7.6}$$

One crucial component of translating the game-theoretical approach of Shapley values to the domain of FI is choosing the function v in Equation (7.1). This function assigns for each set of features S a value $v(S)$ that characterises the 'worth' of the set S. How this function is defined, has a critical impact on the resulting FI. We choose to define the 'worth' of a set S to be the BP dependency of Y on the set S, which is denoted by $\mathrm{Dep}\,(Y|S)$ [16]. Here, $\mathrm{Dep}\,(Y|S) := \mathrm{Dep}\,(Y|Z_S(\mathcal{D}))$ where $\mathcal{D}$ denotes the entire dataset with all features and $Z_S(\mathcal{D})$ is the reduction of the dataset to include only the subset of features S. Let Ω_{f} be the set of all feature variables, and $N_{\mathrm{v}} := |\Omega_{\mathrm{f}}|$. Now, for every $S \subseteq \Omega_{\mathrm{f}}$, we define:

$$v(S) := \mathrm{Dep}\,(Y|S)\,. \tag{7.7}$$

In other words, the value of set S is exactly how *dependent* the target variable Y is on the features in S. The difference $v(S \cup \{i\}) - v(S)$ in Equation (7.1) can now be viewed as the increase in dependency of Y on the set of features, when feature i is also known. The resulting Shapley values using the BP dependency function as characteristic function are defined to be the BP-FI outcome. For each feature i, we get:

$$\begin{aligned} \mathrm{FI}(i) &:= \sum_{S \subseteq \Omega_{\mathrm{f}} \setminus \{i\}} \frac{|S|! \cdot (N_{\mathrm{v}} - 1 - |S|)!}{N_{\mathrm{v}}!} \cdot (v(S \cup \{i\}) - v(S)) \\ &= \sum_{S \subseteq \Omega_{\mathrm{f}} \setminus \{i\}} w(S, N_v) \cdot D(\{i\}, Y, S). \end{aligned} \tag{7.8}$$

Note that when Y is *almost surely constant* (i.e., $\mathbb{P}(Y = y) = 1$), $\mathrm{Dep}\,(Y|S)$ is undefined for any feature set S (see Equation (7.3)).

We argue that it is natural to assume that $\mathrm{FI}(i)$ is also undefined, as every feature attributes everything and nothing at the same time. In the remainder of this chapter, we assume that Y is not a.s. constant.

7.3 Properties of BP-FI

Recall that it is hard to evaluate FI methods, as the ground truth FI is often unknown [2, 74, 154, 171]. With this in mind, we want to show that the BP-FI method has many desirable properties. We also give some synthetic cases where the BP-FI method gives a natural expected outcome. The BP-FI method is based on *Shapley values*, which are a unique solution based on four axioms [164]. These axioms already give many characteristics that are preferable for a FI method. Additionally, using the BP dependency function ensures that it has extra desirable properties. In this section, we prove properties of the BP-FI method and discuss why these are relevant and useful.

Property 1 (Efficiency)**.** The sum of all FI scores is equal to the total dependency of Y on all features:

$$\sum_{i \in \Omega_\mathrm{f}} \mathrm{FI}(i) = \mathrm{Dep}\left(Y | \Omega_\mathrm{f}\right).$$

Proof. Shapley values are *efficient*, meaning that all the value is distributed among the players. Thus,

$$\sum_{i \in \Omega_\mathrm{f}} \mathrm{FI}(i) = v(\Omega_\mathrm{f}) = \mathrm{Dep}\left(Y | \Omega_\mathrm{f}\right).$$

Relevance. With our approach, we try to answer the question 'How much did each feature contribute to the total dependency?'. The total 'payoff' is in our case the total dependency. It is therefore natural to divide the entire payoff (but not more than that) amongst all features.

Corollary 1.1. If adding a RV X to the dataset does not give any additional information (i.e., $\mathrm{Dep}\left(Y | \Omega_\mathrm{f} \cup X\right) = \mathrm{Dep}\left(Y | \Omega_\mathrm{f}\right)$), then the sum of all FI remains the same.

Proof. This directly follows from Property 1.

Relevance. If the collective knowledge remains the same, the same amount of credit is available to be divided amongst the features. Only when new information is added, an increase in combined credit is warranted. A direct result of this corollary is that adding a *clone* (i.e., $X^{\text{clone}} := X$) of a variable X to the dataset will never increase the total sum of FI.

Property 2 (Symmetry)**.** If for every $S \subseteq \Omega_{\text{f}} \setminus \{i, j\}$ it holds that $\text{Dep}\,(Y|S \cup \{i\}) = \text{Dep}\,(Y|S \cup \{j\})$, then $\text{FI}(i) = \text{FI}(j)$.

Proof. Shapley values are defined *symmetrically*, meaning that if $v(S \cup \{i\}) = v(S \cup \{j\})$ for every $S \subseteq \Omega_{\text{f}} \setminus \{i, j\}$, it follows that $\text{FI}(i) = \text{FI}(j)$. Thus, it automatically follows that BP-FI is also symmetric.

Relevance. If two variables are interchangeable, meaning that they *always* contribute equally to the dependency, it is only sensible that they obtain the same FI. This is a desirable property for a FI method, as two features that contribute equally should obtain the same FI.

Property 3 (Range)**.** For any RV X, it holds that $\text{FI}(X) \in [0, 1]$.

Proof. The BP dependency function is *non-increasing* under functions of X [16], which means that for any measurable function f it holds that

$$\text{Dep}\,(Y|f(X)) \leq \text{Dep}\,(Y|X)\,.$$

Take $f := Z_S$, which is the function that reduces $\mathcal{D}$ to the subset of features in S. Using the non-increasing property of BP dependency function, it follows that:

$$\begin{aligned}\text{Dep}\,(Y|S) &= \text{Dep}\,(Y|Z_S(\mathcal{D})) = \text{Dep}\left(Y|Z_S(Z_{S\cup\{i\}}(\mathcal{D}))\right) \\ &\leq \text{Dep}\left(Y|Z_{S\cup\{i\}}(\mathcal{D})\right) = \text{Dep}\,(Y|S \cup \{i\})\,, \qquad (7.9)\end{aligned}$$

Examining Equation (7.8), we observe that every FI value must be greater or equal to zero, as $D(\{i\}, Y, S) = \text{Dep}\,(Y|S \cup \{i\}) - \text{Dep}\,(Y|S) \geq 0$.

One of the properties of the BP dependency function is that for any X, Y it holds that $\mathrm{Dep}(Y|X) \in [0,1]$ [16]. Using Property 1, the sum of all FI values must therefore be in $[0,1]$, as $\sum_{i\in\Omega_\mathrm{f}} \mathrm{FI}(i) = \mathrm{Dep}(Y|\Omega_\mathrm{f}) \in [0,1]$. This gives an upper bound for the FI values, which is why we can now conclude that $\mathrm{FI}(X) \in [0,1]$ for any RV X.

Relevance. It is essential for interpretability that a FI method is bounded by known bounds. For example, a FI score of 4.2 cannot be interpreted properly, when the upper or lower bound is unknown.

Property 4 (Bounds)**.** Every $\mathrm{FI}(X)$ with $X \in \Omega_\mathrm{f}$ is bounded by

$$\frac{\mathrm{Dep}(Y|X)}{N_\mathrm{v}} \leq \mathrm{FI}(X) \leq \mathrm{Dep}(Y|\Omega_\mathrm{f}).$$

Proof. The upper bound follows from Properties 1 and 3, as

$$\mathrm{Dep}(Y|\Omega_\mathrm{f}) = \sum_{i\in\Omega_\mathrm{f}} \mathrm{FI}(i) \geq \mathrm{FI}(X),$$

where the last inequality follows since $\mathrm{FI}(i) \in [0,1]$ for all $i \in \Omega_\mathrm{f}$.

The lower bound can be established using the inequality from Equation (7.9) within Equation (7.8). This gives

$$\begin{aligned}\mathrm{FI}(X) &= \sum_{S\subseteq\Omega_\mathrm{f}\setminus\{X\}} w(S, N_v)\cdot D(X,Y,S)\\ &\geq \frac{0!\cdot(N_\mathrm{v}-0-1)!}{N_\mathrm{v}!}\cdot(\mathrm{Dep}(Y|\varnothing\cup\{X\}) - \mathrm{Dep}(Y|\varnothing))\\ &= \frac{\mathrm{Dep}(Y|X)}{N_\mathrm{v}}.\end{aligned}$$

Relevance. These bounds are useful for upcoming proofs.

Property 5 (Zero FI)**.** For any RV X, it holds that

$$\mathrm{FI}(X) = 0 \Leftrightarrow \mathrm{Dep}(Y|S\cup\{X\}) = \mathrm{Dep}(Y|S) \text{ for all } S \in \Omega_\mathrm{f}\setminus\{X\}.$$

Proof. $\Leftarrow$: When $\mathrm{Dep}\,(Y|S \cup \{X\}) = \mathrm{Dep}\,(Y|S)$ for all $S \in \Omega_\mathrm{f} \setminus \{X\}$, it immediately follows from Equation (7.8) that

$$\begin{aligned}\mathrm{FI}(X) &= \sum_{S \subseteq \Omega_\mathrm{f} \setminus \{X\}} w(S, N_v) \cdot D(X, Y, S) \\ &= \sum_{S \subseteq \Omega_\mathrm{f} \setminus \{X\}} w(S, N_v) \cdot 0 \\ &= 0.\end{aligned}$$

$\Rightarrow$: Assume that $\mathrm{FI}(X) = 0$. It follows from the proof of Property 3 that $D(X, Y, S) \geq 0$ for every $S \subseteq \Omega_\mathrm{f} \setminus \{X\}$. If $D(X, Y, S^*) > 0$ for some given $S^* \in \Omega_\mathrm{f} \setminus \{X\}$, it follows from Equation (7.8) that

$$\begin{aligned}\mathrm{FI}(X) &= \sum_{S \subseteq \Omega_\mathrm{f} \setminus \{X\}} w(S, N_v) \cdot D(X, Y, S) \\ &\geq w(S^*, N_v) \cdot D(X, Y, S^*) \\ &> 0.\end{aligned}$$

This gives a contradiction with the assumption that $\mathrm{FI}(X) = 0$, thus it is not possible that such an S^* exists. This means that $\mathrm{Dep}\,(Y|S \cup \{X\}) = \mathrm{Dep}\,(Y|S)$ for all $S \in \Omega_\mathrm{f} \setminus \{X\}$.

Relevance. When a feature *never* contributes any information, it is only fair that it does not receive any FI. The feature can be removed from the dataset, as it has no effect on the target variable. On the other hand, when a feature has a FI of zero, it would be unfair to this feature if it does in fact contribute information somewhere. It should then be rewarded some FI, albeit small it should be larger than zero.

Null-independence The property that a feature gets zero FI, when $\mathrm{Dep}\,(Y|S \cup \{X\}) = \mathrm{Dep}\,(Y|S)$ for all $S \in \Omega_\mathrm{f} \setminus \{X\}$ is the same notion as a *null player* in game theory. Berkelmans et al. [16] show that $\mathrm{Dep}\,(Y|X) = 0$, when Y is *independent* of X. To be a *null player* requires a stricter definition of independence, which we call *null-independence*. Y is null-independent on X if $\mathrm{Dep}\,(Y|S \cup \{X\}) = \mathrm{Dep}\,(Y|S)$ for all $S \in \Omega_\mathrm{f} \setminus \{X\}$. In other words, X is null-independent if and only if $\mathrm{FI}(X) = 0$.

Corollary 5.1. Independent feature $\not\Rightarrow$ null-independent feature.

Proof. Take, e.g., the dataset consisting of two binary features $X_1, X_2 \sim \mathcal{U}(\{0,1\})$ and a target variable $Y = X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$ which is the XOR of X_1 and X_2. Individually, the variables do not give any information about Y, whereas collectively they fully determine Y. In the proof of Property 15, we show that this leads to $\mathrm{FI}(X_1) = \mathrm{FI}(X_2) = \frac{1}{2}$, whilst $\mathrm{Dep}\,(Y|X_1) = \mathrm{Dep}\,(Y|X_2) = 0$. Thus, X_1 and X_2 are *independent*, but not *null-independent*.

Corollary 5.2. Independent feature $\Leftarrow$ null-independent feature.

Proof. When X is *null-independent*, it holds that $\mathrm{FI}(X) = 0$. Using Property 4, we obtain

$$0 = \mathrm{FI}(X) \geq \frac{\mathrm{Dep}\,(Y|X)}{N_{\mathrm{v}}} \Leftrightarrow \mathrm{Dep}\,(Y|X) = 0.$$

Thus, when X is *null-independent*, it is also *independent*.

Corollary 5.3. Almost surely constant variables get zero FI.

Proof. If X is *almost surely constant* (i.e., $\mathbb{P}(X = x) = 1$), it immediately follows that $\mathrm{Dep}\,(Y|S \cup \{X\}) = \mathrm{Dep}\,(Y|S)$ for any $S \subseteq \Omega_{\mathrm{f}} \setminus \{X\}$, as the distribution of Y is not affected by X.

Property 6 (FI equal to one)**.** When $\mathrm{FI}(X) = 1$, it holds that $\mathrm{Dep}\,(Y|X) = 1$ and all other features are null-independent.

Proof. As the BP dependency function is bounded by $[0,1]$ [16], it follows from Property 1 that $\sum_{i \in \Omega_{\mathrm{f}}} \mathrm{FI}(i) \leq 1$. Noting that each FI must be in $[0,1]$ due to Property 3, we find that

$$\mathrm{FI}(X) = 1 \Rightarrow \mathrm{FI}(X') = 0 \text{ for all } X' \in \Omega_{\mathrm{f}} \setminus \{X\}.$$

Thus all other features are *null-independent*. Next, we show that $\mathrm{Dep}\,(Y|X) = 1$ must also hold, when $\mathrm{FI}(X) = 1$. Assume that

$\mathrm{Dep}\,(Y|X) < 1$. Using Equation (7.8) we find that

$$\begin{aligned}
1 = \mathrm{FI}(X) &= \sum_{S\subseteq\Omega_\mathrm{f}\setminus\{X\}} w(S, N_\mathrm{v})\cdot D(X,Y,S) \\
&= \sum_{S\subseteq\Omega_\mathrm{f}\setminus\{X\}:|S|>0} w(S, N_\mathrm{v})\cdot D(X,Y,S) + w(\varnothing, N_\mathrm{v})\cdot D(X,Y,\varnothing) \\
&\leq \sum_{S\subseteq\Omega_\mathrm{f}\setminus\{X\}:|S|>0} w(S, N_\mathrm{v})\cdot (1-0) + w(\varnothing, N_\mathrm{v})\cdot (\mathrm{Dep}\,(Y|X) - 0) \\
&< \sum_{S\subseteq\Omega_\mathrm{f}\setminus\{X\}} w(S, N_\mathrm{v}) \\
&= \sum_{k=0}^{N_\mathrm{v}-1} \binom{N_\mathrm{v}-1}{k} \cdot \frac{k!\cdot(N_\mathrm{v}-k-1)!}{N_\mathrm{v}!} \\
&= \sum_{k=0}^{N_\mathrm{v}-1} \frac{(N_\mathrm{v}-1)!}{k!\cdot(N_\mathrm{v}-1-k)!} \cdot \frac{k!\cdot(N_\mathrm{v}-k-1)!}{N_\mathrm{v}!} \\
&= \sum_{k=0}^{N_\mathrm{v}-1} \frac{1}{N_\mathrm{v}} \\
&= 1.
\end{aligned}$$

Note that the inequality step follows from the range of the BP dependency function (i.e., $[0,1]$). The largest possible addition is when $\mathrm{Dep}\,(Y|S\cup\{X\}) - \mathrm{Dep}\,(Y|S) = 1 - 0 = 1$. This result gives a contradiction, as $1 < 1$ cannot be true, which means that $\mathrm{Dep}\,(Y|X) = 1$.

Relevance. When a variable gets a FI of one, the rest of the variables should be zero. Additionally, it should mean that this variable contains the necessary information to fully determine Y, which is why $\mathrm{Dep}\,(Y|X) = 1$ should hold.

Property 7. $\mathrm{Dep}\,(Y|X) = 1 \not\Rightarrow \mathrm{FI}(X) = 1$.

Proof. As counterexample, examine the case where there are multiple variables that fully determine Y. Properties 1 and 3 must still hold. Thus, if FI is one for every variable that fully determines Y, we get

$$\sum_{i\in\Omega_\mathrm{f}} \mathrm{FI}(i) \geq 1 + 1 \neq 1 = \mathrm{Dep}\,(Y|\Omega_\mathrm{f})\,,$$

which is a contradiction.

Relevance. This property is important for interpretation of the FI score. When $\mathrm{FI}(X) \neq 1$, it cannot be automatically concluded that Y is not fully determined by X.

If Y is fully determined by X, we call X *fully informative*, as it gives all information that is necessary to determine Y.

Property 8 (Max FI when fully informative)**.** If X is fully informative, it holds that $\mathrm{FI}(i) \leq \mathrm{FI}(X)$ for any $i \in \Omega_{\mathrm{f}}$.

Proof. Assume that there exists a feature i such that $\mathrm{FI}(i) > \mathrm{FI}(X)$, when Y is fully determined by X. To attain a higher FI, somewhere in the sum of Equation (7.8), a higher gain must be made by i compared to X. Observe that for any $S \subseteq \Omega_{\mathrm{f}} \setminus \{i, X\}$ it holds that

$$D(\{i\}, Y, S) \leq 1 - \mathrm{Dep}\,(Y|S) = D(X, Y, S).$$

For any $S \subseteq \Omega_{\mathrm{f}} \setminus \{i\}$ with $X \in S$, it holds that

$$\mathrm{Dep}\,(Y|S \cup \{i\}) - \mathrm{Dep}\,(Y|S) = 1 - 1 = 0.$$

The last step follows from Equation (7.9), as the dependency function is increasing, thus $\mathrm{Dep}\,(Y|S \cup \{i\}) = 1$. In other words, no possible gain can be achieved with respect to X in the Shapley values. Therefore, it cannot hold that $\mathrm{FI}(i) > \mathrm{FI}(X)$.

Relevance. Whenever a variable fully determines Y, it should attain the highest FI. What would a FI higher than such a score mean? It gives more information than the maximal information? When this property would not hold, it would result in a confusing and difficult interpretation process.

Property 9 (Limiting the outcome space)**.** For any measurable function f and RV X, replacing X with $f(X)$ never increases the assigned FI to this variable.

Proof. The BP dependency function is non-increasing under functions of X [16]. This means that for any measurable function g, it

holds that

$$\mathrm{Dep}\,(Y|g(X)) \leq \mathrm{Dep}\,(Y|X)\,.$$

Choose g to be the function that maps the union of any feature set S and the original RV X to the union of S and the replacement $f(X)$. In other words $g(S \cup \{X\}) = S \cup \{f(X)\}$ for any feature set S. It then follows that:

$$\mathrm{Dep}\,(Y|S \cup \{f(X)\}) = \mathrm{Dep}\,(Y|g(S \cup \{X\})) \leq \mathrm{Dep}\,(Y|S \cup \{X\})\,,$$

and thus

$$D(f(X), Y, S) \leq D(X, Y, S),$$

for any $S \subseteq \Omega_{\mathrm{f}} \setminus \{X\}$. Thus, using Equation (7.8), we can conclude that replacing X with $f(X)$ never increases the assigned FI.

Relevance. This is an important observation for preprocessing. Whenever a variable is binned, it would receive less (or equal) FI when less bins are used. It could also potentially provide a useful upper bound, when the FI is already known before replacing X with $f(X)$.

Corollary 9.1. For any measurable function f and RV X, when $X = f(X')$ for another RV X', replacing feature X by feature X' will never decrease the assigned FI.

Proof. When $X = f(X')$ holds, it follows again (similar to Property 9) that

$$\mathrm{Dep}\,(Y|S \cup \{X\}) = \mathrm{Dep}\,(Y|S \cup \{f(X')\}) \leq \mathrm{Dep}\,(Y|S \cup \{X'\})$$

for any $S \subseteq \Omega_{\mathrm{f}} \setminus \{X\}$. Therefore, using Equation (7.8), observe that replacing X with X' never decreases the assigned FI.

Shapley values have additional properties when the characteristic function v is *sub-additive* and/or *super-additive* [139]. We show that our function, defined by Equation (7.7), is neither.

Property 10 (Neither sub-additive nor super-additive)**.** Our characteristic function $v(S) = \mathrm{Dep}\,(Y|S)$ is neither *sub-additive* nor *super-additive*.

Proof. Consider the following two counterexamples.

Counterexample sub-additive: A function f is *sub-additive* if for any $S, T \in \Omega_f$ it holds that

$$f(S \cup T) \leq f(S) + f(T).$$

Examine the dataset consisting of two binary features $X_1, X_2 \sim \mathcal{U}(\{0,1\})$ and a target variable $Y = X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$ which is the XOR of X_1 and X_2. Both X_1 and X_2 do not individually give any new information about the distribution of Y, thus $v(X_1) = v(X_2) = 0$ (see properties of the BP dependency function [16]). However, collectively they fully determine Y and thus $v(X_1 \cup X_2) = 1$. We can therefore conclude that v is not sub-additive, as

$$v(X_1 \cup X_2) = 1 \not\leq 0 + 0 = v(X_1) + v(X_2).$$

Counterexample super-additive: A function f is *super-additive* if for any $S, T \in \Omega_f$ it holds that

$$f(S \cup T) \geq f(S) + f(T).$$

Consider the dataset consisting of two binary features $X \sim \mathcal{U}(\{0,1\})$ and a *clone* $X^{\text{clone}} := X$, where the target variable Y is defined as $Y := X$. Note that both X and X^{clone} fully determine Y, thus $v(X) = v(X^{\text{clone}}) = 1$ (see properties of the BP dependency function [16]). Combining X and X^{clone} also fully determines Y, which leads to:

$$v(X \cup X^{\text{clone}}) = 1 \not\geq 1 + 1 = v(X) + v(X^{\text{clone}}).$$

Thus, v is also not super-additive.

Relevance. If the characteristic function v is *sub-additive*, it would hold that $\text{FI}(X) \leq v(X)$ for any $X \in \Omega_f$. When v is *super-additive*, it follows that $\text{FI}(X) \geq v(X)$ for any $X \in \Omega_f$. This is sometimes also referred to as *individual rationality*, which means that no player receives less, than what he could get on his own. This makes sense in a game-theoretic scenario with human players that can decide to not play when one could gain more by not cooperating. In our

case, features do not have a free will, which makes this property not necessary. The above proof shows that v is in our case neither *sub-additive* nor *super-additive*, which is why we cannot use their corresponding bounds.

Property 11 (Adding features can increase FI)**.** When an extra feature is added to the dataset, the FI of X can increase.

Proof. Consider the previously mentioned XOR dataset, where $X_1, X_2 \sim \mathcal{U}(\{0,1\})$ and $Y = X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$. If at first, X_2 was not in the dataset, the FI of X_1 would be zero, as $\mathrm{Dep}\,(Y|X_1) = 0$. However, if X_2 is added to the dataset, the FI of X_1 increases to $\frac{1}{2}$ (see Property 15). The FI of a feature can thus increase if another feature is added.

Property 12 (Adding features can decrease FI)**.** When an extra feature is added to the dataset, the FI of X can decrease.

Proof. Consider the dataset given by $X \sim \mathcal{U}(\{0,1\})$ and $Y := X$. It immediately follows that $\mathrm{FI}(X) = 1$. However, when a *clone* is introduced ($X^{\mathrm{clone}} := X$), it holds that $\mathrm{FI}(X) = \mathrm{FI}(X^{\mathrm{clone}})$, because of Property 8. Additionally, it follows from Property 1 that $\mathrm{FI}(X) + \mathrm{FI}(X^{\mathrm{clone}}) = 1$. Thus, $\mathrm{FI}(X) = \frac{1}{2}$, and the FI of a variable can therefore be decreased if another variable is added.

Relevance. It is important to observe that the FI of a variable is dependent on the other features (Properties 11 and 12). Adding or removing features could change the FI, which one needs to be aware of.

Property 13 (Cloning does not increase FI)**.** For any RV $X \in \Omega_{\mathrm{f}}$, adding an identical variable $X^{\mathrm{c}} := X$ (cloning) to the dataset, does not increase the FI of X.

Proof. Let $\mathrm{FI}_{\mathrm{c}}(X)$ denote the FI of X after the clone X^{c} is added. To abbreviate the derivation, we again utilize Equation (7.5) to

define $w(S, N_\mathrm{v})$. Using Equation (7.8), we find

$$\begin{aligned}
\mathrm{FI_c}(X) &= \sum_{S \subseteq \Omega_\mathrm{f} \cup \{X^\mathrm{c}\} \setminus \{X\}} w(S, N_\mathrm{v}+1) \cdot D(X, Y, S) \\
&\overset{(a)}{=} \sum_{S \subseteq \Omega_\mathrm{f} \cup \{X^\mathrm{c}\} \setminus \{X\} : X^\mathrm{c} \in S} w(S, N_\mathrm{v}+1) \cdot D(X, Y, S) \\
&\quad + \sum_{S \subseteq \Omega_\mathrm{f} \cup \{X^\mathrm{c}\} \setminus \{X\} : X^\mathrm{c} \notin S} w(S, N_\mathrm{v}+1) \cdot D(X, Y, S) \\
&\overset{(b)}{=} \sum_{S \subseteq \Omega_\mathrm{f} \cup \{X^\mathrm{c}\} \setminus \{X\} : X^\mathrm{c} \in S} w(S, N_\mathrm{v}+1) \cdot 0 \\
&\quad + \sum_{S \subseteq \Omega_\mathrm{f} \cup \{X^\mathrm{c}\} \setminus \{X\} : X^\mathrm{c} \notin S} w(S, N_\mathrm{v}+1) \cdot D(X, Y, S) \\
&= \sum_{S \subseteq \Omega_\mathrm{f} \setminus \{X\}} w(S, N_\mathrm{v}+1) \cdot D(X, Y, S).
\end{aligned}$$

Equality (a) follows by splitting the sum over all subsets of $\Omega_\mathrm{f} \cup \{X^\mathrm{c}\} \setminus \{X\}$ whether X^c is part of the subset or not. Adding X to a subset that already contains the clone X^c does not change the BP dependency function, which is why Equality (b) follows. The takeaway from this derivation is that the sum over all subsets $S \subseteq \Omega_\mathrm{f} \cup \{X^\mathrm{c}\} \setminus \{X\}$ reduces to the sum over $S \subseteq \Omega_\mathrm{f} \setminus \{X\}$.

Comparing the new $\mathrm{FI_c}(X)$ with the original $\mathrm{FI}(X)$ gives

$$\begin{aligned}
\mathrm{FI}(X) - \mathrm{FI_c}(X) &= \sum_{S \subseteq \Omega_\mathrm{f} \setminus \{X\}} w(S, N_\mathrm{v}) \cdot D(X, Y, S) \\
&\quad - \sum_{S \subseteq \Omega_\mathrm{f} \setminus \{X\}} w(S, N_\mathrm{v}+1) \cdot D(X, Y, S).
\end{aligned}$$

Using Equation (7.5), we find that

$$\frac{w(S, N_\mathrm{v}+1)}{w(S, N_\mathrm{v})} = \frac{\frac{|S|! \cdot (N_\mathrm{v}+1-|S|-1)!}{(N_\mathrm{v}+1)!}}{\frac{|S|! \cdot (N_\mathrm{v}-|S|-1)!}{N_\mathrm{v}!}} = \frac{N_\mathrm{v} - |S|}{N_\mathrm{v}+1} < 1,$$

thus $\mathrm{FI}(X) - \mathrm{FI_c}(X) \geq 0$ with equality if and only if $\mathrm{FI}(X) = 0$. Therefore, we can conclude that cloning a variable cannot increase the FI of X and will decrease the FI when X is *null-independent*.

Relevance. We consider this a natural property of a good FI method, as no logical reason can be found why adding the exact same information would lead to an increase in FI for the original variable. The information a variable contains only becomes less valuable, as it becomes common knowledge.

Property 14 (Order does not change FI)**.** The order of the features does not affect the individually assigned FI. Consider the datasets $[X_1, X_2, \ldots, X_{N_\text{v}}]$ and $[Z_1, Z_2, \ldots, Z_{N_\text{v}}]$, where $Z_{\pi(i)} = X_i$ for some permutation π. It holds that $\text{FI}(X_i) = \text{FI}(Z_{\pi(i)})$ for any $i \in \{1, \ldots, N_\text{v}\}$.

Proof. Note that the order of features nowhere plays a roll in the definition of BP-FI (Equation (7.8)). The BP dependency function is also independent of the given order, which is why this property trivially holds.

Relevance. This is a very natural property of a good FI. Consider what would happen if the FI is dependent on the order in the dataset. Should all possible orders be evaluated and averaged to receive a final FI? We cannot find any arguments why someone should want FI to be dependent on the order of features.

Datasets

Next, we consider a few datasets, where we derive the theoretical outcome for the BP-FI. These datasets are also used in Section 7.4.3 to test FI methods. It is very hard to evaluate FI methods, as the ground truth is often unknown. However, we believe that the FI outcomes on these datasets are all natural and defendable. However, it remains subjective what one considers to be the 'correct' FI values.

Property 15 (XOR dataset)**.** Consider the following dataset consisting of two binary features $X_1, X_2 \sim \mathcal{U}(\{0, 1\})$ and a target variable $Y = X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$ which is the XOR of X_1 and X_2. It holds that

$$\text{FI}(X_1) = \text{FI}(X_2) = \frac{1}{2}.$$

Proof. Observe that for the features on their own we have $\mathrm{Dep}(Y|X_1) = \mathrm{Dep}(Y|X_2) = 0$ and together we have $\mathrm{Dep}(Y|X_1 \cup X_2) = 1$. With Equation (7.8), it follows that

$$\begin{aligned}
\mathrm{FI}(X_1) &= \sum_{S \subseteq \{1,2\} \setminus X_1} \frac{|S|!(1-|S|)!}{2!} \left(\mathrm{Dep}(Y|S \cup X_1) - \mathrm{Dep}(Y|S)\right) \\
&= \frac{|\{\varnothing\}|!(1-|\{\varnothing\}|)!}{2!} \left(\mathrm{Dep}(Y|\{\varnothing\} \cup X_1) - \mathrm{Dep}(Y|\{\varnothing\})\right) \\
&\quad + \frac{|\{X_2\}|!(1-|\{X_2\}|)!}{2!} \left(\mathrm{Dep}(Y|X_1 \cup X_2) - \mathrm{Dep}(Y|X_2)\right) \\
&= \frac{1}{2}\left(\mathrm{Dep}(Y|X_1) - 0\right) + \frac{1}{2}\left(\mathrm{Dep}(Y|X_1 \cup X_2) - \mathrm{Dep}(Y|X_2)\right) \\
&= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot (1-0) \\
&= \frac{1}{2}.
\end{aligned}$$

Using Property 1, it follows that $\mathrm{FI}(X_2) = 1 - \mathrm{FI}(X_1) = \frac{1}{2}$.

Relevance. This XOR formula is discussed and used to test FI methods in [61]. However, they only test for equality ($\mathrm{FI}(X_1) = \mathrm{FI}(X_2)$), not the specific value. Due to *symmetry*, we would also argue that both X_1 and X_2 should get the same FI, as they fulfill the same role. Together, they fully determine Y, which is why the total FI should be one (see Property 6). Dividing this equally amongst the two variables, gives a logical desirable FI outcome of $\frac{1}{2}$ for each variable.

Property 16 (Probability dataset)**.** Consider the following dataset consisting of $Y = \lfloor X_S/2 \rfloor$ and $X_i = Z_i + (S-1)$ with $Z_i \sim \mathcal{U}(\{0,2\})$ for $i = 1,2$ and $\mathbb{P}(S=1) = p$, $\mathbb{P}(S=2) = 1-p$. It holds that

$$\mathrm{FI}(X_1) = p \text{ and } \mathrm{FI}(X_2) = 1-p.$$

Proof. Observe that by Equation (7.4)

$$\begin{aligned}
\mathrm{UD}\,(X_1, Y) &= \sum_{x_1 \in \{0,1,2,3\}} p_{X_1}(x_1) \cdot \sum_{y \in \{0,1\}} \left| p_{Y|X_1=x_1}(y) - p_Y(y) \right| \\
&= \sum_{x_1 \in \{0,2\}} p_{X_1}(x_1) \cdot \sum_{y \in \{0,1\}} \left| p_{Y|X_1=x_1}(y) - \frac{1}{2} \right| \\
&\quad + \sum_{x_1 \in \{1,3\}} p_{X_1}(x_1) \cdot \sum_{y \in \{0,1\}} \left| p_{Y|X_1=x_1}(y) - \frac{1}{2} \right| \\
&= \sum_{x_1 \in \{0,2\}} \frac{p}{2} \cdot \left(\left| 1 - \frac{1}{2} \right| + \left| 0 - \frac{1}{2} \right| \right) \\
&\quad + \sum_{x_1 \in \{1,3\}} \frac{1-p}{2} \cdot \sum_{y \in \{0,1\}} |p_Y(y) - p_Y(y)| \\
&= p.
\end{aligned}$$

Similarly, it follows that $\mathrm{UD}\,(X_2, Y) = 1 - p$.

$$\begin{aligned}
\mathrm{UD}\,(Y, Y) &= \sum_{y' \in \{0,1\}} p_Y(y') \cdot \sum_{y \in \{0,1\}} \left| p_{Y|Y=y'}(y) - p_Y(y) \right| \\
&= \sum_{y' \in \{0,1\}} \frac{1}{2} \cdot \left(\left| 1 - \frac{1}{2} \right| + \left| 0 - \frac{1}{2} \right| \right) \\
&= 1.
\end{aligned}$$

From Equation (7.3), it follows that $\mathrm{Dep}\,(Y|X_1) = p$ and $\mathrm{Dep}\,(Y|X_2) = 1 - p$. Additionally, note that knowing X_1 and X_2 fully determines Y, thus $\mathrm{Dep}\,(Y|X_1 \cup X_2) = 1$. With Equation (7.8), we now find

$$\begin{aligned}
\mathrm{FI}(X_1) &= \sum_{S \subseteq \{X_1, X_2\} \setminus X_1} \frac{|S|! \cdot (1 - |S|)!}{2!} \cdot D(X_1, Y, S) \\
&= \frac{|\varnothing|! \cdot (1 - |\varnothing|)!}{2!} \cdot D(X_1, Y, \varnothing) \\
&\quad + \frac{|\{X_2\}|! \cdot (1 - |\{X_2\}|)!}{2!} \cdot D(X_1, Y, X_2) \\
&= \frac{1}{2} \cdot (p - 0) + \frac{1}{2} \cdot (1 - (1 - p)) \\
&= \frac{p}{2} + \frac{p}{2} = p.
\end{aligned}$$

Using Property 1, it follows that $\mathrm{FI}(X_2) = 1 - \mathrm{FI}(X_1) = 1 - p$.

Relevance. At first glance, it is not immediately clear why these FI values are natural, which is why we discuss this dataset in more detail. S can be considered a selection parameter that determines if X_1 or X_2 is used for Y with probability p and $1-p$, respectively. X_i is constructed in such a way that it is uniformly drawn from $\{0, 2\}$ or $\{1, 3\}$ depending on S. However, as $Y = \lfloor X_S/2 \rfloor$, it holds that $X_S = 0$ and $X_S = 1$ give the same outcome for Y. The same holds for $X_S = 2$ and $X_S = 3$. Therefore, note that the distribution of Y is independent of the selection parameter S. Knowing X_1 gives the following information. First, S can be derived from the value of X_1. When $X_1 \in \{0, 2\}$ it must hold that $S = 1$, and if $X_1 \in \{1, 3\}$ it follows that $S = 2$. Second, when $S = 1$ it means that Y is fully determined by X_1. If $S = 2$, knowing that $X_1 = 1$ or $X_1 = 3$ does not provide any additional information about Y. With probability p knowing X_1 will fully determine Y, whereas with probability $1 - p$, it will provide no information about the distribution of Y. The outcome $\text{FI}(X_1) = p$, is therefore very natural. The same argumentation applies for X_2, which leads to $\text{FI}(X_2) = 1 - p$.

7.4 Comparing with existing methods

In the previous section, we showed that BP-FI has many desirable properties. Next, we evaluate for a large collection of FI methods if the properties hold for several synthetic datasets. Note that these datasets can only be used as counterexample, not as proof of a property. First, we discuss the in Section 7.4.1 the FI methods that are investigated. Second, we give the datasets (Section 7.4.2) and explain how they are used to test the properties (Section 7.4.3). The results are discussed in Section 7.4.4.

7.4.1 Alternative FI methods

A wide range of FI methods have been suggested for all kinds of situations. It is therefore first necessary to discuss the major categorical differences between them.

Global vs. local An important distinction to make for FI methods is whether they are constructed for *local* or *global* explanations. *Global* FI methods give an importance score for each feature over the entire dataset, whereas *local* FI methods explain which variables were important for a single example [64]. The global and local scores do not have to coincide: 'features that are globally important may not be important in the local context, and vice versa' [128]. This research is focused on global FI methods, but sometimes a local FI approach can be averaged out to obtain a global FI. For example, in [101] a local FI method is introduced called *Tree SHAP*. It is also used globally, by averaging the absolute values of the local FI.

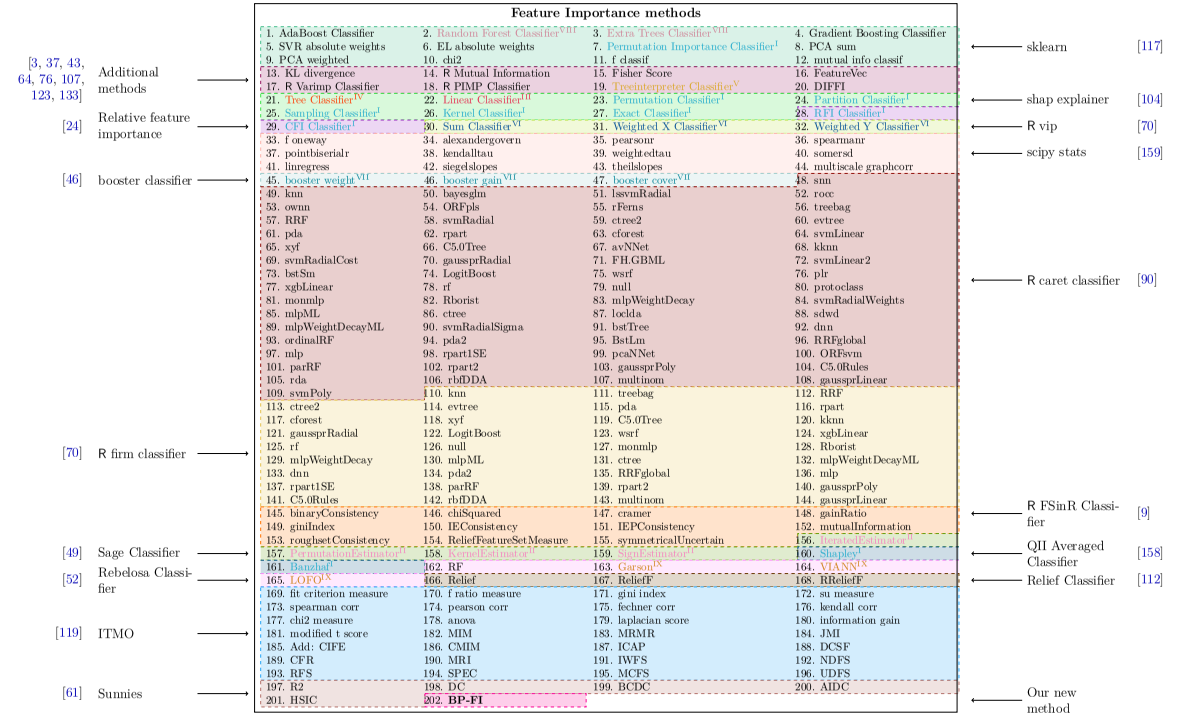
Model-specific vs. model-agnostic A distinction within FI methods can be made between model-*specific* and -*agnostic* methods. *Model-specific* methods aim to find the FI using a prediction model such as a neural network or random forest, whereas *model-agnostic* methods do not use a prediction model. The BP-FI is model-agnostic, which therefore gives insights into the dataset. Whenever a model-specific method is used, the focus lies more on gaining information about the prediction model, not the dataset. In our tests, we use both model-specific and -agnostic methods.

Classification vs. regression Depending on the exact dataset, the target variable is either *categorical* or *numerical*, which is precisely the difference between *classification* and *regression*. Not all existing FI methods can handle both cases. In this chapter, we generate synthetic *classification* datasets, so we only examine FI methods that are intended for these cases. An additional problem with regression datasets, is that continuous variables need to be converted to discrete bins. This conversion could drastically change the FI scores, which makes it harder to draw fair conclusions.

Collection We have gathered the largest known collection of FI methods from various sources [3, 9, 24, 37, 43, 46, 49, 52, 61, 64, 70, 76, 90, 104, 107, 112, 117, 119, 123, 133, 158, 159] or implemented them ourselves. This has been done with the following policy: Whenever code of a *classification* FI method was available in R

or Python or the implementation was relatively straightforward, it was added to the collection. This resulted in 196 *base* methods and 468 total methods, as some base methods can be combined with *multiple* machine learning approaches or selection objectives, see Table 7.1. However, beware that most methods also contain additional parameters, which are not investigated in this chapter. The *default* values for these parameters are always used.

Table 7.1: All evaluated FI methods: List of all FI methods that are evaluated in the experiments. The coloured methods work in combination with multiple options: *Logistic Regression*[I, II, III], *Ridge*[I, II], *Linear Regression*[I, II], *Lasso*[I, II], *SGD Classifier*[I, III], *MLP Classifier*[I, II], *K Neighbors Classifier*[I, II], *Gradient Boosting Classifier*[I, II, IV], *AdaBoost Classifier*[I, II], *Gaussian NB*[I, II], *Bernoulli NB*[I, II], *Linear Discriminant Analysis*[I, II], *Decision Tree Classifier*[I, II, IV, V], *Random Forest Classifier*[I, II, IV, V], *SVC*[I], *CatBoost Classifier*[I, II], *LGBM Classifier*[I, II, IV], *XGB Classifier*[I, II, IV, VII], *XGBRF Classifier*[I, II, IV, VII], *ExtraTree Classifier*[IV, V], *ExtraTrees Classifier*[IV, V], *plsda*[VI], *splsda*[VI], *gini*[VIII], *entropy*[VIII], *NN1*[IX], *NN2*[IX]. This leads to a total of 468 FI methods from various sources [3, 9, 24, 37, 43, 46, 49, 52, 61, 64, 70, 76, 90, 104, 107, 112, 117, 119, 123, 133, 158, 159] or self-implemented.



7.4.2 Synthetic datasets

Next, we briefly discuss the datasets that are used to test the properties described in Section 7.3 for alternative FI methods. In Section 7.A.1, we introduce each dataset and explain how they are generated. To draw fair conclusions, the datasets are not drawn randomly, but *fixed*. To give an example of how we do generate a dataset, we examine dataset 1 *Binary system* (see Section 7.A.1), where the target variable Y is defined as $Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$ with $X_i \sim \mathcal{U}(\{0,1\})$ for all $i \in \{1,2,3\}$. To get interpretable results, we draw each combination of X and Y values the *same number* of times. An example can be seen in Table 7.2. For most datasets, we draw 1,000 samples in total. However datasets 6 and 7 consist of 2,000 samples to ensure null-independence. The datasets have been selected to be computationally inexpensive and to test many properties (see Section 7.4.3) with a limited number of datasets. An overview of the generated datasets can be found in Table 7.3 including the corresponding outcome of BP-FI. Section 7.A.1 provides more technical details about the features and target variables.

Table 7.2: Fixed draw: Example of how the datasets are drawn. Instead of drawing each possible outcome uniformly at random, we draw each combination an equal fixed number of times.

Outcome				# Drawn	
X_1	X_2	X_3	Y	Fixed	Uniform
0	0	0	0	125	133
0	0	1	4	125	129
0	1	0	2	125	121
0	1	1	6	125	109
1	0	0	1	125	136
1	0	1	5	125	124
1	1	0	3	125	115
1	1	1	7	125	133

7.4.3 Property evaluation

In Section 7.4.1, we gathered a collection of existing FI methods. In this section, we evaluate if these FI methods have the same desirable and proven properties of the BP-FI method (see Section 7.3). Due to the sheer number of FI methods (468), it is unfeasible to prove each property for every method. Instead, we divise tests to find

Table 7.3: Overview of datasets: An overview of the generated datasets and the corresponding BP-FI outcome. The details of these datasets can be found in Section 7.A.1. They are used to evaluate if existing FI methods adhere to the same properties as BP-FI (see Section 7.4.3).

	dataset	Variables	BP-FI outcome
Binary system	1. - base	(X_1, X_2, X_3)	$(0.333, 0.333, 0.333)$
	2. - clone	$(X_1^{\text{clone}}, X_1, X_2, X_3)$	$(0.202, 0.202, 0.298, 0.298)$
	3. - clone + 1x fully info.	$(X_1^{\text{clone}}, X_1, X_2, X_3, X_4^{\text{full}})$	$(0.148, 0.148, 0.183, 0.183, 0.338)$
	4. - clone + 2x fully info.	$(X_1^{\text{clone}}, X_1, X_2, X_3, X_4^{\text{full}}, X_5^{\text{full}})$	$(0.117, 0.117, 0.136, 0.136, 0.248, 0.248)$
	5. - clone + 2x fully info. (different order)	$(X_3, X_4^{\text{full}}, X_5^{\text{full}}, X_1^{\text{clone}}, X_1, X_2)$	$(0.136, 0.248, 0.248, 0.117, 0.117, 0.136)$
Null-independent system	6. - base	$(X_1^{\text{null-indep.}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}})$	$(0.000, 0.000, 0.000)$
	7. - constant variable	$(X_1^{\text{null-indep.}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}}, X_4^{\text{const, null-indep.}})$	$(0.000, 0.000, 0.000, 0.000)$
Increasing bins	8. - base	$(X_1^{\text{bins=10}}, X_2^{\text{bins=50}}, X_3^{\text{bins=1,000, full}})$	$(0.297, 0.342, 0.361)$
	9. - more variables	$(X_1^{\text{bins=10}}, X_2^{\text{bins=20}}, X_3^{\text{bins=50}}, X_4^{\text{bins=100}}, X_5^{\text{bins=1,000, full}})$	$(0.179, 0.193, 0.204, 0.208, 0.216)$
	10. - clone (different order)	$(X_3^{\text{bins=1,000, full}}, X_2^{\text{bins=50}}, X_1^{\text{bins=10}}, X_3^{\text{clone, full}})$	$(0.262, 0.253, 0.223, 0.262)$
Dependent system	11. - 1x fully info.	$(X_1^{\text{full}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}})$	$(1.000, 0.000, 0.000)$
	12. - 2x fully info.	$(X_1^{\text{full}}, X_2^{\text{full}}, X_3^{\text{null-indep.}})$	$(0.500, 0.500, 0.000)$
	13. - 3x fully info.	$(X_1^{\text{full}}, X_2^{\text{full}}, X_3^{\text{full}})$	$(0.333, 0.333, 0.333)$
XOR dataset	14. - base	(X_1, X_2)	$(0.500, 0.500)$
	15. - single variable	$(X_1^{\text{null-indep.}})$	(0.000)
	16. - clone	$(X_1^{\text{clone}}, X_1, X_2)$	$(0.167, 0.167, 0.667)$
	17. - null-independent	$(X_1, X_2, X_3^{\text{null-indep.}})$	$(0.500, 0.500, 0.000)$
Probability dataset	18-28. - for $p \in \{0, 0.1, \ldots, 1\}$	(X_1, X_2)	$(p, 1-p)$

counterexamples of these properties using generated datasets (see Section 7.4.2). Due to the number of tests (18), we only discuss the parts that are not straightforward, as most test directly measure the corresponding property. An overview of each test can be found in Section 7.A.2. A summary of the tests can be found in Table 7.4, where it is outlined for each test which property is tested on which datasets.

Computational errors To allow for computational errors, we tolerate a margin of $\epsilon = 0.01$ in each test. If, e.g., a FI value should be zero, a score of 0.01 or -0.01 is still considered a *pass*, whereas a FI value of 0.05 is counted as a *fail*. Usually, this works in the favour of the FI method. However, in Test 9 we evaluate if the FI method assigns zero FI to variables that are not null-independent. In this case, we consider $|\text{FI}(X)| \leq \epsilon$ to be *zero*, as the datasets are constructed in such a way that variables are either null-independent or far from being null-independent.

Running time We limit the running time to one hour per dataset on an i7-12700K processor, whilst four algorithms are running simultaneously. The datasets consist of a small number of features with a very limited outcome space and the number of samples is either 1,000 or 2,000, which is why one hour is a reasonable amount of time.

NaN or infinite values In some cases, a FI method assigns NaN or $\pm\infty$ to a feature. How we handle these values depends on the test. E.g., we consider NaN to fall outside the range $[0, 1]$ (Test 4 and 5), but when we evaluate if the sum of FI values remains stable (Test 2) or if two symmetric features receive the same FI (Test 3), we consider twice NaN or twice $\pm\infty$ to be the same.

Property 9 (Limiting the outcome space) Property 9 states that applying any measurable function f to a RV X cannot increase the FI. In other words, $\text{FI}(X) \geq \text{FI}(f(X))$ holds. This property is tested using dataset 8, 9, and 10 (see Table 7.4). These datasets contain variables that are the outcome of binning the target variable

Table 7.4: Overview of experiments: To evaluate if existing FI methods have the same properties as the BP-FI, we use the tests from Section 7.A.2 on the datasets from Section 7.A.1. ✓means that the test is performed on this dataset. ↕(i) denotes that this dataset is used as baseline or in conjunction with dataset i. The details of the tests and datasets can be found in the appendix.

Test (Section 7.A.2)	Evaluates: Property/Corollary	dataset (Section 7.A.1)																											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1.	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2.	1.1	↕(2-5)	✓	✓	✓	✓	↕(7)	✓	↕(9-10)	✓	✓	↕(12-13)	✓	✓	↕(16-17)		✓	✓												
3.	2	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓		✓	✓						✓					
4.	3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5.	3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6.	4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7.	4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8.	5						✓	✓				✓	✓			✓		✓	✓										✓
9.	5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10.	6											✓																	
11.	8			✓	✓	✓			✓	✓	✓	✓	✓	✓					✓										✓
12.	9								✓	✓	✓																		
13.	11	↕(2)	✓↕(3)	✓↕(4)	✓		↕(7)	✓	↕(9-10)	✓	✓				✓↕(16-17)	↕(14)	✓	✓											
14.	12	↕(2)	✓↕(3)	✓↕(4)	✓		↕(7)	✓	↕(9-10)	✓	✓				✓↕(16-17)	↕(14)	✓	✓											
15.	13	↕(2)	✓						↕(10)		✓						↕(16)		✓										
16.	14				↕(5)	✓													↕(28)	↕(27)	↕(26)	↕(25)	↕(24)		✓	✓	✓	✓	✓
17.	15														✓			✓											
18.	16																		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

using different number of bins. This is how Property 9 is tested, as it should hold that $\text{FI}(X_i) \geq \text{FI}(X_j)$, whenever X_i has more bins than X_j.

Properties 11 and 12 (Adding features can increase/decrease FI) In all other tests, the goal is to find a counterexample of the property. However, Test 13 and 14 are designed to evaluate if a feature gets an increased/decreased FI when a feature is added. This increase/decrease should be more than ϵ. The datasets are chosen in such a way that both an increase and decrease could occur (according to the BP-FI). Only for these tests, we consider the test failed if no counterexample (increase/decrease) is found.

7.4.4 Evaluation results

Best performing methods The top 20 FI methods that pass the most tests are given in Table 7.6. Out of 18 tests, the BP-FI passes all tests, which is as expected as we have proven in Section 7.3 that the BP-FI actually has these properties. Classifiers from *R FSinR Classifier* and *ITMO* fill 11 of the top 20 spots. Out of 11 R FSinR Classifier methods, six are in the top 20, which is quite remarkable. However, observe that the gap between the BP-FI method and the second best method is $18 - 11 = 7$ passed tests. Additionally, 424 out of 468 methods fail more than half of the tests. Figure 7.1 shows how frequently each number of passed tests occurs. A detailed overview of where each top 20 method fails, can be seen in Table 7.5. Note again that in Test 13 and 14 it is considered a fail if adding features never increase or decrease the FI, respectively. It could be that these methods are in fact capable of increasing or decreasing, but for some reason do not with our datasets. Strikingly, most of these methods perform bad on the datasets with a desirable outcome (Test 17 and 18). Adding a variable without additional information (Test 2), also often leads to a change in total FI.

Test 1 In this test, it is evaluated if the sum of FI values is the same as the sum for BP-FI. At first, this seems a rather strict requirement. However, it holds for all datasets that were used that $\text{Dep}\left(Y|\Omega_{\text{f}}\right)$ is either zero or one. Thus, we essentially evaluate if

Table 7.5: Overview of the results: Each FI method is evaluated using the tests outlined in Section 7.A.2, which evaluates if the method adheres to the same properties as the BP-FI (see Section 7.3). This table summarizes out of 468 FI methods how many *pass* or *fail* the test. A distinction is made for the top 20 passing methods. Failing the test means that a counterexample is found. Note that passing the test does not 'prove' that the FI method actually has the property. *No result* indicates that the test could not be executed, because the running time of the FI method was too long or an error occurred.

	Test																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Overall																		
# Passed	1	92	45	438	200	97	132	283	97	31	141	241	243	314	365	172	13	5
# Failed	466	369	421	29	267	370	335	184	370	413	326	98	216	145	58	288	421	459
# No result	1	7	2	1	1	1	1	1	1	24	1	129	9	9	45	8	34	4
Top 20																		
# Passed	1	10	15	20	19	7	18	18	2	13	17	20	4	6	20	17	2	4
# Failed	19	10	5	0	1	13	2	2	18	7	3	0	16	14	0	3	17	16
# No result	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Table 7.6: Top 20: Out of 468 FI methods, these 20 methods pass the 18 tests given in Section 7.A.2 the most often. These tests are designed to examine if a FI method adheres to the same properties as the BP-FI, given in Section 7.3. *Passed* means that the datasets from Section 7.A.1 do not give a counterexample. Certainly, this does not mean that the FI method is proven to actually have this property. *Failed* means that a counterexample was found. *No result* indicates that the test could not be executed, because the running time of the FI method was too long or an error occurred.

Method		Combined result		
		# Passed	# Failed	# No result
202.	BP-FI	18	0	0
147.	cramer	11	7	0
148.	gainRatio	11	7	0
153.	roughsetConsistency	11	7	0
155.	symmetricalUncertain	11	7	0
172.	su measure	11	7	0
88.	sdwd	10	7	1
3.	Extra Trees Classifier	10	8	0
116.	rpart	10	8	0
126.	null	10	8	0
145.	binaryConsistency	10	8	0
152.	mutualInformation	10	8	0
161.	Banzhaf Ridge	10	8	0
197.	R2	10	8	0
162.	RF	10	8	0
166.	Relief	10	8	0
173.	spearman corr	10	8	0
188.	DCSF	10	8	0
189.	CFR	10	8	0
191.	IWFS	10	8	0

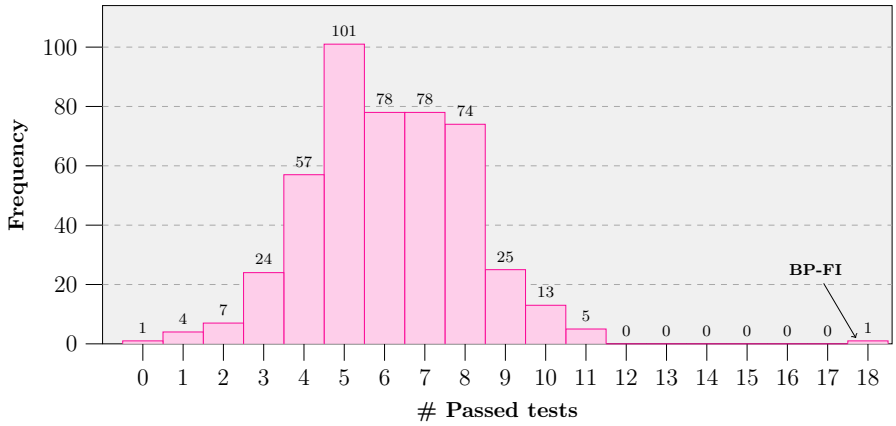


Figure 7.1: Frequency of total passed test: Histogram of the number of passed tests (out of 18) for the 468 FI methods.

the sum of FI is equal to one, when all variables collectively fully determine Y and zero if all variables are null-independent. The tests show that no FI method is able to pass this test, except for the BP-FI. To highlight some of the methods that came close: *162. Rebelosa Classifier RF, 2. Random Forest Classifier entropy, 2. Random Forest Classifier gini* only fail for the datasets where the sum should be zero (because of null-independence) and *1. AdaBoost Classifier* only does not pass on three of the four datasets based on the XOR function (see Section 7.A.1), where the sum should be one, but was zero instead. FI method *51. lssvmRadial* came closest with two fails. For the null-independent datasets (dataset 6 and 7), it gives each feature a FI of 0.5, making the sum larger than zero.

Test 2 In Figure 7.2, a breakdown is given of where the sum of the FI values is unstable. The most errors are made with the *Binary system* datasets, when a fully informative feature is added. In total, 92 methods passed the test, whereas 369 failed. From these 369 methods, 279 fail with at least one increase of the sum, whereas 232 methods fail with at least one decrease. An alarming number of FI methods thus assign significantly more or less FI when a variable is added that does not contain any additional information. More or less credit is given out, whilst the collective knowledge is stable and does not warrant an increase or decrease in credit. Additionally, when

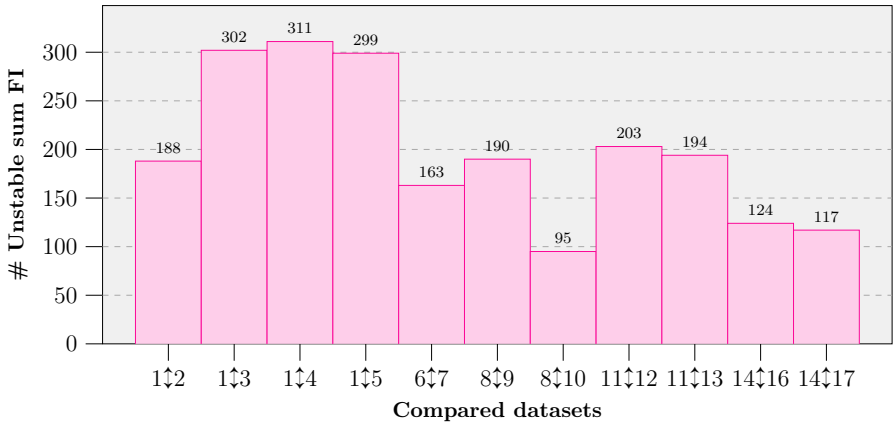


Figure 7.2: Unstable sum FI: Whenever a variable is added that does not give any additional information, the sum of all FI should remain stable. For each comparison, we determine how often this is not the case out of 468 FI methods.

the initial and final sum both contain a NaN value, it is considered as a pass. Three out of 92 would have not passed without this rule. If only the initial or the final sum contained NaN, it is considered a fail, because the sum is not the same. Only five methods fail solely by this rule: *15. Fisher Score*, *11. f classif*, *178. anova*, *179. laplacian score* and *192. NDFS*.

Test 11 Figure 7.3 shows how often each variable is within an ϵ-bound of the largest FI in the dataset. Fully informative variables should attain the largest FI, according to Property 8. In total, we observe that the fully informative variables are often the largest FI with respect to the other variables. However, there still remain many cases where they are not. 326 FI methods fail this test, thus definitively not having Property 8. This makes interpretation difficult, when a variable can get more FI than a variable which fully determines the target variable. What does it mean, when a variable is more important than a variable that gives perfect information?

Test 10, 17, 18 These tests all evaluate if the FI method assigns a specific value to a feature. From Table 7.5, we observe that not

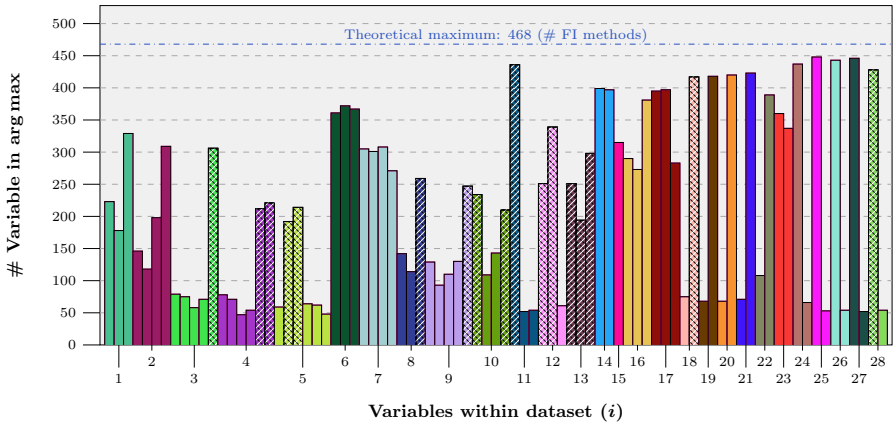


Figure 7.3: Argmax FI: For each variable in every dataset, we determine how often it receives the largest FI (within an ϵ-bound for $\epsilon = 0.01$) with respect to the other variables in the dataset. Fully informative variables should attain the largest FI (see Property 8). All fully informative variables are shaded.

many methods are able to pass these tests. This is not surprising, as they have not been thoroughly tested yet to give a specific value. This is one of the important contributions of this chapter, which is why we want to elaborate on the attempts that have been made in previous research. A lot of synthetic datasets for FI have been proposed [2, 3, 8, 24, 36, 38, 53, 60, 61, 65, 74, 78, 81, 83, 96, 100, 101, 103, 106, 110, 115, 140, 141, 144, 149, 154, 163, 171, 172], but no specific desirable FI values were given. Most commonly, synthetic datasets are generated to evaluate the ability of a FI method to find *noisy* features [8, 36, 65, 74, 78, 81, 96, 140, 144, 149, 163, 171]. The common general concept of such a dataset is that the target variable is *independent* of certain variables. The FI values are commonly evaluated by comparing the FI values of independent variables with dependent variables with the goal to establish if the FI method is able to find independent variables. If the FI method actually predicts the exact desirable FI is not considered. Next, we highlight the papers where some comment about the desired FI is made. Lundberg et al. [101] give two similar datasets, where one variable *increases* in importance. They evaluate multiple FI methods to see if the same behaviour is reflected in the outcome

of these methods. This shows that some commonly used methods could assign lower importance to a variable, when it should actually be increasing. Giles et al. [65] also design multiple artificial datasets to represent different scenarios, where comments are made about which variables should obtain more FI. Sundararajan et al. [149] remark that if every feature value is *unique*, that all variables get *equal* attributions for a FI method (CES) even if the function is not symmetric in the variables. If a tiny amount of noise is added to each feature, all features would get identical attributions. However, no assessment is done on the validity of this outcome. Owen et al. [115] give the following example. Let $f(x_1, x_2) = 10^6 x_1 + x_2$ with $x_1 = 10^6 x_2$, where they argue that, despite the larger variance of x_1, both variables are equally important, as the function can be written as a function of x_1 alone, but also only as a function of x_2. Although we have previously seen that 'written as a function of' is not a good criterion (due to dependencies), we agree with the authors that the FI should be equal. Another example is given by Owen et al. [115], where $\mathbb{P}(x_1 = 0, x_2 = 0, y = y_0) = p_0$, $\mathbb{P}(x_1 = 1, x_2 = 0, y = y_1) = p_1$, and $\mathbb{P}(x_1 = 0, x_2 = 1, y = y_2) = p_2$ are the possible outcomes. If $p_0 = 0$, it is stated in [115] that the Shapley relative importance of x_1 is $\frac{1}{2}$, which is 'what it must be because there is then a bijection between x_1 and x_2'. This is an interesting observation, as most papers do not comment about the validity of an outcome. Additionally, when $y_1 = y_2$ (and $y_0 \neq y_1$), Owen et al. [115] argue that the most important variable, is the one with the largest variance. Fryer et al. [61] also create a binary XOR dataset (see dataset 14). They evaluate seven FI methods for this specific dataset. The role of X_1 and X_2 is symmetric, thus the assigned FI should also be identical. It is shown that six out of seven methods do indeed give a symmetrical result. However, the exact FI value varies greatly. *SHAP* gives FI of 3.19, whereas *Shapley DC* assigns 0.265 as FI. Only symmetry is checked, not the accuracy of the FI method. In conclusion, existing research was not focused on predicting the exact accurate FI values. It is therefore not surprising that FI methods fail these accuracy tests so often. Table 7.7 outlines in more detail how often the variables are assigned a FI value outside an ϵ-bound (with $\epsilon = 0.01$) of the desired outcome. With dataset 11, the FI methods mostly struggle with assigning

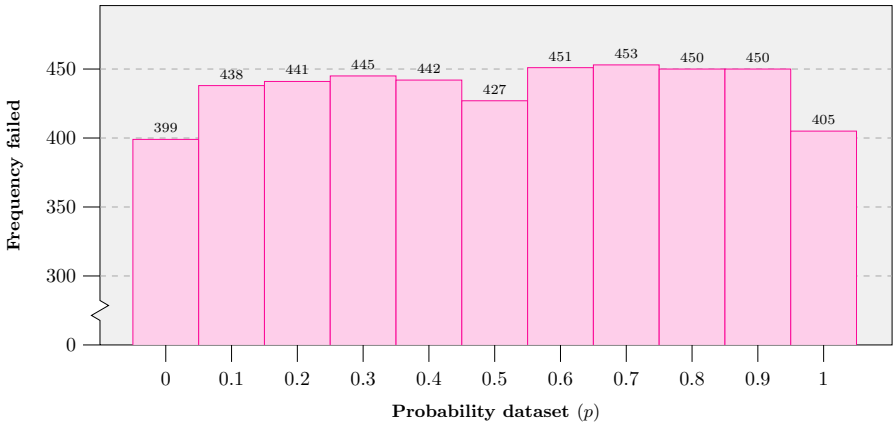


Figure 7.4: Breakdown Test 18 per dataset: In Test 18 a FI method needs to assign the correct FI values for every probability dataset (see Section 7.A.1). In this figure, we breakdown per dataset how often a FI method fails.

1 to the fully informative variable. In total, 413 methods failed Test 10. For dataset 14 and 17, the two XOR variables fail about as often. Comparing these two datasets, it is interesting to note that the XOR variables fail more often, when a null-independent variable is added. In total, 421 methods failed Test 17. Test 18 is hard, as the FI method should assign the correct values for all probability datasets (see Section 7.A.1). Only five methods are able to pass this test: *152. mutualInformation*, *153. roughsetConsistency*, *162. RF*, *175. fechner corr*, and *202. BP-FI*. These five methods also pass Test 10. However, besides BP-FI, there is only one method that also satisfies Test 17, which is *162. RF*. The other three methods all assign only zeros for dataset 14 and 17, not identifying the value that the XOR variables hold, when their information is combined. In Figure 7.4, a breakdown is given for each probability dataset how often FI methods fail. An unexpected result, is that the dataset with probability $p < \frac{1}{2}$ and the dataset with probability $1-p$ do not fail as often. Consistently, $p < \frac{1}{2}$ fails less often than its counterpart $1 - p$, although the datasets are the same up to a reordering of the features and the samples. This effect can also be seen in Table 7.7.

Table 7.7: Specific outcomes: Test 10, 17 and 18 all evaluate if a FI method gives a specific outcome for certain datasets. In this table, it is outlined how often each variable of these datasets is assigned a value outside an ϵ-bound (with $\epsilon = 0.01$) of the desired outcome.

Dataset	Desirable outcome	# Non-desirable outcome					
		not NaN			NaN		
		X_1	X_2	X_3	X_1	X_2	X_3
11	$(1, 0, 0)$	360	89	88	4	4	4
14	$(\frac{1}{2}, \frac{1}{2})$	353	351	-	5	5	-
17	$(\frac{1}{2}, \frac{1}{2}, 0)$	369	364	90	5	5	5
18	$(0, 1)$	82	352	-	4	4	-
19	$(\frac{1}{10}, \frac{9}{10})$	412	434	-	3	3	-
20	$(\frac{2}{10}, \frac{8}{10})$	434	438	-	3	3	-
21	$(\frac{3}{10}, \frac{7}{10})$	435	441	-	3	3	-
22	$(\frac{4}{10}, \frac{6}{10})$	439	436	-	3	3	-
23	$(\frac{5}{10}, \frac{5}{10})$	423	422	-	3	3	-
24	$(\frac{6}{10}, \frac{4}{10})$	448	447	-	3	3	-
25	$(\frac{7}{10}, \frac{3}{10})$	449	446	-	3	3	-
26	$(\frac{8}{10}, \frac{2}{10})$	446	444	-	3	3	-
27	$(\frac{9}{10}, \frac{1}{10})$	444	435	-	3	3	-
28	$(1, 0)$	352	86	-	5	5	-

No result Focussing on the *no result* row of Table 7.5, there is one base method named *158. KernelEstimator* in combination with *Lasso* that in all cases did not work or exceeded running time. The large number of no results in Test 12 stem mostly from slow running times on the three datasets that are used in the test. At least 63 methods were too slow for each dataset, which automatically means that the test cannot be executed.

7.5 Discussion and future research

Whilst it is recommended to use our new FI method, it is important to understand the limitations and potential pitfalls. Below we elaborate on both the shortcomings of the approach proposed, and the related challenges for further research. We start by discussing by some matters that one needs to be aware of when applying the BP-FI (Section 7.5.1). Next, we discuss some choices that were made for the experiments in Section 7.5.2. Finally, we elaborate on

other possible research avenues in Section 7.5.3.

7.5.1 Creating awareness

Binning Berkelmans et al. [16] explained that the way in which continuous data is discretized can have a considerable effect on the BP dependency function, which is why all datasets that were used in our research are *discrete*. If a feature has too many unique values (due to poor binning), it will receive a higher FI from BP-FI, as more information can be stored in the unique values (see Property 9). On the other hand, when too few bins are chosen, an important feature can receive low FI, as the information is lost due to the binning. Future research should investigate and test which binning algorithms give the closest results to the underlying FI.

Too few samples Consider the following dataset: for random variables $X_i, Y \sim \mathcal{U}(\{0, 1, \ldots, 9\})$ i.i.d. for $i \in \{1, \ldots, 5\}$. Note that all features are null-independent, as Y is just uniformly drawn without considering the features in any way. If $n_s = \infty$, the desired outcome would therefore be $(0, 0, 0, 0, 0)$. However, when *not enough* samples are given in the dataset, the features will get nonzero FI. Considering that the total number of different feature values is 10^5, combining all features *does* actually give information about Y, when $n_s \ll 10^5$. For any possible combination of features, it is unlikely that it occurs more than once in the dataset. Therefore, knowing all feature values would (almost surely) determine the value of Y. Property 1 gives that the sum of all FI should therefore be one. All feature variables are also *symmetric* (Property 2), which is why the desired outcome is $(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$ instead. This example shows that one should be aware of the influence of the number of samples on the resulting FI. Variables that do *not* influence Y can still contain information, when not enough samples are provided. In this way, insufficient samples could lead to wrong conclusions, if one is not wary of this phenomenon.

Counterintuitive dependency case The *Berkelmans-Pries* dependency of Y on X measures how much probability mass of Y is shifted by knowing X. However, two similar shifts in probability

mass could lead to different predictive power. To explain this, we examine the following dataset. $X_1, X_2 \sim \mathcal{U}(\{0,1\})$ with

$$\mathbb{P}(Y = y | X_1 = x_1, X_2 = x_2) = \begin{cases} 1/4 & \text{if } (x_2, y) = (0,0), \\ 3/4 & \text{if } (x_2, y) = (0,1), \\ 5/8 & \text{if } (x_1, x_2, y) = (0,1,0), \\ 3/8 & \text{if } (x_1, x_2, y) = (0,1,1), \\ 7/8 & \text{if } (x_1, x_2, y) = (1,1,0), \\ 1/8 & \text{if } (x_1, x_2, y) = (1,1,1). \end{cases}$$

Knowing the value of X_2 shifts the distribution of Y. Before, Y was split 50/50, but when the value of X_2 is known, the labels are either split 25/75 or 75/25, depending on the value of X_2. Knowing X_1 gives even more information, as e.g., knowing $X_1 = X_2 = 1$ makes it more likely that $Y = 0$. However, the shift in distribution of Y is the same for knowing only X_2 and X_1 combined with X_2, which results in $\text{Dep}(Y|X_2) = \text{Dep}(Y|X_1 \cup X_2)$. This is a counter-intuitive result. Globally, knowing X_2 or $X_1 \cup X_2$ gives the same shift in distribution, but locally we can predict Y much better if we know X_1 as well. We are unsure how this effects the BP-FI. In this case, it follows that $\text{FI}(X_1 \cup X_2) > \text{FI}(X_2)$, which is desirable. It is not unthinkable that a solution can be found to modify the dependency function in order to get a more intuitive result for such a case. Think, e.g., of a different distance metric, that incorporates the local accuracy given the feature values or a conditional variant, which not only tests for independence, but also for conditional independence. These are all critical research paths that should be investigated.

Using FI for feature selection *Feature selection* (FS) is 'the problem of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept' [87]. Basically, the objective is to find a subset of all features that gives the best performance for a given model, as larger feature sets could decrease the accuracy of a model [91]. Many FI methods actually stem from a FS procedure. However, it is important to stress that *high* FI means that it should automatically be selected as feature. Shared knowledge with other features could render the feature less useful than expected. The other way around, *low* FI features should not

automatically be discarded. In combination with other features, it could still give some additional insights that other features are not able to provide. Calculation of BP-FI values could also provide insight into which group of K features Y is most dependent on. To derive the result of BP-FI, all dependencies of Y on a subset $S \subseteq \Omega_f$ are determined. If only K variables are selected, it is natural to choose

$$S_K^* \in \underset{S \subseteq \Omega_f : |S| = K}{\arg\max} \{\mathrm{Dep}\,(Y|S)\}.$$

These values are stored as an intermediate step in BP-FI, thus S_K^* can be derived quickly thereafter.

Larger outcome space leads to higher FI We have proven that a larger outcome space can never lead to a decrease in FI for BP-FI. This means, that features with more possible outcomes are more likely to attain a higher FI, depending on the distribution. There is a difference between a feature that has many possible outcomes that are *almost never* attained, and a feature where many possible outcomes are *regularly* observed. We do not find this property undesirable, as some articles suggest [146, 171], as we would argue that a feature *can* contain more information by storing the information in additional outcomes, which would lead to an non-decreasing FI.

7.5.2 Experimental design choices

Regression To avoid binning issues, we only considered classification models and datasets. There are many more regression FI methods, that should be considered in a similar fashion. However, to draw clear and accurate conclusions, it is first necessary to understand how binning affects the results. Sometimes counter-intuitive results can occur due to binning, that are not necessarily wrong. In such a case, it is crucial that the FI method is not depreciated.

Run-time In the experiments, it could happen that a FI method had *no result*, due to an excessive run-time or incompatible FI scores. The maximum run-time for each algorithm was set to one hour

per dataset on an i7-12700K processor with 4 algorithms running simultaneously. The maximum run-time was necessary due to the sheer number of FI methods and datasets. Running four algorithms in parallel could unfairly penalise the run-time, as the processor is sometimes limited by other algorithms. In some occurrences, other parallel processes were already finished, which could potentially lower the run-time of an algorithm. There is a potential risk here, that accurate (but slow) FI methods are not showing up in the results. However, our synthetic datasets are relatively small with respect to the number of samples and the number of features, and we argue that one hour should be reasonable. Depending on the use case, sometimes a long time can be used to determine a FI value, whereas in other cases it could be essential to determine it rather quickly. Especially for larger datasets, it could even be unfeasible to run some FI methods. BP-FI uses Shapley values, which are exponentially harder to compute when the number of features grow. Approximation algorithms should be developed to faster estimate the true BP-FI outcome. Quick approximations could be useful if the run-time is much faster and the approximation is decent enough. Already, multiple papers have suggested approaches to approximate Shapley values faster [2, 40, 81, 97, 147]. These approaches save time, but at what cost? A study could be done to find the best FI method given a dataset and an allowed running time.

Stochasticity methods One factor we did not incorporate, is the *stochasticity* of some FI methods. Some methods do not predict the same FI values, when it is repeatedly used. As example, *79. rf* predicted for dataset 3 (12.1, 11.7, 17.9, 15.2, 37.7) rounded to the first decimal. Running the method again gives a different result: (11.4, 12.0, 17.4, 15.6, 37.1), as this method uses a stochastic *random forest*. In principle, it is *undesirable* that a FI method is stochastic, as we believe that there should be a unique assignment of FI given a dataset. Due to the number of FI methods and datasets, we did not repeat and averaged each FI method. This would however give a better view on the performance of stochastic FI methods.

Parameter tuning All FI methods were used with *default* parameter values. Different parameter values could lead to more or less

failed tests. However, the *ideal* parameter setting is not known beforehand, making it necessary to search a wide range of parameters. This was not the focus of our research, but future research could try to understand and learn which parameter values should be chosen for a given dataset.

Ranking FI methods In Table 7.6, the 20 FI methods that passed the most tests were highlighted. However, it is important to stress that not every test is equally difficult. Depending on the user, some properties could be more or less relevant. It is, e.g., much harder to accurately predict the specific values for 11 datasets (Test 18), than to always predict non-negatively (Test 4). Every test is weighed equally, but this does not necessarily represent the difficulty of passing the tests accurately. However, we note that *175. fechner corr* is the only FI method that passed Test 18, that ended up outside the top 20. We stress that we focused on finding out if FI methods adhere to the properties, not necessarily finding the best and most fair ranking.

7.5.3 Additional matters

Global vs. local BP-FI is designed to determine the FI *globally*. However, another important research area focuses on *local* explanations. These explanations should provide information about why a specific sample has a certain target value instead of a different value. They provide the necessary interpretability that is increasingly demanded for practical applications. This could give insights for questions like: 'If my income would be higher, could I get a bigger loan?', 'Does race play a role in this prediction?', and 'For this automated machine learning decision, what were the critical factors?'. Many local FI methods have been proposed, and some even use Shapley values. A structured review should be made about all proposed local methods, similar to our approach for global FI methods to find which local FI methods actually produce accurate explanations.

BP-FI can be modified to provide local explanations. For example, we can make the characteristic function localized in the following way.

Let $Y_{S,z}$ be Y restricted to the event that $X_i = z_i$ for $i \notin S$, let us similarly define $X_{S,z}$. Then, we can define a localized characteristic function by:

$$v_z(S) := \text{Dep}\left(Y_{S,z}|X_{S,z}\right). \tag{7.10}$$

When dealing with continuous data, assuming equality could be too strict. In this case, a precision vector parameter ϵ can be used, where we define $Y_{S,z,\epsilon}$ to be Y restricted to the event that $|X_i - z_i| \leq \epsilon_i$ for $i \notin S$, and in the same way we define $X_{S,z,\epsilon}$. We then get the following localized characteristic function:

$$v_{z,\epsilon}(S) := \text{Dep}\left(Y_{S,z,\epsilon}|X_{S,z,\epsilon}\right). \tag{7.11}$$

Additionally, there are at least two possible ways how BP-FI can be adapted to be used for local explanations if some distance function $d(i,j)$ and parameter δ are available to determine if sample j is close enough to i to be considered 'local'. We can (I) discard all samples where $d(i,j) > \delta$ and/or (II) generate samples, such that $d(i,j) \leq \delta$ for all generated samples. Then, we can use BP-FI on the remaining samples and/or the generated samples, which would give local FI. Note that there should still be enough samples, as we have previously discussed that too few samples could lead to different FI outcomes. However, there are many more ways how BP-FI can be modified to be used for local explanations.

Model-specific FI BP-FI is in principle model-agnostic, as the FI is determined of the dataset, not the FI for a prediction model. However, BP-FI can still provide insights for any specific model. By replacing the target variable with the predicted outcomes of the model, we can apply BP-FI to this new dataset, which gives insight into which features are useful in the prediction model. Additionally, one can compare these FI results with the original FI (before replacing the target variable with the predicted outcomes) to see in what way the model changed the FI.

Additional properties In this chapter, we have proven properties of BP-FI. However, an in-depth study could lead to finding more useful properties. This holds both for BP-FI as well as the dependency function it is based on. Applying isomorphisms, e.g.,

does not change the dependency function. Therefore, the BP-FI is also stable under isomorphisms. Understanding what properties BP-FI has is a double-edged sword. Finding useful properties shows the power of BP-FI and finding undesirable behavior could lead to a future improvement.

Additional datasets Ground truths are often unknown for FI. In this chapter, we have given two kinds of datasets where the desirable outcomes are natural. It would however, be useful to create a *larger* collection of datasets both for global and local FI with an exact ground truth. We recognize that this could be a tall order, but we believe that it is essential to further improve FI methods.

Human labelling In some articles [104, 128], humans are used to evaluate explanations. An intriguing question to investigate is if humans are good at predicting FI. The BP-FI can be used as baseline to validate the values that are given by the participants. Are humans able to identify the correct order of FI? Even more difficult, can they predict close to the actual FI values?

7.5.4 Summary

We started by introducing a novel FI method named *Berkelmans-Pries* FI, which combines *Shapley* values and the *Berkelmans-Pries* dependency function [16]. In Section 7.3, we proved many useful properties of BP-FI. We discussed which FI methods already exist and introduced datasets to evaluate if these methods adhere to the same properties. In Section 7.4.3, we explain how the properties are tested. The results show that BP-FI is able to pass *many more* tests than any other FI method from a large collection of FI methods (468), which is a significant step forwards. Most methods have not previously been tested to give exact results due to missing ground truths. In this chapter, we provide several specific datasets, where the desired FI can be derived. From the tests, it follows that previous methods are not able to accurately predict the desired FI values. In Section 7.5, we extensively discussed the shortcomings of this chapter, and the challenges for further research. There are many challenging research opportunities that should be explored to

further improve interpretability and explainability of datasets and machine learning models.

7.A Appendix

7.A.1 Datasets

In this appendix, we discuss how the datasets are generated that are used in the experiments. We use *fixed draw* instead of *uniformly random* to draw each dataset *exactly* according to its distribution. This is done to remove stochasticity from the dataset in order to get precise and interpretable results. An example of the difference between fixed draw and uniformly random can be seen in Table 7.2. The datasets consist of 1,000 samples, except for datasets 6 and 7 which contains 2,000 samples to ensure null-independence. The datasets are designed to be computationally inexpensive, whilst still being able to test many properties (see Section 7.4.3). Below, we outline the formulas that are used to generate the datasets and give the corresponding FI values of our novel method BP-FI.

Dataset 1: ***Binary system***
Feature variable(s): $X_i \sim \mathcal{U}(\{0,1\})$ i.i.d. for $i \in \{1,2,3\}$
Target variable: $Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$.
Order: (X_1, X_2, X_3).
BP-FI: $(0.333, 0.333, 0.333)$.

Dataset 2: ***Binary system with clone***
Feature variable(s): $X_i \sim \mathcal{U}(\{0,1\})$ i.i.d. for $i \in \{1,2,3\}$ and $X_1^{\text{clone}} := X_1$.
Target variable: $Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$.
Order: $(X_1^{\text{clone}}, X_1, X_2, X_3)$.
BP-FI: $(0.202, 0.202, 0.298, 0.298)$.

Dataset 3: ***Binary system with clone and one fully informative variable***
Feature variable(s): $X_i \sim \mathcal{U}(\{0,1\})$ i.i.d. for $i \in \{1,2,3\}$ and $X_1^{\text{clone}} := X_1$ and $X_4^{\text{full}} := Y^2$.
Target variable: $Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$.
Order: $(X_1^{\text{clone}}, X_1, X_2, X_3, X_4^{\text{full}})$.
BP-FI: $(0.148, 0.148, 0.183, 0.183, 0.338)$.

Dataset 4: ***Binary system with clone and two fully informative variables***
Feature variable(s): $X_i \sim \mathcal{U}(\{0,1\})$ i.i.d. for $i \in \{1,2,3\}$ and $X_1^{\text{clone}} := X_1$ and $X_4^{\text{full}} := Y^2$, $X_5^{\text{full}} := Y^3$.
Target variable: $Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$.
Order: $(X_1^{\text{clone}}, X_1, X_2, X_3, X_4^{\text{full}}, X_5^{\text{full}})$.
BP-FI: $(0.117, 0.117, 0.136, 0.136, 0.248, 0.248)$.

Dataset 5: ***Binary system with clone and two fully informative variables different order***
Feature variable(s): $X_i \sim \mathcal{U}(\{0,1\})$ i.i.d. for $i \in \{1,2,3\}$ and $X_1^{\text{clone}} := X_1$ and $X_4^{\text{full}} := Y^2$, $X_5^{\text{full}} := Y^3$.
Target variable: $Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$.
Order: $(X_3, X_4^{\text{full}}, X_5^{\text{full}}, X_1^{\text{clone}}, X_1, X_2)$.
BP-FI: $(0.136, 0.248, 0.248, 0.117, 0.117, 0.136)$.

Dataset 6: ***Null-independent system***
Feature variable(s): $X_i^{\text{null-indep.}} \sim \mathcal{U}(\{0,1\})$ i.i.d. for $i \in \{1,2,3\}$.
Target variable: $Y \sim \mathcal{U}(\{0,1\})$.
Order: $(X_1^{\text{null-indep.}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}})$.
BP-FI: $(0.000, 0.000, 0.000)$.

Dataset 7: ***Null-independent system with constant variable***
Feature variable(s): $X_i^{\text{null-indep.}} \sim \mathcal{U}(\{0,1\})$ i.i.d. for $i \in \{1,2,3\}$ and $X_4^{\text{const, null-indep.}} := 1$.
Target variable: $Y \sim \mathcal{U}(\{0,1\})$.
Order: $(X_1^{\text{null-indep.}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}}, X_4^{\text{const, null-indep.}})$.
BP-FI: $(0.000, 0.000, 0.000, 0.000)$.

Dataset 8: ***Uniform system increasing bins***
Feature variable(s): Let $\mathcal{L}_i := \{0, 1/(i-1), \ldots, 1\}$ be an equally spaced set. Define:

$$\begin{aligned} X_1^{\text{bins=10}} &:= \underset{x_1 \in \mathcal{L}_{10}}{\arg\max}\{Y \geq x_1\}, \\ X_2^{\text{bins=50}} &:= \underset{x_2 \in \mathcal{L}_{50}}{\arg\max}\{Y \geq x_2\}, \\ X_3^{\text{bins=1,000, full}} &:= \underset{x_3 \in \mathcal{L}_{1,000}}{\arg\max}\{Y \geq x_3\}. \end{aligned}$$

Target variable: $Y \sim \mathcal{U}(\mathcal{L}_{1,000})$.
Order: $(X_1^{\text{bins=10}}, X_2^{\text{bins=50}}, X_3^{\text{bins=1,000, full}})$.
BP-FI: $(0.297, 0.342, 0.361)$.

Dataset 9: ***Uniform system increasing bins more variables***
Feature variable(s): Let $\mathcal{L}_i := \{0, 1/(i-1), \ldots, 1\}$ be an equally spaced set. Define:

$$\begin{aligned} X_1^{\text{bins=10}} &:= \underset{x_1 \in \mathcal{L}_{10}}{\arg\max}\{Y \geq x_1\}, \\ X_2^{\text{bins=20}} &:= \underset{x_2 \in \mathcal{L}_{20}}{\arg\max}\{Y \geq x_2\}, \\ X_3^{\text{bins=50}} &:= \underset{x_3 \in \mathcal{L}_{50}}{\arg\max}\{Y \geq x_3\}, \\ X_4^{\text{bins=100}} &:= \underset{x_4 \in \mathcal{L}_{100}}{\arg\max}\{Y \geq x_4\}, \\ X_5^{\text{bins=1,000, full}} &:= \underset{x_5 \in \mathcal{L}_{1,000}}{\arg\max}\{Y \geq x_5\}. \end{aligned}$$

Target variable: $Y \sim \mathcal{U}(\mathcal{L}_{1,000})$.
Order: $(X_1^{\text{bins=10}}, X_2^{\text{bins=20}}, X_3^{\text{bins=50}}, X_4^{\text{bins=100}}, X_5^{\text{bins=1,000, full}})$.
BP-FI: $(0.179, 0.193, 0.204, 0.208, 0.216)$.

Dataset 10: ***Uniform system increasing bins with clone different order***
Feature variable(s): Let $\mathcal{L}_i := \{0, 1/(i-1), \ldots, 1\}$ be an equally spaced set. Define:

$$\begin{aligned} X_1^{\text{bins=10}} &:= \underset{x_1 \in \mathcal{L}_{10}}{\arg\max}\{Y \geq x_1\}, \\ X_2^{\text{bins=50}} &:= \underset{x_2 \in \mathcal{L}_{50}}{\arg\max}\{Y \geq x_2\}, \\ X_3^{\text{bins=1,000, full}} &:= \underset{x_3 \in \mathcal{L}_{1,000}}{\arg\max}\{Y \geq x_3\}, \\ X_3^{\text{clone, full}} &:= X_3^{\text{bins=1,000, full}}. \end{aligned}$$

Target variable: $Y \sim \mathcal{U}(\mathcal{L}_{1,000})$.
Order: $(X_3^{\text{bins=1,000, full}}, X_2^{\text{bins=50}}, X_1^{\text{bins=10}}, X_3^{\text{clone, full}})$.
BP-FI: $(0.262, 0.253, 0.223, 0.262)$.

Dataset 11: ***Dependent system: 1x fully informative variable***
Feature variable(s): $X_1^{\text{full}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}} \sim \mathcal{U}(\{1, 2\})$.
Target variable: $Y := X_1^{\text{full}}$.
Order: $(X_1^{\text{full}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}})$.
BP-FI: $(1.000, 0.000, 0.000)$.

Dataset 12: ***Dependent system: 2x fully informative variable***
Feature variable(s): $X_1^{\text{full}}, X_3^{\text{null-indep.}} \sim \mathcal{U}(\{1, 2\})$ and $X_2^{\text{full}} := Y^2$.
Target variable: $Y := X_1^{\text{full}}$.
Order: $(X_1^{\text{full}}, X_2^{\text{full}}, X_3^{\text{null-indep.}})$.
BP-FI: $(0.500, 0.500, 0.000)$.

Dataset 13: ***Dependent system: 3x fully informative variable***
Feature variable(s): $X_1^{\text{full}} \sim \mathcal{U}(\{1, 2\})$ and $X_2^{\text{full}} := Y^2$, $X_3^{\text{full}} := Y^3$.
Target variable: $Y := X_1^{\text{full}}$.
Order: $(X_1^{\text{full}}, X_2^{\text{full}}, X_3^{\text{full}})$.
BP-FI: $(0.333, 0.333, 0.333)$.

Dataset 14: ***XOR dataset***
Feature variable(s): $X_1, X_2 \sim \mathcal{U}(\{1, 2\})$.
Target variable: $Y := X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$.
Order: (X_1, X_2).
BP-FI: $(0.500, 0.500)$.

Dataset 15: ***XOR dataset one variable***
Feature variable(s): $X_1^{\text{null-indep.}} \sim \mathcal{U}(\{1, 2\})$.
Target variable: $Y := X_1^{\text{null-indep.}} \cdot (1 - X_2) + X_2 \cdot (1 - X_1^{\text{null-indep.}})$ with $X_2 \sim \mathcal{U}(\{1, 2\})$.
Order: $(X_1^{\text{null-indep.}})$.
BP-FI: (0.000).

Dataset 16: ***XOR dataset with clone***
Feature variable(s): $X_1, X_2 \sim \mathcal{U}(\{1, 2\})$ and $X_1^{\text{clone}} := X_1$.
Target variable: $Y := X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$.
Order: $(X_1^{\text{clone}}, X_1, X_2)$.
BP-FI: $(0.167, 0.167, 0.667)$.

Dataset 17: ***XOR dataset with null independent***
Feature variable(s): $X_1, X_2 \sim \mathcal{U}(\{1,2\})$ and $X_3^{\text{null-indep.}} \sim \mathcal{U}(\{0,3\})$.
Target variable: $Y := X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$.
Order: $(X_1, X_2, X_3^{\text{null-indep.}})$.
BP-FI: $(0.500, 0.500, 0.000)$.

Dataset 18-28: ***Probability datasets***
Feature variable(s): $X_i = Z_i + S$ with $Z_i \sim \mathcal{U}(\{0,2\})$ i.i.d. for $i = 1, 2$ and $\mathbb{P}(S = 1) = p$, $\mathbb{P}(S = 2) = 1 - p$.
Target variable: $Y = \lfloor X_S/2 \rfloor$.
Order: (X_1, X_2).
BP-FI: $(p, 1 - p)$.

7.A.2 Tests

This appendix gives an overview of the tests that are used for each FI method to evaluate if they adhere to the properties given in Section 7.3. Most tests are straightforward, but additional explanations are given in Section 7.4.3.

Test 1: ***Efficiency sum BP-FI***
Evaluates: Property 1.
Explanation: We evaluate if the sum of all FI is equal to the sum of the *Berkelmans-Pries* dependency function of Y on all features. When a FI value of NaN or infinite is assigned, the sum is automatically not equal to the sum for BP-FI.

Test 2: ***Efficiency stable***
Evaluates: Corollary 1.1.
Explanation: Whenever a variable is added to a dataset, we examine if the sum of all FI changes. If a variable does not give any additional information compared to the other variables, the sum of all FI should stay the same.

Test 3: ***Symmetry***
Evaluates: Property 2.
Explanation: In some datasets, there are *symmetrical* variables (see Property 2). We determine for all symmetrical variables if they receive identical FI.

Test 4: ***Range (lower)***
Evaluates: Property 3.
Explanation: We examine for all FI outcomes if they are greater or equal to zero.
Test 5: ***Range (upper)***

Evaluates: Property 3.
Explanation: We examine for all FI outcomes if they are smaller or equal to one.

Test 6: ***Bounds BP-FI (lower)***
Evaluates: Property 4.
Explanation: We evaluate if the bounds given in Property 4 also hold for other FI methods. Every $\text{FI}(X)$ with $X \in \Omega_{\text{f}}$ can be lower bounded for BP-FI by $\frac{\text{Dep}(Y|X)}{N_{\text{v}}} \leq \text{FI}(X)$.

Test 7: ***Bounds BP-FI (upper)***
Evaluates: Property 4.
Explanation: We evaluate if the bounds given in Property 4 also hold for other FI methods. Every $\text{FI}(X)$ with $X \in \Omega_{\text{f}}$ can be upper bounded for BP-FI by $\text{FI}(X) \leq \text{Dep}(Y|\Omega_{\text{f}})$.

Test 8: ***Null-independent implies zero FI***
Evaluates: Property 5.
Explanation: In some datasets, there are *null-independent* variables. In these cases, we investigate if they also receive zero FI.

Test 9: ***Zero FI implies null-independent***
Evaluates: Property 5.
Explanation: When a variable gets zero FI, it should hold that such a feature is null-independent.

Test 10: ***One fully informative, two null-independent***
Evaluates: Property 6.
Explanation: dataset 11 (see Section 7.A.1) consists of a fully dependent target variable $Y := X_1^{\text{full}}$ and two null-independent variables $X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}}$. We test if $\text{FI}(X_1^{\text{full}}) = 1$ and $\text{FI}(X_2^{\text{null-indep.}}) = \text{FI}(X_3^{\text{null-indep.}}) = 0$.

Test 11: ***Fully informative variable in argmax FI***
Evaluates: Property 8.
Explanation: Whenever a fully informative feature exists in a dataset, there should not be a feature that attains a higher FI.

Test 12: ***Limiting the outcome space***
Evaluates: Property 9.
Explanation: To evaluate if applying a measurable function f to a RV X could increase the FI, we examine the datasets where the same RV is binned using different bins. The binning can be viewed as applying a function f. Whenever less bins are used, the FI should not increase.

Test 13: ***Adding features can increase FI***
Evaluates: Property 11.
Explanation: Whenever a feature is added to a dataset, we examine if this ever increases the FI of an original variable. If the FI never increases, we consider this a fail.

Test 14: ***Adding features can decrease FI***
Evaluates: Property 12.
Explanation: Whenever a feature is added to a dataset, we examine if this ever decreases the FI of an original variable. If the FI never decreases, we consider this a fail.

Test 15: ***Cloning does not increase FI***
Evaluates: Property 13.
Explanation: We evaluate if adding a clone to a dataset increase the FI of the original variable.

Test 16: ***Order does not change FI***
Evaluates: Property 14.
Explanation: We check if the order of the variables changes the assigned FI.

Test 17: ***Outcome XOR***
Evaluates: Property 15.
Explanation: This test evaluates the specific outcome of two datasets. For dataset 14 the desired outcome is $(1/2, 1/2)$ and $(1/2, 1/2, 0)$ for dataset 17. A FI method fails this test when one of the FI values falls outside the ϵ-bound of the desired outcome.

Test 18: ***Outcome probability datasets***
Evaluates: Property 16.
Explanation: This test evaluates the specific outcomes of all probability datasets (datasets 18-28). The desired outcome for probability p is $(p, 1-p)$. A FI method fails this test when one of the FI values falls outside the ϵ-bound of the desired outcome.

Chapter 7 The Berkelmans-Pries Feature Importance method: a generic measure of informativeness of features

8

Extending the Berkelmans-Pries dependency function to a conditional setting with early performance testing on medication data

8.1 Introduction

In Chapter 6 we extensively discussed the notion of dependency. However, we have not yet addressed the notion of 'conditional dependency', i.e. for random variables, X_A, X_B, and X_C (taking values in E_A, E_B, E_C respectively), if X_C is known, what is the dependency relation between X_A and X_B?

One might naively think that if X_A and X_B are conditionally dependent given X_C that $\mathrm{Dep}\,(X_A|(X_B, X_C)) > \mathrm{Dep}\,(X_A|X_C)$. However, regrettably this is not necessarily the case.

For example, consider the following, with X_A, X_B, X_C all binary with distribution as in Table 8.1, then $\mathrm{UD}\,(X_A, (X_B, X_C)) = \frac{1}{2} = \mathrm{UD}\,(X_A, X_C)$. So though our dependency measure measures dependency, no conclusions can be drawn regarding conditional dependency. So we need another measure if we want to quantify conditional de-

Table 8.1: Distribution for example where there is conditional dependence, but the BP dependency function does not change.

$\mathbb{P}(X_A = a, X_B = b, X_c = c)$	$a = 0$	$a = 1$
$b = 0, c = 0$	1/32	7/32
$b = 0, c = 1$	5/32	3/32
$b = 1, c = 0$	3/32	5/32
$b = 1, c = 1$	7/32	1/32

pendency.

8.2 Conditional Berkelmans-Pries dependency function

In the discrete case, the definition of conditional independence is clear: for any $(a, b, c) \in E_A \times E_B \times E_C$ it must hold that

$$\mathbb{P}(X_A = a|X_C = c)\mathbb{P}(X_B = b|X_C = c) = \mathbb{P}(X_A = a, X_B = b|X_C = c),$$

or, if we allow X_A, X_B to be more general, we have conditional independence if and only if for all $c \in E_C$ we have

$$\mu_{X_A, X_B|\{X_C=c\}} = \mu_{X_A|\{X_C=c\}} \times \mu_{X_B|\{X_C=c\}}.$$

This last formulation looks a lot like the formulation for independence, but restricted to the set $\{X_C = c\}$. Thus suggesting that there is a reasonable measure of conditional dependency of A on B based on restricted dependencies. In other words, some transformation of

$$d_c = \text{Dep}\left(X_B|_{\{X_C=c\}}|X_A|_{\{X_C=c\}}\right).$$

One possible such transformation is simply

$$\text{Dep}_{\text{Cond},X_C}(X_B|X_A) = \frac{1}{\sum_c w_c} \sum_{c:w_c \neq 0} w_c d_c,$$

where w_c is $\mathbb{P}(X_C = c)$ if $X_B|_{\{X_C=c\}}$ is non-trivial (so d_c is defined), and 0 otherwise. Note, that if $w_c = 0$ for all c, this is not defined, but that is fine, since similarly to the general BP dependency

function, we then have simultaneously full conditional dependency as conditional independency.

The advantage of this formulation is that a lot of properties from the general BP dependency function translate to the conditional setting. Namely:

- Property II.1: It is asymmetric;
- Property II.2: It is between 0 and 1;
- Conditional version of Property II.3: It is 0 if and only if we have conditional independence;
- Conditional version of Property II.4: If X_B is completely determined by X_A and X_C it is equal to 1 (or undefined if X_B is completely determined by X_C);
- Property II.7: It is invariant under isomorphisms applied to X_A and X_B;
- Property II.8: It is non-increasing under measurable functions being applied to X_A.

Proof The first property: consider the case that X_C is almost surely constant. Then the expression reduces to the expression for $\mathrm{Dep}\,(X_B|X_A)$. Then using the asymmetric example from Section 6.5 shows asymmetricity.

The second property: $[0,1]$ is convex, therefore a weighted average of elements of $[0,1]$ is still in $[0,1]$.

The third property: we have conditional independence if and only if for all $c \in E_C$ we have that $X_A|_{\{X_C=c\}}$ and $X_B|_{\{X_C=c\}}$ are independent. But this is the case if and only if for all $c \in E_C$ we have $d_c = \mathrm{Dep}\left(X_B|_{\{X_C=c\}}|X_A|_{\{X_C=c\}}\right) = 0$. However this last statement is true if and only if $\mathrm{Dep}_{\mathrm{Cond},X_C}(X_B|X_A) = 0$.

The fourth property: if X_A, X_C completely determine X_B, then for any $c \in E_C$ we have that $X_B|_{\{}X_C = c\}$ is completely determined by $X_A|_{X_C=c}$. Therefore for all $c \in E_C$ we have $d_c = 1$ or d_c is undefined (due to $X_B|_{\{X_C=c\}}$ being trivial). So the weighted average of the d_c is therefore also equal to 1, or undefined in the case that for all

$c \in E_C$ we have that $X_B|_{\{X_C=c\}}$ is trivial. This last case is exactly the case that X_B is completely determined by X_C.

The seventh property: since isomorphisms applied to X_A and X_B induce isomorphisms on $X_A|_{\{X_C=c\}}$, $X_B|_{\{X_C=c\}}$, we have that the d_c are invariant, and therefore the conditional variant of the BP dependency functionis invariant.

The eight property: measurable functions on X_A induce measurable functions on $X_A|_{\{X_C=c\}}$. Therefore d_c can only reduce under such a measurable function. So the conditional variant of the BP dependency functioncannot increase.

Properties 5 and 6 for the conditional setting It is not immediately obvious what the conditional extension of Property II.5 would be, but one possible extension would be the following: for $X_1, X_2, X_3, \ldots, X_N$ conditionally independent on X_C (and non-trivial when restricted to $X_C = c$ for any c), and S a random variable taking values in $1, \ldots, N$ conditionally independent of $X_1, X_2, \ldots, X_N$ on X_C, and finally have $X_A = X_S$, then we should have $\mathrm{Dep}_{\mathrm{Cond},X_C}(X_i|X_A) = \mathbb{P}(S = i)$.

Then this holds for the conditional case since by Property II.5 of the unconditional case we have

$$d_c = \mathrm{Dep}\left(X_A|_{\{X_C=c\}}|X_i|_{\{X_C=c\}}\right) = \mathbb{P}(S = i|X_C = c),$$

and therefore, since w_c is not 0 for some c, we have

$$\begin{aligned}\mathrm{Dep}_{\mathrm{Cond},X_C}(X_i|X_A) &= \frac{1}{\sum_c w_c}\sum_{c:w_c\neq 0} w_c d_c \\ &= \sum_c \mathbb{P}(S = i|X_C = c)\mathbb{P}(X_C = c) \\ &= \mathbb{P}(S = i).\end{aligned}$$

Finally, there might exist an extension (that preserves all the properties listed above) to more general X_C, thus satisfying Property II.6,

however this remains an open problem at the time of writing.

8.3 Testing on real-world data

In this section we will evaluate this conditional dependency based on the real world data from Chapter 5 and compare it to the more conventional methodology used there to see whether this dependency measure is useful in a practical setting.

There are two questions we would like to consider: first, can the conditional BP dependency functioncompete when the assumptions of the conventional methodology are satisfied? Second, in the case these assumptions are not satisfied, can it find dependencies where the conventional methodology fails?

8.3.1 Methodology of the comparison

For each of the medications (M) tested, we evaluate the conditional BP dependency functionof the outcome of suicide on medication use conditioned on age, sex, and mental healthcare usage ($d(M)$). We then calculate p-values based on a naive simulation approach. We generate datasets with the same population size for each combination of age, sex, mental healthcare usage, and medication use. We then simulate the amount of suicides based on the suicide rate within the combination of age, sex, and mental healthcare usage. In other words assuming medication use and suicide are conditionally independent given age, sex, and mental healthcare usage. We then calculate the conditional BP dependency functionof suicide on medicaton usage given age, sex, and mental healthcare usage. This gives us for each simulation i the simulated dependency $\hat{d}_i(M)$, resulting in a p-value of

$$p_{sim}(M) = \frac{1}{N_{sim}+1}(|\{i : 1 \leq i \leq N_{sim}, \hat{d}_i(M) \geq d(M)\}| + 1).$$

We do this for N_{sim} equal to 1,000, 10,000, 100,000 and two naive dynamic stopping variants.

There are two naive dynamic stopping variants, one for threshold of $\alpha = 0.05$ (Dynamic_1), and one for threshold of $\alpha = 0.000352$

(Dynamic_2). For Dynamic_1 we stop after 1,000 simulations if the p-value is larger than 0.10, and continue to 10,000 simulations otherwise. For Dynamic_2 we stop after 1,000 simulations if the p-value is larger than 0.01, and continue to 100,000 simulations otherwise.

We then compare this to the results for conditional logistic regression. Are there medications for which the dependency methodology would lead to early rejection, but the conditional logistic regression does find a result? Are there medications for which conditional logistic regression fails to reject the null hypothesis, but the dependency methodology does reject independence?

Finally we compare run-times of the two methodologies. Since the server we run the analyses in has shared resources, we alternate the two methodologies so the impact of extra load by other users is more fairly distributed. This is also the reason that the logistic regression times can differ significantly between different simulation counts, although exactly the same operations are executed.

8.3.2 Results

Table 8.2: Runtimes of the conditional dependency under different simulation methods, compared to the runtimes of the conditional logistic regression.

Simulation method	Average run-time conditional dependency (s)	Average run-time conditional logistic regression (s)
1000 simulations	2.786	8.798
10,000 simulations	29.834	13.906
100,000 simulations	273.701	12.365
Dynamic_1	15.360	10.477
Dynamic_2	98.842	10.108

Table 8.2 shows the runtimes of the conditional dependency, compared to the conditional logistic regression. We note that the runtime of the conditional dependency is mainly determined by the number of simulations used for the p-values. Additionally Dynamic_1

already improves upon the runtime of 10,000 simulations by a factor of 2, whereas Dynamic_2 improves upon the runtime of 100,000 simulations by a factor of 3.

Table 8.3: Comparison of the rejection of the null-hypothesis ($\mathcal{H}_0$) under conditional logistic regression (CLR) and conditional dependency (CD)

Significance level	Rejections of $\mathcal{H}_0$ under both methods	Rejections of $\mathcal{H}_0$ under CLR but not CD	Rejections of $\mathcal{H}_0$ under CD but not CLR
0.05	85	26	2
0.000352	51	51	1

Table 8.4: Conditional dependency values for the various medications, with associated p-values, compared to p-values from logistic regression. Differences concerning significance at the $\alpha = 0.05$ or $\alpha = 0.000352$ level are marked with a '*' for the 0.05 level, '**' for the 0.000352 level, and '***' for both. The lower p-value is marked.

ATC4 code	Conditional dependency	p-value conditional dependency	p-value logistic regression
A01A	0.0036	0.15203	3.5E-6***
A02B	0.0668	0.00001	$< \epsilon$
A03A	0.0077	0.00001	$< \epsilon$
A03F	0.0235	0.00001	$< \epsilon$
A04A	0.0046	0.00003	3.4E-13
A05A	0.0008	0.59284	0.31
A06A	0.0670	0.00001	$< \epsilon$
A07A	0.0039	0.03276	$< \epsilon$***
A07D	0.0026	0.00199	3.0E-13**
A07E	0.0020	0.67224	1.2E-05***
A09A	0.0027	0.00003	$< \epsilon$
A10A	0.0102	0.00001	$< \epsilon$
A10B	0.0063	0.08321	2.2E-08***
A11C	0.0233	0.00001	$< \epsilon$
A11D	0.0003	0.00338	2.0E-13**
A12A	0.0157	0.00001	$< \epsilon$
A12B	0.0024	0.00064	$< \epsilon$**
B01A	0.0282	0.00001	$< \epsilon$
B02A	0.0012	0.14680	0.37
B02B	0.0011	0.15493	3.4E-05***
B03A	0.0078	0.00119	$< \epsilon$**
B03B	0.0098	0.00001	$< \epsilon$
B03X	0.0009	0.36710	0.47
B05B	0.0026	0.00530	9.8E-11**
C01A	0.0022	0.06075	9.6E-8***
C01B	0.0023	0.02588	1.0E-4**
C01C	0.0019	0.38461	0.78
C01D	0.0057	0.00019	7.8E-7
C01E	0.0007	0.09912	5.1E-5***
C02A	0.0008	0.36505	1.1E-3*
C02C	0.0013	0.30905	0.38
C03A	0.0111	0.00001	3.0E-4
C03B	0.0025	0.02213	2.0E-3
C03C	0.0122	0.00001	$< \epsilon$

ATC4 code	Conditional dependency	*p*-value conditional dependency	p-value logistic regression
C03D	0.0042	0.00072	$< \epsilon^{**}$
C03E	0.0017	0.33146	0.66
C05A	0.0043	0.02098	3.4E-10**
C07A	0.0247	0.00001	$< \epsilon$
C07B	0.0016	0.03815*	0.89
C07C	0.0005	0.87103	0.31
C08C	0.0101	0.00007	$< \epsilon$
C08D	0.0041	0.00025	2.1E-10
C09A	0.0121	0.00001	$< \epsilon$
C09B	0.0029	0.10782	0.48
C09C	0.0074	0.01138	1.7E-09**
C09D	0.0056	0.00164*	0.36
C09X	0.0006	0.59176	0.25
C10A	0.0216	0.00001	$< \epsilon$
C10B	0.0016	0.07218	4.1E-2*
D01A	0.0086	0.08024	1.3E-15***
D01B	0.0033	0.06997	2.1E-2*
D02A	0.0123	0.02222	$< \epsilon^{**}$
D02B	0.0013	0.32582	0.24
D04A	0.0012	0.21124	2.7E-3*
D05A	0.0026	0.21376	0.58
D05B	0.0007	0.25637	0.76
D06A	0.0152	0.00004	$< \epsilon$
D06B	0.0061	0.00296	$< \epsilon^{**}$
D07A	0.0162	0.03463	$< \epsilon^{**}$
D07C	0.0006	0.61365	0.10
D07X	0.0086	0.03961	7.6E-11**
D08A	0.0005	0.26527	9.0E-2
D10A	0.0083	0.00048	1.3E-12**
D10B	0.0017	0.05717	8.3E-05***
D11A	0.0046	0.05632	0.11
G01A	0.0043	0.37735	9.5E-07***
G02B	0.0012	0.56813	0.62
G02C	0.0005	0.80848	0.22
G03A	0.0125	0.00007**	3.5E-2
G03B	0.0009	0.04640	2.5E-06**
G03C	0.0045	0.00122	2.2E-16**
G03D	0.0025	0.46411	5.8E-05***
G03F	0.0025	0.00001	$< \epsilon$
G03H	0.0041	0.00197	7.9E-09**
G04B	0.0060	0.00002	$< \epsilon$
G04C	0.0085	0.00001	$< \epsilon$
H01A	0.0003	0.23253	9.2E-08***
H02A	0.0225	0.00001	$< \epsilon$
H02B	0.0011	0.18758	0.13
H03A	0.0112	0.00001	$< \epsilon$
H03B	0.0010	0.78754	0.25
H04A	0.0019	0.05141	2.0E-07***
J01A	0.0125	0.00071	$< \epsilon^{**}$
J01C	0.0346	0.00001	$< \epsilon$
J01D	0.0020	0.00217	$< \epsilon^{**}$
J01E	0.0096	0.00001	$< \epsilon$
J01F	0.0137	0.00003	$< \epsilon$
J01M	0.0136	0.00001	$< \epsilon$
J01X	0.0161	0.00001	$< \epsilon$
J02A	0.0069	0.00124	$< \epsilon^{**}$
J04A	0.0005	0.84729	0.14
J05A	0.0052	0.00021	$< \epsilon^{**}$
J07A	0.0034	0.10472	8.3E-12***
J07B	0.0008	0.19277	2.8E-2*
L01B	0.0015	0.88787	0.99
L02A	0.0014	0.19528	6.3E-05***
L02B	0.0020	0.16891	3.0E-05***
L03A	0.0017	0.00243	2.0E-2
L04A	0.0028	0.46670	0.21
M01A	0.0358	0.00001	$< \epsilon$
M03B	0.0046	0.00001	$< \epsilon$
M04A	0.0031	0.07902	1.4E-4***
M05B	0.0071	0.00001	$< \epsilon$
N01B	0.0088	0.00005	$< \epsilon$
N02A	0.0667	0.00001	$< \epsilon$
N02B	0.0142	0.00001	$< \epsilon$
N02C	0.0071	0.00141	$< \epsilon$
N03A	0.0471	0.00001	$< \epsilon$
N04A	0.0072	0.00001	$< \epsilon$
N04B	0.0045	0.00001	$< \epsilon$

ATC4 code	Conditional dependency	p-value conditional dependency	p-value logistic regression
N05A	0.1076	0.00001	$< \epsilon$
N05B	0.1037	0.00001	$< \epsilon$
N05C	0.0701	0.00001	$< \epsilon$
N06A	0.2037	0.00001	$< \epsilon$
N06B	0.0123	0.00001	$< \epsilon$
N06D	0.0010	0.31454	0.38
N07A	0.0008	0.01236	3.4E-4**
N07B	0.0152	0.00001	$< \epsilon$
N07C	0.0026	0.18741	3.07E-06***
N07X	0.0008	0.00455	2.07E-07**
P01A	0.0051	0.01602	9.71E-10**
P01B	0.0020	0.06088	5.6E-3*
P03A	0.0009	0.88581	0.44
R01A	0.0161	0.00934	3.2E-09**
R03A	0.0253	0.00001	$< \epsilon$
R03B	0.0163	0.00001	$< \epsilon$
R03D	0.0033	0.02613	6.4E-07**
R05C	0.0006	0.09243	3.5E-3*
R05D	0.0095	0.02498	8.9E-11**
R06A	0.0223	0.00001	$< \epsilon$
S01A	0.0117	0.00220	2.2E-16**
S01B	0.0049	0.03649	2.0E-6**
S01C	0.0043	0.12073	8.7E-8***
S01E	0.0054	0.00116	4.4E-14**
S01F	0.0011	0.52280	0.18
S01G	0.0072	0.08258	0.64
S01X	0.0167	0.00001	$< \epsilon$
S02A	0.0024	0.21135	3.4E-5***
S02C	0.0103	0.00387	$< \epsilon$
V01A	0.0017	0.12115	0.74
V03A	0.0008	0.27178	2.8E-2*
V07A	0.0015	0.00178	1.7E-4**

In Table 8.3 we see a summary of the results. In Table 8.4. we see the conditional dependency for the various medications as well as associated p-values. We also see the p-values that came out of the conditional logistic regression model. We note that in most cases these agree on whether the null hypothesis can be rejected or not. There are a number exceptions where the assumptions of conditional logistic regression are sufficiently satisfied to not reject in the conditional dependency case but do reject in the conditional logistic regression case. There are also some where the reverse holds, chief of which would be C07B (beta blocking agents and thiazides), which the conditional dependency would reject based on $\alpha = 0.05$ but which conditional logistic regression gives a p-value of 0.89. When looking at the sensitivity checks from Section 5.A.6 we do indeed see a violation of the core assumption that the change in risk is unidirectional: in the mental health cohort it is associated with an increase in risk, whereas in the non-mental healthcare cohort it is associated with a reduction in risk. The same is true for C09D ('angiotensin II receptor blockers (ARBs), combinations'), though to a lesser extent. At the 0.000352 level we see similar behaviour

with G03A ('hormonal contraceptives for systemic use').

8.4 Discussion

In this chapter we extended the notion of the BP dependency functionto the setting of conditional dependency. We showed it held some nice theoretical results. When tested on real world data against a conventional methodology, we saw that they had comparable runtimes, with the conditional dependency lagging behind somewhat. This was caused mostly by the naive way the p-value was calculated. Given the vast improvements from a naive early stopping algorithm it is very plausible that a smarter early stopping algorithm would have runtimes on par with or even beating conditional logistic regression. Additionally, by investigating the distribution of these simulations a structure might be discovered that would allow one to calculate p-values based on the mean and variance estimators alone.

When it came to detecting dependency structures, we observed that the conditional dependency measure was disadvantaged in the cases where the assumptions of the conditional logistic regression were satisfied (which was most of them), but was much better at detecting dependency structures in the cases where they were violated, especially when the effects in the different subpopulations were almost exactly equal but in opposing directions, as seen with C07B and to a lesser extent with C09D and G03A.

In conclusion, based on both theoretical and numerical results, we believe the conditional dependency has the potential to complement existing methodology, and function as an additional tool for statistical analysis.

Summary

This thesis is split into two major parts. In Part I we looked at population data and, using conventional methods such as logistic regression and a novel extension thereof, we considered socio-demographic risk factors for suicide to gain more insight in who dies by suicide, which is essential to be able to effectively deploy selective interventions.

In Chapter 2 we compared characteristics of youth suicide victims and adult suicide victims. We found that the main differences between youths and adults concerned the magnitude of the added risk from the various risk factors. This was larger among adults than among youths. Antoher substantial difference was the method of suicide that was picked.

In Chapter 3 we used a logistic regression approach to test a large collection of characteristics on whether they were indicative of a higher suicide risk. Among these we found males, people of middle age, people living alone, people with high healthcare costs, people on benefits, people with a low income, and those that had no external migration background had an increased risk.

In Chapter 4 we investigated whether or not there were combinations of risk factors that resulted in a disproportionate increase in risk. We found numerous combinations where the risk differed substantially from the assumptions. Among these there was one group with a

suicide risk almost eight times the national average: people that were never married and unfit for work. Additionally, there were two more groups where the suicide risk was more than four times the national average: males that were unfit for work, and people aged 55-69 who lived alone, were never married and had a low household income. Finally, there were two groups which would not have been found based on their individual risk factors, namely widowed males and people between 25 and 40 years old with a low level of education.

In Chapter 5 we considered prescribed medications and what kind of association this had with suicide risk. We found that most medication classes had at least some association with suicide risk. There were two prominent clusters: that of drugs affecting the nervous system, and that of drugs affecting metabolism. Together, the chapters in Part I reveal numerous high risk groups where preventive interventions could play a key role in reducing suicides.

In Part II we took a more theoretical approach. In Chapter 6 we considered the notion of dependency. What does it mean for one type of observation to depend on another, and is there a good way to measure this? We found the answer to the first question to be rather complicated and the answer to the second to be 'no'. We then developed our own way of measuring the abstract notion of 'dependency' and showed that it satisfied multiple basic general properties, unlike the previously proposed methods of measuring this such as the widely used Pearson's correlation coefficient.

In Chapter 7 we combined the proposed dependency measure with the concept of the Shapley value from cooperative game theory. This resulted in a measure for feature importance and attempts to answer an important question in machine learning: how useful is a certain feature in predicting an outcome of interest? We tested certain logical properties one would desire from a measure of feature importance and compared it to existing methodologies. Our proposed notion of feature importance outperformed all others by a significant margin. Together, these two chapters developed methods for dependency and feature importance that meet both theoretical, as well as practical demands, are flexible, and additionally

outperform the existing methods by a wide margin.

Finally, in Chapter 8 we proposed an extension to the setting of *conditional* dependence and showed it satisfied conditional versions of the properties introduced in Chapter 6. We also compared the performance of our notion of conditional dependency to conventional conditional logistic regression on the dataset from Chapter 5. We found comparable detection rates and comparable runtimes, though some associations were picked up by the conventional method and not the conditional dependency, and vice versa. Both methodologies found strong associations between medication usage and suicide, even when age, sex, and mental healthcare usage were accounted for.

D

Discussion and outlook

D.1 What insights have we gained?

Our work contributed to the identification of people at high risk of suicide. We found (and confirmed previously found) numerous high-risk groups of interest, namely males, people of middle age, people on benefits, people with high healthcare costs, people living alone, having a Dutch migration background, and having a low income. We also found a number of sub-populations corresponding to interactions of risk factors, identifying ultrahigh-risk groups with suicide rates up to 88.48 per 100,000, including those never married and unfit for work, males that were unfit for work, and people aged 55-69 that live alone, have a household income in the bottom quartile, and were never married. There were also sub-populations which would not have been found to be of increased risk based on individual risk factors alone, but where the risk factors combined do lead to a heightened risk, namely widowed males and people between 25 and 40 years old with a low level of education.

On the theoretical side, we found that a relatively straightforward dependency measure satisfies properties that seem to be quite basic, and yet which existing measures of dependency regularly violate. We found that this dependency measure could be used to define an effective measure of feature importance which satisfies intuitively

logical properties which most other measures of feature importance fail. We also showed using both conventional methodology as well as a conditional variant of our notion of dependency that there is a clear association between medication usage and suicide.

D.2 Are these insights useful?

There are generally three levels to interventions: universal (targeted at the whole population), selective (targeted at specific high-risk groups), and indicated (targeted at individuals).

The models concerning suicide risk we developed are not remotely good enough for screening on an individual level, because suicide is a rare (yet catastrophic) event. Current state of the art models would still require many false positives to label a true suicidal individual. These models are therefore not useful for indicated interventions.

However, the fact that these models pointed us to previously unknown risk groups allows for implementing selective interventions targeted at said risk groups. Examples of such interventions are screening for signals or training gatekeepers in the relevant fields. Or it might even be possible to develop new interventions specifically targeted at these risk groups.

The dependency measure in Chapter 6 and the feature importance in Chapter 7 are useful in a more general setting than the scope of this thesis, and will be useful in any data driven setting. What it does share with the scope of this thesis is that it is an example of results from research being able to be transferred to other settings and fields. The results from Chapter 5 should help make clear the importance of not only monitoring for possible physical side effects of medications, but also for mental ones.

D.3 What remains to be done?

Here we need to distinguish between four possible avenues to continue from this thesis, three of which concern research, and one

of which concerns implementation. First, research that would also have fitted the scope of this thesis: data driven research into suicide prevention. Second, research in other fields that is based on the results in this thesis. Third, a continuation of the theoretical frameworks developed in this thesis. Fourth, policy considerations based on the results in this thesis.

D.3.1 Data driven research into suicide prevention

Since this thesis was exclusively based on data from Statistics Netherlands, the results were also limited to the Netherlands. It would be interesting to see whether the results as found in this thesis can be reproduced using data from other countries, or if the high-risk groups and combinations of risk factors leading to high risk might be different in other countries.

We were limited to suicides, but lacked any kind of data on non-fatal suicide attempts. It might be interesting to see whether it might be possible to systematically log suicide attempts, for example when they present themselves to first responders. This data could then possibly be connected to the Statistics Netherlands databases, allowing researchers to expand the studies underlying this thesis to suicide attempts.

Since other countries have different types of data, it might be interesting to see what kind of insights could be gained from those databases that complement our own.

Additionally, it might be prudent to examine and log the specific medications a suicide victim was prescribed in the months leading up to the suicide. After all, one of the main limitations in Chapter 5 was a lack of data on specific medications (only knowing up to ATC4 level) as well as only knowing it was prescribed in a certain calendar year.

D.3.2 Other research into suicide prevention

As shown in this thesis, new insights can be gained when looking at a field of research through the eyes of another field. It would therefore be interesting to see what other interdisciplinary approaches might reveal. An interesting avenue to explore would, for example, be looking at the biological aspect of suicide prevention. As we showed in Chapter 5, there are associations to be found between medications and suicide. The extent to which these associations represent causal relationships remains to be investigated. Also, it might be interesting to explore possible (neuro)biologically mechanisms underlying the associations, to better understand biological pathways contributing to suicidal behaviours.

D.3.3 Continuation of theoretical framework

With regard to the theoretical framework introduced in the second half of this thesis, there are a lot of open questions left to be explored, concerning matters such as estimator convergence, estimator distribution, and confidence intervals of the dependency measure. Also there is the question whether there are other possible measures that satisfy the eight formulated requirements. Additionally, there remain questions about the best ways to handle continuous and mixed random variables.

There are also open questions concerning the feature importance. When the number of features grows large calculating the Shapley value precisely becomes computationally infeasible. Are the properties robust to numerical approximations?

D.3.4 Policy considerations

There also remain many questions regarding prevention policies. How can the risk groups found best be targeted? Are there places where they can easily be found? For example, given that numerous high risk groups are associated with unfit for work benefits, one could train the doctors who are responsible for assessing their unfit for work status to be gatekeepers. Or given the high degree of risk all around, one could even train all employees interacting with people

on benefits. Similarly, there are numerous groups that interact with people of low income who could be trained as gatekeepers. The group of widowed males could be preventively targeted by training general practitioners in proper aftercare and signal detection.

The results regarding medications found in Chapter 5 also raise questions on how best to proceed. One possibility could be better monitoring by general practitioners when prescribing said medications. Another possibility would be to be more explicit about the possible mental side effects of medications being prescribed, so that individuals on said medications make the connection.

D.4 Conclusion

This thesis has added a piece of the puzzle to answering the question raised at the start: 'Who are the people who die by suicide?'. We have found socio-demographic groups that have a risk that is up to eight times as high as the general population. We have found medication classes where even corrected for age, sex, and mental healthcare usage, the risk was increased up to nine times. It is now crucial to find ways to reach these risk groups and implement interventions to reduce the risk of suicide among them.

Additionally during the research methodological questions arose. To answer these questions we made a considerable theoretical contribution to the field of data science and machine learning. This contribution took the form of a measure of dependency of random variables, and a measure of feature importance. Both these measures satisfied a great number of desired properties which existing methods failed to do.

There were both practical and theoretical insights gained from the research underlying this thesis, showing the value of bringing in different people with different types of expertise. However, a lot of questions remain to be answered, and a lot of actions remain to be taken to reduce the number of suicides. This requires effort from a wide range of people, from researchers, general practitioners, policy makers, to the general public. As Jan Mokkenstorm, the founder of 113 Suicide Prevention, once said: 'We won't rest until the number

is zero.'

Bibliography

[1] *Aantal flexwerkers in 15 jaar met drie kwart gegroeid*. Feb. 2019. URL: https://www.cbs.nl/nl-nl/nieuws/2019/07/aantal-flexwerkers-in-15-jaar-met-drie-kwart-gegroeid (visited on 15/06/2022) (cited on page 59).

[2] K. Aas, M. Jullum and A. Løland. 'Explaining individual predictions when features are dependent: More accurate approximations to Shapley values'. In: *Artificial Intelligence* 298 (2021), page 103502 (cited on pages 152, 157, 184, 191).

[3] N. Abe and M. Kudo. 'Entropy criterion for classifier-independent feature selection'. In: *Knowledge-Based Intelligent Information and Engineering Systems*. Edited by R. Khosla, R. J. Howlett and L. C. Jain. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pages 689–695 (cited on pages 172, 174, 184).

[4] R. Agarwal, P. Sacre and S. V. Sarma. *Mutual dependence: A novel method for computing dependencies between random variables*. 2015. arXiv: 1506.00673 [math.ST] (cited on pages 109, 110, 112, 136).

[5] B. K. Ahmedani et al. 'Detecting and distinguishing indicators of risk for suicide using clinical records'. In: *Translational Psychiatry* 12.1 (July 2022), pages 1–9 (cited on page 74).

[6] B. K. Ahmedani et al. 'Health care contacts in the year before suicide death'. In: *Journal of General Internal Medicine* 29.6 (June 2014), pages 870–877 (cited on page 74).

[7] A. D. Althouse. 'Adjust for multiple comparisons? It's not that simple'. In: *The Annals of thoracic surgery* 101.5 (2016), pages 1644–1645 (cited on page 18).

[8] A. Altmann, L. Toloşi, O. Sander and T. Lengauer. 'Permutation importance: a corrected feature importance measure'. In: *Bioinformatics* 26.10 (Apr. 2010), pages 1340–1347 (cited on pages 151, 184).

[9] F. Aragón-Royón, A. Jiménez-Vílchez, A. Arauzo-Azofra and J. M. Benítez. *FSinR: an exhaustive package for feature selection*. 2020 (cited on pages 172, 174).

[10] E. Arensman, M. Bennardi, C. Larkin, A. Wall, C. McAuliffe, J. McCarthy, E. Williamson and I. J. Perry. 'Suicide among young people and adults in Ireland: Method characteristics, toxicological analysis and substance abuse histories compared'. In: *PLoS one* 11.11 (2016), e0166881 (cited on pages 15, 25, 29).

[11] G. Arsenault-Lapierre, C. Kim and G. Turecki. 'Psychiatric diagnoses in 3275 suicides: a meta-analysis'. In: *BMC Psychiatry* 4.1 (Nov. 2004), page 37 (cited on page 41).

[12] G. Ayhan, R. Arnal, C. Basurko, V. About, A. Pastre, E. Pinganaud, D. Sins, L. Jehel, B. Falissard and M. Nacher. 'Suicide risk among prisoners in French Guiana: prevalence and predictive factors'. In: *BMC Psychiatry* 17.1 (Dec. 2017), page 156 (cited on pages 33, 46).

[13] L. P. Barbosa, L. Quevedo, G. D. G. da Silva, K. Jansen, R. T. Pinheiro, J. Branco, D. Lara, J. Oses and R. A. da Silva. 'Childhood trauma and suicide risk in a sample of young individuals aged 14–35 years in southern Brazil'. In: *Child Abuse & Neglect* 38.7 (2014), pages 1191–1196 (cited on page 15).

[14] A. Batt, F. Bellivier, B. Delatte, O. Spreux-Varoquaux, D. Cremniter, V. Dubreu et al. 'Suicide: Psychological autopsy, a research tool for prevention'. In: *National Institute for Health and Medical Research Collective Expert Report* (2004) (cited on page 2).

[15] M. Bauwelinck, P. Deboosere, D. Willaert and H. Vandenheede. 'Suicide mortality in Belgium at the beginning of the 21st century: differences according to migrant background'. In: *The European Journal of Public Health* 27.1 (2017), pages 111–116 (cited on pages 23, 29).

[16] G. Berkelmans, J. Pries, R. D. van der Mei and S. Bhulai. 'The BP dependency function: a generic measure of dependence between random variables'. To appear in Journal of Applied Probability 60.4. Dec. 2023 (cited on pages 11, 101, 153, 155, 156, 158–161, 163, 165, 188, 194).

[17] G. Berkelmans, L. J. Schweren, S. Bhulai, R. D. van der Mei and R. Gilissen. 'Identifying populations at ultra-high risk of suicide using a novel machine learning method'. In: *Comprehensive Psychiatry* 123 (2023), page 152380 (cited on pages 11, 45).

[18] G. Berkelmans, L. J. Schweren, R. D. van der Mei, S. Bhulai, R. Gilissen and A. Beekman. 'On the relation between medication prescriptions and suicide'. Submitted for publication (cited on pages 11, 73).

[19] G. Berkelmans, R. D. van der Mei, S. Bhulai and R. Gilissen. 'Identifying socio-demographic risk factors for suicide using data on an individual level'. In: *BMC Public Health* 21.1 (2021), pages 1–8 (cited on pages 10, 33, 46).

[20] G. Berkelmans, R. D. van der Mei, S. Bhulai, S. Merelle and R. Gilissen. 'Demographic risk factors for suicide among youths in the Netherlands'. In: *International Journal of Environmental Research and Public Health* 17.4 (2020), pages 1–11 (cited on pages 10, 15, 41).

[21] M. Bhatt et al. 'Profile of suicide attempts and risk factors among psychiatric patients: A case-control study'. In: *PLOS ONE* 13.2 (Feb. 2018), e0192998 (cited on pages 33, 46, 59).

[22] M. Bierlaire. *A short introduction to PandasBiogeme*. URL: http://transp-or.epfl.ch/documents/technicalReports/Bier23.pdf (cited on page 36).

[23] J. Bilsen. 'Suicide and youth: risk factors'. In: *Frontiers in psychiatry* (2018), page 540 (cited on pages 15, 16).

[24] A. D. Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante and R. Vezzani, editors. *Pattern recognition. ICPR International workshops and challenges*. Springer International Publishing, 2021 (cited on pages 172, 174, 184).

[25] S. Bloch-Elkouby, B. Gorman, A. Schuck, S. Barzilay, R. Calati, L. J. Cohen, F. Begum and I. Galynker. 'The suicide crisis syndrome: A network analysis.' In: *Journal of counseling psychology* 67.5 (2020), page 595 (cited on page 2).

[26] M. Bosak, W. Turaj, D. Dudek, M. Siwek and A. Szczudlik. 'Depressogenic medications and other risk factors for depression among Polish patients with epilepsy'. In: *Neuropsychiatric Disease and Treatment* 11 (2015), pages 2509–2517 (cited on page 74).

[27] K. L. Bower and K. G. Emerson. 'Exploring contextual factors associated with suicide among older male farmers: Results from the CDC NVDRS dataset'. In: *Clinical Gerontologist* 44.5 (Oct. 2021), pages 528–535 (cited on page 59).

[28] *BP dependency function repository*. https://github.com/joris-pries/Official_Dependency_Function (cited on page 103).

[29] *BP feature importance repository*. URL: https://github.com/joris-pries/BP-Feature-Importance (cited on page 153).

[30] R. C. Bradley. 'Basic properties of strong mixing conditions. A survey and some open Questions'. In: *Probability Surveys* 2.none (2005), pages 107–144 (cited on pages 111, 112).

[31] J. A. Bridge, J. B. Greenhouse, D. Ruch, J. Stevens, J. Ackerman, A. H. Sheftall, L. M. Horowitz, K. J. Kelleher and J. V. Campo. 'Association between the release of Netflix's 13 Reasons Why and suicide rates in the United States: An interrupted time series analysis'. In: *Journal of the American Academy of Child & Adolescent Psychiatry* 59.2 (2020), pages 236–243 (cited on pages 24, 29).

[32] W. Bunney, A. Kleinman, T. Pellmar, S. Goldsmith et al. 'Reducing suicide: A national imperative'. In: (2002) (cited on page 3).

[33] C. Burnette, R. Ramchand and L. Ayer. 'Gatekeeper training for suicide prevention: A theoretical model and review of the empirical literature'. In: *Rand health quarterly* 5.1 (2015) (cited on page 3).

[34] T. Callréus, U. Agerskov Andersen, J. Hallas and M. Andersen. 'Cardiovascular drugs and the risk of suicide: a nested case-control study'. In: *European Journal of Clinical Pharmacology* 63.6 (June 2007), pages 591–596 (cited on page 74).

[35] L. Capitani, L. Bagnato and A. Punzo. 'Testing serial independence via density-based measures of divergence'. In: *Methodology And Computing In Applied Probability* 16 (Aug. 2014), pages 627–641 (cited on pages 112, 113).

[36] M. Carletti, C. Masiero, A. Beghi and G. A. Susto. 'Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis'. In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019, pages 21–26 (cited on page 184).

[37] M. Carletti, M. Terzi and G. A. Susto. *Interpretable anomaly detection with DIFFI: Depth-based Isolation Forest Feature Importance*. 2020 (cited on pages 172, 174).

[38] G. Casalicchio, C. Molnar and B. Bischl. 'Visualizing the feature importance for black box models'. In: *Machine Learning and Knowledge Discovery in Databases*. Edited by M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley and G. Ifrim.

Cham: Springer International Publishing, 2019, pages 655–670 (cited on page 184).

[39] G. Castelpietra, M. Gobbato, F. Valent, M. Bovenzi, F. Barbone, E. Clagnan, E. Pascolo-Fabrici, M. Balestrieri and G. Isacsson. 'Somatic disorders and antidepressant use in suicides: A population-based study from the Friuli Venezia Giulia region, Italy, 2003–2013'. In: *Journal of Psychosomatic Research* 79.5 (Nov. 2015), pages 372–377 (cited on page 84).

[40] J. Castro, D. Gómez and J. Tejada. 'Polynomial calculation of the Shapley value based on sampling'. In: *Computers & Operations Research* 36.5 (2009), pages 1726–1730 (cited on page 191).

[41] B. Cavanagh, S. Ibrahim, A. Roscoe, H. Bickley, D. While, K. Windfuhr, L. Appleby and N. Kapur. 'The timing of general population and patient suicide in England, 1997–2012'. In: *Journal of affective disorders* 197 (2016), pages 175–181 (cited on pages 26, 29).

[42] C. M. Celano, O. Freudenreich, C. Fernandez-Robles, T. A. Stern, M. A. Caro and J. C. Huffman. 'Depressogenic effects of medications: a review'. In: *Dialogues in Clinical Neuroscience* 13.1 (2011), pages 109–125 (cited on page 74).

[43] E. Celik. *vita: Variable Importance Testing Approaches*. 2015 (cited on pages 172, 174).

[44] *Centraal Bureau voor de Statistiek*. URL: https://www.cbs.nl (cited on page 5).

[45] J. Cerel, M. M. Brown, M. Maple, M. Singleton, J. Venne, M. Moore and C. Flaherty. 'How many people are exposed to suicide? Not six'. In: *Suicide and Life-Threatening Behavior* 49.2 (Apr. 2019), pages 529–534 (cited on page 1).

[46] T. Chen and C. Guestrin. 'XGBoost: A scalable tree boosting system'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pages 785–794 (cited on pages 172, 174).

[47] P. Cheng, W. Hao and L. Carin. *Estimating Total Correlation with Mutual Information Bounds*. 2020 (cited on page 136).

[48] S. B. Choi, W. Lee, J.-H. Yoon, J.-U. Won and D. W. Kim. 'Risk factors of suicide attempt among people with suicidal ideation in South Korea: a cross-sectional study'. In: *BMC Public Health* 17.1 (June 2017), page 579 (cited on pages 33, 46).

[49] I. Covert, S. M. Lundberg and S.-I. Lee. 'Understanding Global Feature Contributions With Additive Importance Measures'. In: *Advances in neural information processing systems*. Edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin. Volume 33. Curran Associates, Inc., 2020, pages 17212–17223 (cited on pages 172, 174).

[50] C. Crump, K. Sundquist, J. Sundquist and M. A. Winkleby. 'Sociodemographic, psychiatric and somatic risk factors for suicide: a Swedish national cohort study'. In: *Psychological Medicine* 44.2 (Jan. 2014), pages 279–289 (cited on page 84).

[51] A. Datta, S. Sen and Y. Zick. 'Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems'. In: *2016 IEEE Symposium on Security and Privacy (SP)*. 2016, pages 598–617 (cited on page 154).

[52] C. R. de Sá. 'Variance-based feature importance in neural networks'. In: *Discovery Science*. Edited by P. Kralj Novak, T. Šmuc and S. Džeroski. Cham: Springer International Publishing, 2019, pages 306–315 (cited on pages 172, 174).

[53] K. Dhamdhere, A. Agarwal and M. Sundararajan. 'The Shapley Taylor interaction index'. In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020 (cited on page 184).

[54] D. Dilillo, S. Mauri, C. Mantegazza, V. Fabiano, C. Mameli and G. V. Zuccotti. 'Suicide in pediatrics: epidemiology, risk factors, warning signs and the role of the pediatrician in detecting them'. In: *Italian journal of pediatrics* 41.1 (2015), pages 1–8 (cited on page 16).

[55] P. Embrechts, A. J. McNeil and D. Straumann. 'Correlation and dependence in risk management: Properties and pitfalls'. In: *Risk Management: Value at Risk and Beyond*. Cambridge: Cambridge University Press, 2002, pages 176–223 (cited on pages 108, 109).

[56] F. Engel, T. Stadnitski, E. Stroe-Kunold, S. Berens, R. Schäfert and B. Wild. 'Temporal relationships between abdominal pain, psychological distress and coping in patients with IBS - A time series approach'. In: *Frontiers in Psychiatry* 13 (2022), page 768134 (cited on page 83).

[57] F. Fang, K. Fall, M. A. Mittleman, P. Sparén, W. Ye, H.-O. Adami and U. Valdimarsdóttir. 'Suicide and cardiovascular death after a cancer diagnosis'. In: *The New England Journal of Medicine* 366.14 (Apr. 2012), pages 1310–1318 (cited on page 74).

[58] *FOMAT: Procedure overlijden*. URL: http://www.fomat.nl/proc-overlijden.html (visited on 30/08/2019) (cited on page 17).

[59] J. C. Franklin, J. D. Ribeiro, K. R. Fox, K. H. Bentley, E. M. Kleiman, X. Huang, K. M. Musacchio, A. C. Jaroszewski, B. P. Chang and M. K. Nock. 'Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research'. In: *Psychological Bulletin* 143 (2017), pages 187–232 (cited on pages 3, 33, 42, 46, 73).

[60] C. Frye, C. Rowat and I. Feige. 'Asymmetric Shapley Values: incorporating causal knowledge into model-agnostic explainability'. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc., 2020 (cited on page 184).

[61] D. V. Fryer, I. Strumke and H. Nguyen. 'Model independent feature attributions: Shapley values that uncover non-linear dependencies'. In: *PeerJ Computer Science* 7 (2021), e582 (cited on pages 152, 154, 169, 172, 174, 184, 185).

[62] H. Gebelein. 'Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung'. In: *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 21 (6 Jan. 1941), pages 364–379 (cited on pages 110, 112).

[63] G. Geulayov, D. Casey, K. C. McDonald, P. Foster, K. Pritchard, C. Wells, C. Clements, N. Kapur, J. Ness, K. Waters et al. 'Incidence of suicide, hospital-presenting non-fatal self-harm, and community-occurring non-fatal self-harm in adolescents in England (the iceberg model of self-harm): a retrospective study'. In: *The Lancet Psychiatry* 5.2 (2018), pages 167–174 (cited on page 16).

[64] A. Ghorbani, D. Berenbaum, M. Ivgi, Y. Dafna and J. Y. Zou. 'Beyond importance scores: Interpreting tabular ML by visualizing feature semantics'. In: *Information* 13.1 (2022) (cited on pages 172, 174).

[65] O. Giles et al. *Faking feature importance: A cautionary tale on the use of differentially-private synthetic data*. 2022. arXiv: 2203.01363 [cs.LG] (cited on pages 184, 185).

[66] H. C. Gorton, R. T. Webb, N. Kapur and D. M. Ashcroft. 'Non-psychotropic medication and risk of suicide or attempted suicide: a systematic review'. In: *BMJ Open* 6.1 (Jan. 2016), e009074 (cited on page 74).

[67] J. L. Gradus, A. J. Rosellini, E. Horváth-Puhó, A. E. Street, I. Galatzer-Levy, T. Jiang, T. L. Lash and H. T. Sørensen. 'Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark'. In: *JAMA Psychiatry* 77.1 (Jan. 2020), pages 25–34 (cited on pages 34, 46).

[68] A. Gramacki. *Nonparametric kernel density estimation and its computational aspects*. 1st. New York: Springer Publishing Company, Incorporated, 2017 (cited on page 119).

[69] C. W. Granger, E. Maasoumi and J. Racine. 'A dependence metric for possibly nonlinear processes'. In: *Journal of Time Series Analysis* 25.5 (2004), pages 649–669 (cited on page 109).

[70] B. M. Greenwell and B. C. Boehmke. 'Variable importance plots—An introduction to the vip Package'. In: *The R Journal* 12.1 (2020), pages 343–366 (cited on pages 172, 174).

[71] A. Gretton, R. Herbrich, A. Smola, O. Bousquet and B. Schölkopf. 'Kernel methods for measuring independence'. In: *Journal of Machine Learning Research* 6 (2005), pages 2075–2129 (cited on page 109).

[72] W.-Z. Hao, X.-J. Li, P.-W. Zhang and J.-X. Chen. 'A review of antibiotics, depression, and the gut microbiome'. In: *Psychiatry Research* 284 (Feb. 2020), page 112691 (cited on page 74).

[73] E. Hellinger. 'Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen.' In: *Journal für die reine und angewandte Mathematik* 1909.136 (1909), pages 210–271 (cited on page 110).

[74] S. Hooker, D. Erhan, P.-J. Kindermans and B. Kim. 'A benchmark for interpretability methods in deep neural networks'. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cited on pages 152, 153, 157, 184).

[75] H. Hotelling. 'Relations between two sets of variates'. In: *Biometrika* 28.3/4 (1936), pages 321–377 (cited on page 112).

[76] T. Hothorn and A. Zeileis. 'partykit: A modular toolkit for recursive Partytioning in R'. In: *Journal of Machine Learning Research* 16 (2015), pages 3905–3909 (cited on pages 172, 174).

[77] X. Huang, J. D. Ribeiro, K. M. Musacchio and J. C. Franklin. 'Demographics as predictors of suicidal thoughts and behaviors: A meta-analysis'. In: *PloS one* 12.7 (2017), e0180793 (cited on page 16).

[78] V. A. Huynh-Thu, Y. Saeys, L. Wehenkel and P. Geurts. 'Statistical interpretation of machine learning-based feature importance scores for biomarker discovery'. In: *Bioinformatics* 28.13 (Apr. 2012), pages 1766–1774 (cited on page 184).

[79] R. J. Janse, T. Hoekstra, K. J. Jager, C. Zoccali, G. Tripepi, F. W. Dekker and M. van Diepen. 'Conducting correlation analysis: important limitations and pitfalls'. In: *Clinical Kidney Journal* 14.11 (May 2021), pages 2332–2337 (cited on page 108).

[80] H. Joe. 'Relative entropy measures of multivariate dependence'. In: *Journal of the American Statistical Association* 84.405 (1989), pages 157–164 (cited on page 109).

[81] P. V. Johnsen, I. Strümke, S. Riemer-Sørensen, A. T. DeWan and M. Langaas. *Inferring feature importance with uncertainties in high-dimensional data*. 2021. arXiv: 2109.00855 [cs.LG] (cited on pages 184, 191).

[82] T. E. Joiner Jr, K. A. Van Orden, T. K. Witte and M. D. Rudd. *The interpersonal theory of suicide: Guidance for working with suicidal clients*. American Psychological Association, 2009 (cited on page 2).

[83] S. Khodadadian, M. Nafea, A. Ghassami and N. Kiyavash. *Information theoretic measures for fairness-aware feature selection*. 2021 (cited on page 184).

[84] R. Kilian, T. Becker, K. Krüger, S. Schmid and K. Frasch. 'Health behavior in psychiatric in-patients compared with a German general population sample'. In: *Acta Psychiatrica Scandinavica* 114.4 (Oct. 2006), pages 242–248 (cited on page 83).

[85] E. Kim, S.-E. Cho, K.-S. Na, H.-Y. Jung, K.-J. Lee, S.-J. Cho and D.-G. Han. 'Blue Monday is real for suicide: A case–control study of 188,601 suicides'. In: *Suicide and Life-Threatening Behavior* 49.2 (2019), pages 393–400 (cited on pages 26, 29).

[86] G. Kimeldorf and A. R. Sampson. 'Monotone dependence'. In: *The Annals of Statistics* 6.4 (1978), pages 895–903 (cited on page 112).

[87] K. Kira and L. A. Rendell. 'A practical approach to feature selection'. In: *Proceedings of the Ninth International Workshop on Machine Learning*. ML92. Aberdeen, Scotland, United Kingdom: Morgan Kaufmann Publishers Inc., 1992, pages 249–256 (cited on pages 152, 189).

[88] O. J. Kirtley, K. van Mens, M. Hoogendoorn, N. Kapur and D. de Beurs. 'Translating promise into practice: a review of machine learning in suicide research and prevention'. In: *The Lancet Psychiatry* 9.3 (Mar. 2022), pages 243–252 (cited on pages 4, 47).

[89] W. H. Kruskal. 'Ordinal measures of association'. In: *Journal of the American Statistical Association* 53.284 (1958), pages 814–861 (cited on pages 101, 106, 153).

[90] M. Kuhn. *caret: Classification and Regression Training*. 2022 (cited on pages 172, 174).

[91] M. B. Kursa and W. R. Rudnicki. 'Feature selection with the Boruta package'. In: *Journal of Statistical Software* 36 (11 Sept. 2010), pages 1–13 (cited on page 189).

[92] A. D. LaMontagne, L. S. Too, L. Punnett and A. J. Milner. 'Changes in job security and mental health: An analysis of 14 annual waves of an Australian working-population panel survey'. In: *American Journal of Epidemiology* 190.2 (Feb. 2021), pages 207–215 (cited on page 59).

[93] H. O. Lancaster. 'Correlation and complete dependence of random variables'. In: *The Annals of Mathematical Statistics* 34.4 (1963), pages 1315–1321 (cited on page 104).

[94] C.-k. Law and D. De Leo. 'Seasonal differences in the day-of-the-week pattern of suicide in Queensland, Australia'. In: *International journal of environmental research and public health* 10.7 (2013), pages 2825–2833 (cited on pages 26, 29).

[95] E. Lexne, L. Brudin, I. Marteinsdottir, J. J. Strain and P.-O. Nylander. 'Psychiatric symptoms among patients with acute abdominal pain'. In: *Scandinavian Journal of Gastroenterology* 55.7 (July 2020), pages 769–776 (cited on page 83).

[96] X. Li, Y. Wang, S. Basu, K. Kumbier and B. Yu. 'A debiased MDI feature importance measure for random forests'. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cited on page 184).

[97] S. Lipovetsky and M. Conklin. 'Analysis of regression in game theory approach'. In: *Applied Stochastic Models in Business and Industry* 17.4 (2001), pages 319–330 (cited on page 191).

[98] R. T. Liu, R. F. L. Walsh and A. E. Sheehan. 'Prebiotics and probiotics for depression and anxiety: A systematic review and meta-analysis of controlled clinical trials'. In: *Neuroscience and biobehavioral reviews* 102 (July 2019), pages 13–23 (cited on page 83).

[99] *Live life: an implementation guide for suicide prevention in countries*. Genève, Switzerland: World Health Organization, June 2021 (cited on page 3).

[100] Y. Lu, Y. Fan, J. Lv and W. Stafford Noble. 'DeepPINK: reproducible feature selection in deep neural networks'. In: *Advances in neural information processing systems* 31 (2018) (cited on page 184).

[101] S. M. Lundberg, G. G. Erion and S.-I. Lee. 'Consistent individualized feature attribution for tree ensembles'. In: *arXiv preprint arXiv:1802.03888* (2018) (cited on pages 172, 184).

[102] S. M. Lundberg et al. 'Explainable machine-learning predictions for the prevention of hypoxaemia during surgery'. In: *Nature Biomedical Engineering* 2.10 (Oct. 2018), pages 749–760 (cited on page 152).

[103] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee. 'From local explanations to global understanding with explainable AI for trees'. In: *Nature Machine Intelligence* 2.1 (Jan. 2020), pages 56–67 (cited on page 184).

[104] S. M. Lundberg and S.-I. Lee. 'A unified approach to interpreting model predictions'. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pages 476–4777 (cited on pages 154, 172, 174, 194).

[105] P. McPherson, S. Sall, A. Santos, W. Thompson and D. S. Dwyer. 'Catalytic reaction model of suicide'. In: *Frontiers in Psychiatry* 13 (2022) (cited on page 2).

[106] L. Merrick and A. Taly. 'The explanation game: Explaining machine learning models using Shapley values'. In: *Machine Learning and Knowledge Extraction*. Edited by A. Holzinger, P. Kieseberg, A. M. Tjoa and E. Weippl. Cham: Springer International Publishing, 2020, pages 17–38 (cited on page 184).

[107] P. E. Meyer. *infotheo: Information-theoretic measures*. 2022. URL: https://CRAN.R-project.org/package=infotheo (cited on pages 172, 174).

[108] I. Miller. 'The gut-brain axis: historical reflections'. In: *Microbial Ecology in Health and Disease* 29.1 (2018), page 1542921 (cited on page 83).

[109] Y. Molero, A. Cipriani, H. Larsson, P. Lichtenstein, B. M. D'Onofrio and S. Fazel. 'Associations between statin use and suicidality, depression, anxiety, and seizures: a Swedish total-population cohort study'. In: *The Lancet. Psychiatry* 7.11 (Nov. 2020), pages 982–990 (cited on page 74).

[110] C. Molnar, G. König, B. Bischl and G. Casalicchio. *Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach*. 2021. arXiv: 2006.04628 [stat.ML] (cited on page 184).

[111] T. F. Móri and G. J. Székely. 'Four simple axioms of dependence measures'. In: *Metrika* 82.1 (Jan. 2019), pages 1–16 (cited on page 109).

[112] A. Mungo. *sklearn-relief*. Dec. 2017. URL: https://libraries.io/pypi/sklearn-relief (cited on pages 172, 174).

[113] R. C. O'Connor and O. J. Kirtley. 'The integrated motivational–volitional model of suicidal behaviour'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1754 (Sept. 2018), page 20170268 (cited on page 2).

[114] S. O'Neill and R. C. O'Connor. 'Suicide in Northern Ireland: epidemiology, risk factors, and prevention'. In: *The Lancet Psychiatry* 7.6 (June 2020), pages 538–546 (cited on page 74).

[115] A. B. Owen and C. Prieur. 'On Shapley value for measuring importance of dependent inputs'. In: *SIAM/ASA Journal on Uncertainty Quantification* 5.1 (2017), pages 986–1002 (cited on pages 184, 185).

[116] I. Parra-Uribe, H. Blasco-Fontecilla, G. Garcia-Parés, L. Martínez-Naval, O. Valero-Coppin, A. Cebrià-Meca, M. A. Oquendo and D. Palao-Vidal. 'Risk of re-attempts and suicide death after a suicide attempt: A survival analysis'. In: *BMC Psychiatry* 17.1 (Dec. 2017), page 163 (cited on pages 33, 46).

[117] F. Pedregosa et al. 'Scikit-learn: Machine learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pages 2825–2830 (cited on pages 172, 174).

[118] J. A. Phillips and K. Hempstead. 'Differences in U.S. suicide rates by educational attainment, 2000–2014'. In: *American Journal of Preventive Medicine* 53.4 (Oct. 2017), e123–e130 (cited on page 42).

[119] N. Pilnenskiy. *ITMO-FS*. Aug. 2020. URL: https://pypi.org/project/ITMO-FS/ (cited on pages 172, 174).

[120] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. 3rd edition. USA: Cambridge University Press, 2007 (cited on pages 105, 108, 112).

[121] J. Pries, G. Berkelmans, R. D. van der Mei and S. Bhulai. 'The Berkelmans-Pries feature importance method: A generic measure of informativeness of features.' Submitted for publication. 2023 (cited on pages 11, 151).

[122] Q. Puzo, P. Qin and L. Mehlum. 'Long-term trends of suicide by choice of method in Norway: a joinpoint regression analysis of data from 1969 to 2012'. In: *BMC Public Health* 16.1 (2016), pages 1–9 (cited on pages 28, 29).

[123] *Python implementation of Fisher score computing attribute importance*. https://www.codestudyblog.com/cs2112pyc/1223230432.html (cited on pages 172, 174).

[124] S. E. Quirk, L. J. Williams, A. O'Neil, J. A. Pasco, F. N. Jacka, S. Housden, M. Berk and S. L. Brennan. 'The association between diet quality, dietary patterns and depression in adults: a systematic review'. In: *BMC Psychiatry* 13.1 (June 2013), page 175 (cited on page 83).

[125] A. Reneflot, S. L. Kaspersen, L. J. Hauge and J. Kalseth. 'Use of prescription medication prior to suicide in Norway'. In: *BMC Health Services Research* 19 (Apr. 2019), page 215 (cited on page 74).

[126] A. Rényi. 'On measures of dependence'. In: *Acta Mathematica Academiae Scientiarum Hungarica* 10 (1959), pages 441–451 (cited on pages 102, 105, 109–111).

[127] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti. 'Detecting novel associations in large data sets'. In: *Science* 334.6062 (2011), pages 1518–1524 (cited on page 109).

[128] M. T. Ribeiro, S. Singh and C. Guestrin. '"Why should I trust you?": Explaining the predictions of any classifier'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pages 1135–1144 (cited on pages 172, 194).

[129] C. Richardson, K. A. Robb and R. C. O'Connor. 'A systematic review of suicidal behaviour in men: A narrative synthesis of risk factors'. In: *Social Science & Medicine* 276 (May 2021), page 113831 (cited on page 59).

[130] C. Rodway, S.-G. Tham, S. Ibrahim, P. Turnbull, K. Windfuhr, J. Shaw, N. Kapur and L. Appleby. 'Suicide in children and young people in England: a consecutive case series'. In: *The Lancet Psychiatry* 3.8 (2016), pages 751–759 (cited on page 15).

[131] M. Roser, J. Hasell, B. Herre and B. Macdonald. 'War and peace'. In: *Our World in Data* (2016) (cited on page 1).

[132] G. A. Roth et al. 'Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017'. In: *The Lancet* 392.10159 (Nov. 2018), pages 1736–1788 (cited on page 1).

[133] A. Saabas. *TreeInterpreter*. Jan. 2021 (cited on pages 172, 174).

[134] A. Sariaslan, M. Sharpe, H. Larsson, A. Wolf, P. Lichtenstein and S. Fazel. 'Psychiatric comorbidity and risk of premature mortality and suicide among those with chronic respiratory diseases, cardiovascular diseases, and diabetes in Sweden: A nationwide matched cohort study of over 1 million patients and their unaffected siblings'. In: *PLoS Medicine* 19.1 (Jan. 2022), e1003864 (cited on page 74).

[135] N. R. Saunders, M. Lebenbaum, T. A. Stukel, H. Lu, M. L. Urquia, P. Kurdyak and A. Guttmann. 'Suicide and self-harm trends in recent immigrant youth in Ontario, 1996-2012: a population-based longitudinal cohort study'. In: *BMJ Open* 7.9 (2017), e014863 (cited on pages 23, 29).

[136] C. Schmidt. 'Thinking from the gut'. In: *Nature* 518.7540 (Feb. 2015), S12–S14 (cited on page 83).

[137] D. Scott and B. Happell. 'The high prevalence of poor physical health and unhealthy lifestyle behaviours in individuals with severe mental illness'. In: *Issues in Mental Health Nursing* 32.9 (Aug. 2011), pages 589–597 (cited on page 83).

[138] A. Sedano-Capdevila, A. Porras-Segovia, H. J. Bello, E. Baca-Garcia and M. L. Barrigon. 'Use of ecological momentary assessment to study suicidal thoughts and behavior: a systematic review'. In: *Current psychiatry reports* 23.7 (2021), page 41 (cited on page 2).

[139] L. S. Shapley and A. E. Roth, editors. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press, 1988 (cited on pages 152, 154, 164).

[140] K. Shin, T. Kuboyama, T. Hashimoto and D. Shepard. 'sCWC/sLCC: highly scalable feature selection algorithms'. In: *Information* 8.4 (2017), page 159 (cited on page 184).

[141] E. Song, B. L. Nelson and J. Staum. 'Shapley effects for global sensitivity analysis: Theory and computation'. In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1 (2016), pages 1060–1083 (cited on page 184).

[142] R. Soole, K. Kõlves and D. De Leo. 'Suicide in children: a systematic review'. In: *Archives of Suicide Research* 19.3 (2015), pages 285–304 (cited on page 15).

[143] C. B. v. d. Statistiek. *Microdata: Zelf onderzoek doen*. URL: https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen (visited on 28/10/2022) (cited on pages 3, 16, 34).

[144] S. Stijven, W. Minnebo and K. Vladislavleva. 'Separating the wheat from the chaff: on feature selection and feature importance in regression random forests and symbolic regression'. In: *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation*. 2011, pages 623–630 (cited on page 184).

[145] J. L. Streeter. 'Gender differences in widowhood in the short run and long run: financial and emotional well-being'. In: *Innovation in Aging* 3(Suppl 1) (Nov. 2019), S736–S736 (cited on page 60).

[146] C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn. 'Bias in random forest variable importance measures: illustrations, sources and a solution'. In: *BMC Bioinformatics* 8 (Jan. 2007), page 25 (cited on page 190).

[147] E. Štrumbelj and I. Kononenko. 'Explaining prediction models and individual predictions with feature contributions'. In: *Knowledge and Information Systems* 41.3 (Dec. 2014), pages 647–665 (cited on page 191).

[148] M. Sugiyama and K. M. Borgwardt. 'Measuring statistical dependence via the mutual information dimension'. In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. IJCAI '13. Beijing, China: AAAI Press, 2013, pages 1692–1698 (cited on page 109).

[149] M. Sundararajan and A. Najmi. 'The many Shapley values for model explanation'. In: *Proceedings of the 37th International Conference on Machine Learning*. Edited by H. D. III and A. Singh. Volume 119. Proceedings of Machine Learning Research. PMLR, June 2020, pages 9269–9278 (cited on pages 184, 185).

[150] G. J. Székely and M. L. Rizzo. 'Brownian distance covariance'. In: *The Annals of Applied Statistics* 3.4 (2009), pages 1236–1265 (cited on pages 109, 112, 146).

[151] S. B. Teasdale, P. B. Ward, K. Samaras, J. Firth, B. Stubbs, E. Tripodi and T. L. Burrows. 'Dietary intake of people with severe mental illness: systematic review and meta-analysis'. In: *The British Journal of Psychiatry* 214.5 (May 2019), pages 251–259 (cited on page 83).

[152] *The Big Nightlife Study 2016*. URL: https://www.trimbos.nl/aanbod/webwinkel/product/af1494-het-grote-uitgaansonderzoek-2016 (visited on 14/11/2019) (cited on page 25).

[153] D. Tjøstheim, H. Otneim and B. Støve. *Statistical dependence: Beyond Pearson's ρ*. 2018. arXiv: 1809.10455 [math.ST] (cited on page 137).

[154] S. Tonekaboni, S. Joshi, K. Campbell, D. K. Duvenaud and A. Goldenberg. 'What went wrong and when? Instance-wise feature importance for time-series black-box models'. In: *Advances in Neural Information Processing Systems*. Edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin. Volume 33. Curran Associates, Inc., 2020, pages 799–809 (cited on pages 152, 157, 184).

[155] R. Uher. 'Gene-environment interactions in severe mental illness'. In: *Frontiers in Psychiatry* 5 (May 2014) (cited on page 46).

[156] D. D. van Bergen, M. Eikelenboom and P. P. van de Looij-Jansen. 'Attempted suicide of ethnic minority girls with a Caribbean and Cape Verdean background: Rates and risk factors'. In: *BMC Psychiatry* 18.1 (2018), pages 1–8 (cited on page 23).

[157] A. T. Vazsonyi, J. Mikuška and Z. Gaššová. 'Revisiting the immigrant paradox: Suicidal ideations and suicide attempts among immigrant and non-immigrant adolescents'. In: *Journal of Adolescence* 59 (2017), pages 67–78 (cited on pages 23, 29).

[158] H. X. Vinh. *QII tool*. Feb. 2019. URL: https://pypi.org/project/qii-tool/ (cited on pages 172, 174).

[159] P. Virtanen et al. 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python'. In: *Nature Methods* 17 (2020), pages 261–272 (cited on pages 172, 174).

[160] D. C. Voaklander, B. H. Rowe, D. M. Dryden, J. Pahal, P. Saar and K. D. Kelly. 'Medical illness, medication use and suicide in seniors: a population-based case-control study'. In: *Journal of Epidemiology and Community Health* 62.2 (Feb. 2008), pages 138–146 (cited on page 74).

[161] G. Walsh, G. Sara, C. Ryan and M. Large. 'Meta-analysis of suicide rates among psychiatric in-patients'. In: *Acta Psychiatrica Scandinavica* 131.3 (2015), pages 174–184 (cited on page 23).

[162] S. Watanabe. 'Information theoretical analysis of multivariate correlation'. In: *IBM Journal of Research and Development* 4.1 (1960), pages 66–82 (cited on page 112).

[163] B. D. Williamson and J. Feng. 'Efficient nonparametric statistical inference on population feature importance using Shapley values'. In: *Proceedings of Machine Learning Research* 119 (July 2020), pages 10282–10291 (cited on page 184).

[164] E. Winter. 'The shapley value'. In: *Handbook of Game Theory with Economic Applications*. Edited by R. Aumann and S. Hart. 1st edition. Volume 3. Elsevier, 2002. Chapter 53, pages 2025–2054 (cited on page 157).

[165] World Health Organization. *Preventing suicide: a global imperative*. Geneva: World Health Organization, 2014 (cited on pages 1, 45).

[166] J. Yang, G. He, S. Chen, Z. Pan, J. Zhang, Y. Li and J. Lyu. 'Incidence and risk factors for suicide death in male patients with genital-system cancer in the United States'. In: *European Journal of Surgical Oncology* 45.10 (Oct. 2019), pages 1969–1976 (cited on page 59).

[167] P. S. Yip, E. Caine, S. Yousuf, S.-S. Chang, K. C.-C. Wu and Y.-Y. Chen. 'Means restriction for suicide prevention'. In: *The Lancet* 379.9834 (2012), pages 2393–2399 (cited on page 3).

[168] G. Zalsman, K. Hawton, D. Wasserman, K. van Heeringen, E. Arensman, M. Sarchiapone, V. Carli, C. Höschl, R. Barzilay, J. Balazs et al. 'Suicide prevention strategies revisited: 10-year systematic review'. In: *The Lancet Psychiatry* 3.7 (2016), pages 646–659 (cited on page 3).

[169] R. L. Zelkowitz, T. Jiang, E. Horváth-Puhó, A. E. Street, T. L. Lash, H. T. Sørensen, A. J. Rosellini and J. L. Gradus. 'Predictors of nonfatal suicide attempts within 30 days of discharge from psychiatric hospitalization: Sex-specific models developed using population-based registries'. In: *Journal of Affective Disorders* 306 (June 2022), pages 260–268 (cited on page 74).

[170] L. Zheng et al. 'Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records'. In: *Translational Psychiatry* 10.1 (Feb. 2020), pages 1–10 (cited on pages 42, 43, 46).

[171] Z. Zhou and G. Hooker. 'Unbiased measurement of feature importance in tree-based methods'. In: *ACM Transactions on Knowledge Discovery from Data* 15.2 (Jan. 2021) (cited on pages 152, 157, 184, 190).

[172] A. Zien, N. Krämer, S. Sonnenburg and G. Rätsch. 'The feature importance ranking measure'. In: *Machine Learning and Knowledge Discovery in Databases*. Edited by W. Buntine, M. Grobelnik, D. Mladenić and J. Shawe-Taylor. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pages 694–709 (cited on page 184).