University of Bremen

Faculty of Human and Health Sciences

Using decision trees to identify intersectional subgroups at risk for cancer screening non-attendance: three European case studies

Dissertation

for the academic degree of

Doctor Public Health (Dr. PH)

submitted by

Núria Pedrós Barnils

Submitted: 20th November 2024

Defended: 26th February 2025

Supervisor: Prof. Dr. Benjamin Schüz

1. Reviewer: Prof. Dr. Hajo Zeeb (University of Bremen)

2. Reviewer: Prof.in Dr. Gertraud Stadler (Charité - Universitätsmedizin Berlin)

Acknowledgements

Benjamin, thanks for the productive stupidity, the ignorance by choice, the openness to different perspectives and the constant support. You've been a great supervisor! Per, thanks for introducing me to intersectionality a decade ago, your crazy brain, the intersectionality club and the Marianne. Dani, gràcies per l'honor de ser la teva sisena filla, per ensenyar-me a investigar i introduir-me a la cooperació internacional. Thanks to my dear research group colleagues, Inga, Christopher, Chen-Chia, Urte, Himal, Victoria and Jonathan, for sharing the ups and downs throughout these three years. Inga, thanks for your endless patience and big heart –without you, we would get lost. Emily, danke, dass du mich in die Entscheidungsbäume eingeführt hast und für unsere ständigen schönen Diskussionen. Himal, thank you for the proofreading and the unga unga. Antonioi, gracias por el diamante en bruto y salchi-perri. Miriam, Mireia, Marc, Eva, Pau, Ameer i Katrin, gràcies per Eduale, i per seguir tirant endavant el nostre projecte educatiu al Líban – shukran kteer. Mire, Paula, Almi, Christian, Marcel, Emma, Afro, Giuli, Doro i Janissa la vostra amistat és la base del meu dia a dia, quina sort la meva. Mama, gràcies per donar-me la vida, ensenyar-me a mirar i a escoltar. Papa, gràcies pel teu suport incondicional i per ser el fan número u. Sarah, gràcies per l'acceptació, l'evolució i el cor - ets el meu tros de pa preferit!

Table of content

List of figures

List of tables

Abbreviations

AIHDA	Analysis of Individual Heterogeneity and Discriminatory Accuracy
AUC	Area Under the receiver operating characteristics Curve
BCS	Breast Cancer Screening
CART	Classification And Regression Tree
CHAID	Chi-square Automatic Interaction Detector
CI	Confidence Interval
CIT	Conditional Inference Tree
CNED14	Clasificación Nacional de Educación (2014)
CRC	Colorectal Cancer
CSDH	Conceptual Framework for Action on the Social Determinants of Health
DA	Discriminatory Accuracy
EC	European Commission
EHIS	European Health Interview Survey
EU	European Union
FIT	Faecal Immunochemical Test
GALI	Global Activity Limitations Indicator
gFOBT	guaiac Faecal Occult Blood Test
ICC	Intra-Class Correlation
ISCED-2011	International Standard Classification of Education (2011)
MAIHDA	Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy
OLS	Ordinary Least Squares
OR	Odds Ratio
OSP	Organised Screening Program
p	p-value
PROGRESS-Plus	Place of residence, Race, ethnicity, culture and language, Occupation, Gender/sex, Education, Socioeconomic status, Social capital and Plus
XGBoost	Extreme Gradient Boosting

Abstract

As in many relevant public health areas, attendance in cancer screening programs is stratified by social dimensions. Current additive approaches to identify at-risk groups for cancer screening non-attendance fail to capture the complexity of individuals' experiences of discrimination that lead to health inequalities. In fact, social dimensions interact, shaping experiences of discrimination in accessing cancer screening. This dissertation aims to advance the research of complex interacting social inequalities by proposing three analytical strategies. Here, data-driven decision tree methods were employed under the framework of intersectionality to identify intersectional subgroups at risk of not attending cancer screening appointments in three European case studies: breast cancer screening (BCS) in Germany, BCS in Spain and colorectal cancer (CRC) screening in Sweden. Throughout the four manuscripts of this dissertation, the following analytical strategies were explored: (i) comparison between a decision tree-based approach and an evidence-informed approach (BCS Germany), (ii) the use of decision trees for reducing intersectional complexity (CRC screening Sweden), and (iii) using decision trees as predictive tools (BCS Spain).

The findings highlight relevant intersections of individual and regional variables within decentralised health systems that significantly improve understanding cancer screening attendance inequalities. In Spain, regions played a significant role in predicting BCS attendance, reflecting the OSP implementation timeline and the economic inequalities between regions. In fact, the region of residence has acted as a risk factor for the most economically disadvantaged regions. In Germany, cohabitating with a partner was a protective factor for BCS attendance, and women residing in specific regions were at higher risk of never attending BCS than women in the same intersectional group from other regions. In Sweden, regions played a role in CRC screening attendance, so that in regions with implemented organised screening programs, only age-related inequalities were found. In contrast, in regions with opportunistic screening, inequalities in the intersections of gender, migration background and income were identified.

This dissertation meaningfully contributes to advancing the identification of intersectional subgroups at risk of cancer screening non-attendance. It does so by allowing an inductive variable selection of relevant predictors (no risk of overrepresentation and stigmatisation of specific social dimensions) and modelling non-linear and nuanced interactions between categories across intersectional subgroups. Furthermore, it contributes to the discussion of

quantitative intersectionality methods. Here, it proposes using decision trees to reduce complex intersectional matrices and improve results’ interpretability, facilitating knowledge translation with diverse stakeholders.

In a nutshell, this dissertation significantly contributes to understanding inequalities in cancer screening in three different European countries, to the methodological field of quantitative intersectionality, and it proposes new ways of asking relevant public health questions, thus providing a new methodological tool for public health practitioners.

Chapter 1. Theoretical and methodological positioning

1.1 Health inequalities - *raison de recherche*

This dissertation, as many others, exists solely because of the persistence of health inequalities. Here, I frame health inequalities as unfair and unavoidable differences in health outcomes caused by an unjust distribution of the underlying social determinants of health (Whitehead, 1992). These social determinants have, in turn, their own causes, such as cultural and historical systems of oppression (Rose, 1993). Often, public health researchers, especially quantitative researchers, including the author, focus their attention on the causes of health inequalities and oversee the causes of the causes as they uncomfortably lie at the edges of our domain of expertise (Braveman & Gottlieb, 2014).

Indeed, obstacles and barriers experienced by individuals are multifaceted depending on the community they belong to and their context. If we want to reach out to these individuals and improve, for example, their accessibility to healthcare and, ultimately, their health, we need to be able to understand these experienced barriers adequately (Rose, 1993).

In public health, our first step to assess health inequalities is to map them across several social dimensions within a certain population, asking questions such as: "Who is at higher risk of an outcome?". A set of social dimensions drawn from the literature is used to build an inference model (e.g. regression analysis) to answer this research question. Then, we assess the increased risk of a specific outcome by estimating the effects of a certain social dimension while controlling for the effects of other covariates. Later, if we are interested in estimating two or more social dimensions simultaneously, we add the effect sizes calculated in the multivariate regression (Bauer, 2014).

However, this approach fails to capture the individual's experiences of discrimination based on their social dimensions (Bowleg, 2012). Surely, no one is just a woman, a person with a migration background, or someone from a low social class. On the contrary, individuals have simultaneously a gender, a migration status, and a social class that affects their experiences of discrimination, and hence their health status, in unique ways. Existing frameworks and methodologies allow us to rethink how we model health inequalities at a population level, asking for a broader range of questions to, hopefully, better assess complex and heterogeneous experiences of social inequalities (Bauer, 2014).

In this dissertation, I want to advance the research on how to assess complex interacting social inequalities to better target at-risk populations and address health inequalities. For this reason, the thesis explores the use of data-driven methods, such as decision trees, under the framework of intersectionality to identify intersectional subgroups at risk of not attending cancer screening appointments in three different European countries.

Following, Chapter 1 explores the concept of intersectionality, its integration into public health research, and the most popular methods employed in quantitative intersectionality.

1.2 The theory of intersectionality

During the 20th century, Black women activists in the US often found their concerns and needs overlooked by feminist and civil rights movements (Hooks, 1952). To illustrate, the first wave of feminism in the US fought for 70 years to secure women's suffrage. Although the 19th Amendment was ratified in August 1920, granting women the right to vote, this initial victory was limited to white women. Black people, including women, remained disenfranchised due to Jim Crow laws, which persisted until 1965 (Fremon, 2000). The concept of Black women simultaneously fighting multiple systems of oppression emerged within Black feminist circles in the 1970s. This notion was explicitly articulated by the Combahee River Collective in their 1977 statement (Collective, 1977), and by Angela Davis in *Women, Race & Class* in 1981 (Davis, 1983).

In the academic literature, the concept of intersectionality was first coined into a theory by law scholar Kimberlé Crenshaw in 1989, exemplified in the case *DeGraffenreid v General Motors* (Crenshaw, 1989). In this law case, a group of black women sued the company General Motors for being dismissed based on gender discrimination and racial discrimination. However, the plaintiffs lost the case as the judge ruled that General Motors was not perpetuating gender discrimination (as white women were not being dismissed) nor racial discrimination (as black men were not being dismissed). Here, Kimberlé Crenshaw argued that since there was no available frame to think, see and judge both systems of oppression simultaneously (i.e. sexism and racism), discrimination suffered by DeGraffenreid and colleagues could not be acknowledged.

The conceptualisation that the intersection of several systems of oppression leads to unique and complex experiences of inequalities gained popularity across different disciplines and social dimensions. Fields such as sociology (Collins & Guo, 2021), politics (Brown et al., 2021), and public health (Bowleg, 2012) have embraced the theory, expanding the assessed systems of oppression (e.g. ableism (Frederick & Shifrer, 2018), classism (Harris & Bartlow, 2015), heterosexism (Yep, 2002)) and, therefore, the affected social dimensions (e.g. disability, social class, sexual orientation).

1.2.1 Intersectionality in public health

While initially labelled a theory, intersectionality is utilised as an analytical framework in the field of public health. Intersectionality does not entail core elements or variables that can be operationalised or empirically tested such as in traditional social scientific theories. Instead, it is adopted as an analytical lens or paradigm to conceptualise how multiple social dimensions intersect at the micro level, reflecting interlocking systems of oppression or privilege at the macro level (Bowleg, 2012).

The operationalisation of this framework in public health has taken place through qualitative (Bowleg, 2008), quantitative (Bauer, 2014) and mixed methods research (Grace, 2014). In this dissertation, an intersectional framework is adopted for conducting quantitative research.

1.2.2 Quantitative intersectionality in public health

In quantitative intersectionality, several concerns arise regarding the appropriate use of an intersectional lens throughout the research process (Bowleg, 2012; Creswell & Creswell, 2017). Figure 1 exemplifies some of these relevant considerations for each stage of the research cycle.

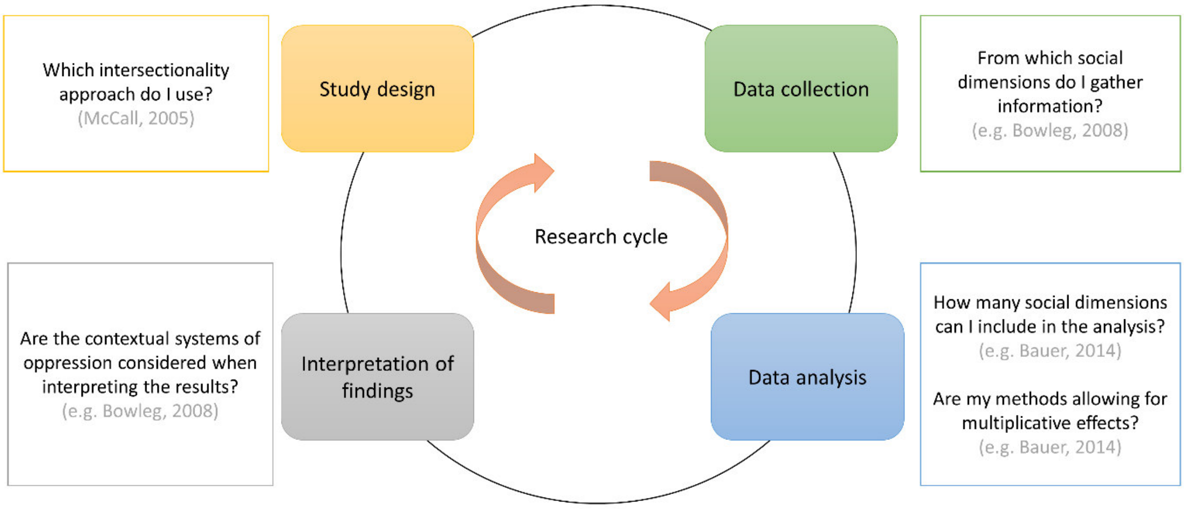

Which intersectionality approach do I use?
(McCall, 2005)
Study design
Data collection
From which social dimensions do I gather information?
(e.g. Bowleg, 2008)
Research cycle
Are the contextual systems of oppression considered when interpreting the results?
(e.g. Bowleg, 2008)
Interpretation of findings
Data analysis
How many social dimensions can I include in the analysis?
(e.g. Bauer, 2014)
Are my methods allowing for multiplicative effects?
(e.g. Bauer, 2014)

Figure 1. An example of relevant intersectional questions at each stage of the research cycle (own figure)

During the *study design* phase, a key aspect to consider is deciding which intersectional approach to employ. McCall (2005) identified three approaches that significantly influence the research focus and methodology (McCall, 2005):

- Intracategorical approach: explore experiences of individuals within a social category
- Intercategorical approach: explore experiences of individuals between social categories
- Anticategorical approach: criticises the integrity of categorical distinctions, calling for a methodology that deconstructs analytical categories

Quantitative intersectionality health research often employs an intercategorical approach, focusing on inequalities between different social groups. However, this approach faces several challenges, such as *collecting data* (i.e. social dimensions) that represent individuals' experiences of discrimination. In the present dissertation, the social dimensions gathered by the employed secondary data (i.e. the European Health Interview Survey) were considered representative enough of individuals' experiences of discrimination. In Chapter 4, these social dimensions are enumerated.

Then, *data analysis* must align with the principles of intersectionality, particularly the tenet on the interconnected and mutually influencing nature of social dimensions (McCall, 2005). This means that an individual's experiences, such as health inequalities, cannot be fully understood by examining only one social dimension (like gender) or by simply adding up different social positions (such as ethnicity and gender). It is important to account for the effects of simultaneously being at the intersection of several social positions (Bauer, 2014).

The translation of this intersectional jargon into statistical language requires that, for example, when examining the risk of not attending breast cancer screening (BCS) among women of different ages and social classes simultaneously, we cannot simply add the risks associated with each category (i.e. additive approach). That is, when an additive approach is taken, the risk for each variable (or categories within a variable) is calculated in a regression while holding the other included variables constant. Then, estimates of these risks are added to calculate the overall risk of a specific social group (i.e. young women from disadvantaged social class) (Bauer, 2014). Conceptually, this oversimplifies and ignores the interplay between social positions contributing to health and social inequalities (Crenshaw, 1991; Cuadraz & Uttal, 1999). Methodologically, it does not allow for categories to interact, hence missing the relevant effects of those interactions (Bauer, 2014).

To account for these interactions and capture "the unique experiences of individuals" (Crenshaw, 1989), a multiplicative approach is needed. Here, statistical methods need to capture the effect of the interactions. The initial methods implemented in quantitative intersectionality that aim to account for the intersectional effects were: regression with interactions and regression with cross-classified categories. Following the previous example, to assess the risk for young women from a disadvantaged social class of not attending BCS, a regression with interaction terms would estimate the overall risk by adding the risk of being young plus the risk of being socially disadvantaged plus the risk of being young and socially disadvantaged simultaneously. A regression with cross-classification would regress an exposure variable created from the combination of all categories of both variables of interest (e.g. young-low, young-high, old-low, old-high), setting one of the intersectional subgroups as the reference group (Block Jr et al., 2023).

Over the past two decades, there have been extensive methodological discussions on how to describe and quantify inequalities among intersectional subgroups (i.e. descriptive intersectionality) (Bauer, 2014). Bauer et al.'s (2021) systematic review summarises various statistical methods used in quantitative intersectionality (Bauer et al., 2021). Every proposed method treats the complexity of intersecting social dimensions differently. Consequently, the intersectional focus of the applied statistical methods varies.

Table 1, following Bauer et al.'s (2021) systematic review, displays the most common statistical methods used in quantitative intersectionality grouped by their intersectional focus: intersectional mapping of inequalities, isolating the "intersectional effect" within inequalities, addressing the "tyranny of the averages" or performing data-driven simplification.

Table 1. Statistical methods used in quantitative intersectionality by intersectional-focused function

Intersectional focus	**Statistical methods**
Intersectional mapping	Regression with cross-classification
Isolating the "intersectional effect"	Regression with interactions
"Tyranny of the averages"	Analysis of Individual Heterogeneity and Discriminatory Accuracy (AIHDA) Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA)
Data-driven simplification	Latent variable analysis Cluster analysis Decision trees

Regression with cross-classification is the simplest form of *intersectionally mapping* inequalities' estimations. First, an exposure variable is created by combining all categories of pre-selected variables. Then, a regression is performed, taking one of the intersectional subgroups of the cross-classified variable as a reference group and estimating the effect sizes for every other intersectional subgroup (Pedros Barnils et al., 2020).

Regression with interaction terms allows *isolating the "intersectional effect"* of those individuals who are at the intersection of two disadvantaged categories. By isolating the effect of, for example, being in two disadvantaged categories through the interaction term, it is possible to quantify the additional risk of being in two disadvantaged categories simultaneously. The interaction term is easy to interpret when a two-way interaction is tested. However, interpreting the interaction terms becomes more complicated when more variables are tested (i.e. three or four-way interactions) (Block Jr et al., 2023).

Merlo et al. (2017) brought to the quantitative intersectionality community the concern of the *tyranny of the averages* (Merlo et al., 2017), previously described in epidemiology (Tabery, 2011) and medical sciences (Pepe et al., 2004). The *tyranny of the averages* defends that standard measures of association fail to capture within-group variability, and accordingly, statistical estimates should be reported together with measures of individual variance that represent intra-group heterogeneity. They postulate this argument through, on the one hand, questioning the insufficiency of the "statistical significance" criterion to consider two groups to be meaningfully different. Here, they argue that large sample sizes can lead to statistically significant differences even when the actual differences are small and insignificant. On the other hand, they posit that only by computing an estimation coefficient of a group, the intragroup variance is overlooked, that is, the individual heterogeneity within the group. In turn, the attributed estimation to an individual (i.e. the group average estimation) might not be accurate (i.e. we do not know how close the group's average estimation is to the individual estimation). A very illustrative example is available elsewhere ((Merlo et al., 2017), page 693). Put simply, it is necessary to find ways to maximise between-group heterogeneity and within-group homogeneity with regard to the outcome. One approach is to compute a discriminatory accuracy (DA) measure, which quantifies the ability of a group to accurately classify individuals based on outcomes (e.g. for binary outcomes to classify cases and non-cases) (Merlo et al., 2017).

Two statistical methodologies used in quantitative intersectionality are based on the above-described notion: the analysis of individual heterogeneity and discriminatory accuracy (AIHDA) (Merlo et al., 2023) and the multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) (Merlo, 2018). The main difference between both is the use of traditional regressions in AIHDA (i.e. regression with cross-classification) and the use of multilevel regression analyses in MAIHDA (i.e. level 1 individual variables, level 2 intersectional subgroups) as well as the different measures for DA. AIHDA employs the area under the receiver operating characteristics curve (AUC) to measure the DA of the model. The AUC is obtained by plotting the true-positive fraction of the model (i.e. sensitivity) against the false-positive fraction of the model (i.e. 1-specificity) for various thresholds of the predicted probability of the outcome (Bamber, 1975). The AUC indicates how well the groups of the model discriminate between two or more classes, and its value ranges from 0.5 (no DA) to 1 (perfect DA). MAIHDA adopts multilevel discrimination measures, such as the Intra-Class Correlation (ICC), to assess the proportion of total variance attributable between groups rather than within groups. A high ICC indicates that individuals within a group are more similar to each other than to individuals in other groups. On the contrary, low ICC indicates high within-group variance and, thus, low DA regarding the outcome from the groups (Goldstein et al., 2002).

In building any of the models presented above, a set of exposure variables must be selected, which requires a decision on how many and which social dimensions to examine. This choice is usually guided by existing evidence, such as published literature. However, due to the relatively recent emergence of intersectional approaches in public health, there often is a lack of clear theoretical or empirical guidance for selecting the relevant specific social dimensions for analysis while excluding others (Bauer et al., 2021). To address this limitation (which accumulates into a garbage-in-garbage-out- problem), a recent trend in the field is using data-driven methods to select relevant variables and then identify relevant intersectional subgroups (Bauer et al., 2021; Bauer et al., 2022).

Furthermore, when many social dimensions are included in the analysis (i.e. high-dimensional interactions), methods such as regression with cross-classification encounter methodological challenges. Here, large matrices of intersectional groups with small cell sizes limit the statistical power of the analysis. In addition, interpreting and assessing numerous intersectional groups can be difficult, hindering the development of clear guidance for policymakers and practitioners (Mahendran et al., 2022). Potential approaches to address this complexity are methods that perform *data-driven simplification*, that is*,* data-driven selection of relevant variables and simplification of the resulting intersectional subgroups. These methods include latent variable analyses, clustering methods, and decision trees.

Latent variable analysis and clustering methods, despite their differences in purpose and technique, offer several common advantages for *data-driven simplification* in quantitative intersectionality. Bauer et al.'s (2022) systematic review stated that these methods effectively identify overlapping and co-occurring social positions or experiences from a range set of social dimensions, particularly when variables are measured on a continuous scale (avoiding the need for arbitrary cut-offs) (Bauer et al., 2022). However, unlike decision trees, these identified subgroups do not directly lead to clearly defined categories based on social dimensions (Battista et al., 2023; Bauer et al., 2022).

1.3 Decision trees for intersectional subgroup identification

In the present dissertation, decision trees are employed to identify intersectional subgroups of people at higher risk of not attending cancer screening appointments. Therefore, the present section introduces the decision tree method and its use in public health and quantitative intersectionality research.

Decision trees are inductive, rule-based methods that exploratively classify or predict outcomes. They create a hierarchical structure of binary (occasionally ternary or quaternary) partitions from continuous and/or categorical predictors. These partitions are intended to minimise the difference between the predicted and observed values in consequent subsets of data (Breiman et al., 2017).

By incorporating social dimensions as predictors, decision trees identify the optimal aggregation of these predictors to classify or predict an outcome. This process includes selecting certain predictors for building the decision tree and grouping their categories based on the increased homogeneity of the resulting groups (Venkatasubramaniam et al., 2017). The decision tree results are then nuanced intersectional subgroups, making decision trees highly relevant for public health (Battista et al., 2023) and quantitative intersectionality (Mahendran et al., 2022).

Decision trees have been applied in public health as explorative tools to select relevant variables and identify homogeneous subgroups for binary outcomes (Battista et al., 2023; Freitas et al., 2012; Mena, Bolte, & Advance Gender study, 2021), as well as integrated into quantitative intersectionality methodological discussions (Bauer et al., 2021; Mahendran et al., 2022). In addition to its exploratory use, decision trees have also been employed as predictive tools. Methodologically, using decision trees as predictive tools means that a subset of the data (training set) is used for building the decision tree and then tested on unseen data (test set) for prediction performance. This allows for the generalisation of the predictions to unseen data, that is, for predicting future events (Banerjee et al., 2019).

Several authors have used decision trees as prediction tools in public health (Esmaily et al., 2018; Fu et al., 2022). Although several algorithms have shown better predictive accuracy than decision trees, such as ensemble algorithms and neural networks (Sahoo et al., 2020), decision trees provide a good balance between accuracy and explainability. That is, the predictive performance of their models is reasonable while the underlying structure remains comprehensible (Loyola-Gonzalez, 2019). This made decision trees the most popular white-box models nowadays. White-box models are models with good explainability and interpretability, that is, models that non-experts can easily understand without needing additional models or other features.

In this dissertation, decision trees are employed as explorative and predictive tools. Whereas the dissertation does not aim to discuss the advantages and disadvantages of both approaches (see Bowe (2022) for a comprehensive overview of the field of early prevention (44)), a discussion on decision trees in both paradigms and their implications for public health can be found in Chapter 6.

1.3.1 Gaps in the literature in decision trees for intersectional subgroup identification

In the field of public health, decision trees have been employed for identifying at-risk populations, and their use has been highly encouraged in health promotion and prevention for targeted intervention development (Delgado-Gallegos et al., 2023; Eagle et al., 2022). Likewise, following a comprehensive comparison with traditional regression models, they have been identified as essential methodologies for the population health researchers' repertoire (Battista et al., 2023).

Nevertheless, only a few studies have employed decision trees for identifying intersectional subgroups at higher risk of never attending cancer screening appointments. It is particularly relevant to ascertain intersectional subgroups of individuals never attending cancer screening appointments to develop more targeted interventions for promoting screening attendance. In the Portuguese context, Freitas et al. (2012) examined individual and environmental factors that predict BCS attendance among targeted women (Freitas et al., 2012), Yet, they did not adopt an intersectional perspective in their study.

Discussions on using decision trees under the framework of intersectionality have been rare (Bauer et al., 2021). Cairney (2014) compared the performance of logistic regression with a decision tree for mental health service use in Canada, highlighting the capability of the decision tree to identify complex and unsuspected interactions in line with the framework of intersectionality (Cairney et al., 2014). Mena, Bolte, & AdvanceGender Study (2021) discussed the use of decision trees for monitoring frequent mental distress embedded within distinct sex/gender theoretical concepts in health research that include intersectionality-informed variables as well as solution-linked variables (i.e. here, uniquely taking an intersectional anti-categorical approach) (Mena, Bolte, & AdvanceGender Study, 2021).

This dissertation uses decision trees for intersectional subgroup identification of people at risk of not attending cancer screening appointments. Besides, applications of decision trees are explored through three different methodological approaches. These are explained and expanded in Chapter 4 and Chapter 5.

1.4 Cumulative Dissertation Outline

This dissertation provides an overview of the research context of the individual studies presented in Appendix A. **Chapter one** provides an overview of the theoretical and methodological positionality of the dissertation to place the research contribution. **Chapter two** provides the empirical foundation of the dissertation with an overview of cancer screening programs in the European Union, a justification of the selected case studies, and a summary of inequalities in attending cancer screening in the three case studies. Finally, existing knowledge gaps addressed in this dissertation are highlighted. **Chapter three** describes the dissertation's overarching research aim and objectives and elucidates how the different manuscripts contribute to the research objectives. **Chapter four** synthesises the data and methods applied in the four manuscripts of this dissertation. **Chapter five** presents the results of the different manuscripts. **Chapter six** first describes the main empirical and methodological findings of the dissertation. Second, it expands on the multifaceted implications of the results of the dissertation for quantitative intersectionality, public health research and practice and policy. Third, it outlines the strengths and limitations of the thesis. **Chapter seven** drafts the future research outlook. Finally, **Chapter eight** concludes the dissertation.

Note A: Throughout the PhD, I collaborated closely with my co-authors, who provided valuable guidance and support. Hence, sentences related to our shared work are framed from a "we" perspective, whereas sections specific to this dissertation are framed from an "I" perspective.

Note B: Along the thesis, the word "variable" is used in an estimative inference context, and "predictor" is used in a predictive context. In chapters where the use of decision trees is discussed for both paradigms, such as introducing decision trees in Chapter 1 and discussing and formulating implications of this method in Chapter 6, both words are used interchangeably.

Note C – *Manuscripts I, II,* and *IV* were carried out at the Prevention and Health Promotion research group from the Institute for Public Health and Nursing Research at the University of Bremen (Germany). *Manuscript III* was developed at the Department of Global Health and Epidemiology at the University of Umeå (Sweden).

This implied a different conceptualisation and language when discussing inequalities/inequities in attending cancer screening services. At the research group Prevention and Health Promotion from the Institute for Public Health and Nursing Research at the University of Bremen, it was agreed to use the wording "inequalities", whereas at the Department of Global Health and Epidemiology at the University of Umeå is more often used the concept "inequities". The second encourages using this concept given the ethical implications of social justice implicit in equity (Braveman & Gruskin, 2003). The first encourages equality, given the broad use of this term in health research. Throughout the dissertation, I have used the concept of inequalities for major consistency with the publications, except when presenting the research aim of the paper developed in Sweden. Nevertheless, my intentionality in the word inequalities is the referred to inequities by Whitehead (1992), that is, health differences that "are not only unnecessary and avoidable but, in addition, are considered unfair and unjust" (Whitehead, 1992).

Another difference in conceptualisation between both schools of public health is the framework to place and think about inequalities. On the one hand, at the Prevention and Health Promotion research group from the Institute for Public Health and Nursing Research at the University of Bremen, PROGRESS-Plus characteristics were used since they allow for a broad empirical exploration of inequalities at various social layers (Oliver et al., 2008) On the other hand, at the Department of Global Health and Epidemiology at the University of Umeå, the use of the Conceptual Framework for Action on the Social Determinants of Health was encouraged, in specific, the use of the structural determinants of health as they reflect historically downstream expressions of injustices (Solar & Irwin, 2010).

Note D – Although I have used the wording "machine learning" for predictive algorithms in m*anuscript IV,* in this dissertation, I do not differentiate decision trees based on a research field. Instead, I distinguished both applications of decision trees based on their goal: exploration or prediction. This decision is based on the lack of a clear-cut between both research fields (LMU, 2022). In fact, there have been decades of discussions on this matter

(Breiman, 2001b), and since I did not intend to position myself in this discussion, I named decision trees' applications based on their aim.

Chapter 2. Empirical background

Chapter 2 presents the empirical background of the dissertation, starting with a general introduction to cancer screening programs in the European Union (EU), a discussion of why intersectional approaches are relevant within this context and follow by the specific cancer screenings studied in the three European cases and outlining the specific literature gaps that the studies address.

2.1 Cancer screening in the European Union

In the EU, cancer was the second leading cause of death in 2021, accounting for a significant 21.6% of all deaths (1,144,420 individuals) (Eurostat, 2024). Studies indicate that nearly 40% of European cancer cases could be prevented if effective primary prevention programs were in place (Behrens et al., 2018; Brown et al., 2018; Soerjomataram et al., 2018). Moreover, cancer mortality could be reduced through early detection approaches such as systematic cancer screenings (Wild et al., 2019).

To reduce cancer mortality rates, the European Commission (EC) has prioritised nationally organised cancer screening programs (Wild et al., 2019). In 2003, they recommended that all Member States implement organised screening programs (OSP) for breast, cervical, and colorectal cancers (Armaroli et al., 2015). OSP allow for an inclusive approach towards the early detection of malignant cells among at-risk populations (Miles et al., 2004). This initiative was further strengthened by 2021 *Europe's Beating Cancer Plan*, which aims to increase screening rates for these cancers to 90% of eligible individuals by 2025 (European Commission, 2021).

The initiation of OSP implementation nationally has widely varied across Member States. Briefly, OSP for breast cancer (BC) were first implemented in Sweden in 1986 and last implemented in Slovakia in 2019. By 2023, Greece and Bulgaria had not yet fully implemented an OSP for breast cancer detection (Cardoso et al., 2023; OECD, 2023a). For cervical cancer, national OSPs were first implemented in Finland in 1963 and last in Spain in 2019. By 2023, Austria, Bulgaria, Cyprus, and Luxembourg had not yet fully implemented OSP for cervical cancer (OECD, 2023b). Finally, the national OSP for colorectal cancer (CRC) was first implemented in Germany in 1974 and last in Sweden in 2022. By 2023, Bulgaria, Greece, Romania and the Slovak Republic had not yet implemented an OSP for colorectal cancer detection (OECD, 2023c).

Attendance rates for health screenings vary significantly across countries and sociodemographic factors. In 2019, only 2.3% of targeted women in Sweden reported never having had a breast examination by X-ray, compared to 71.6% in Romania. Cervical smear test participation varied slightly less, as in Romania, 47.4% of women reported never having been screened, compared to 2.6% in Czechia. CRC screening showed the highest differences, with 93.2% of people reporting never having undergone the procedure in Cyprus, compared to 17.4% in Denmark. Sociodemographic inequalities within the EU 27 further exacerbate these attendance disparities across countries. People with lower education levels, lower incomes, younger people, those living in rural areas and those out of the labour market had significantly lower screening rates for all three cancers. Women with disabilities were less likely to participate in BCS, while individuals without disabilities were less likely to participate in cervical and CRC screening programs Finally, men were slightly less likely than women to attend CRC screening (European Commission, 2024).

A crucial strategy to reduce overall and sociodemographic inequalities in cancer mortality is early and equitable risk detection through national OSP. However, universal access programs and interventions do not always have equitable effects. Research has shown that downstream interventions (Lorenc et al., 2013) and interventions that lack the promotion of informed decision-making (Durand et al., 2014) generate inequalities in the population. It is, therefore, essential to monitor the implementation of OSP in order to guarantee that no individual is excluded, enhancing access for the entire target population (Persaud et al., 2023). Adopting an intersectional perspective in the monitoring of OSP implementation, facilitates the identification of specific subgroups that may encounter overlapping obstacles to accessing cancer screening. Hence, moving beyond the one-size-fits-all paradigm (Caserta et al., 2016) and developing tailored campaigns to the needs of vulnerable groups might aid in effectively allocating resources and equitably reaching all at-risk populations.

2.2 Three European case studies

This dissertation examines inequalities in cancer screening non-attendance through three different case studies: BCS attendance in Germany, BCS in Spain and CRC screening in Sweden, Inequalities in BCS attendance and CRC screening attendance were considered pertinent given its public health relevance and intersectional interest.

BC was the most commonly diagnosed cancer among women in the EU in 2020 (European Commission, 2021), accounting for 28.7% of all female cancers, with an estimated 355,457 new cases and an age-standardized mortality rate of 142.8 per 100,000 women (ECIS, 2020).

CRC was the second most diagnosed cancer among women and third among men in the EU in 2020 (European Commission, 2021), encompassing 13% of all new cancer cases, with an estimated 361,986 new cases and 161,182 deaths (12% of all deaths due to cancer) being the second leading cause of cancer deaths (ECIR, 2024).

The early detection of both cancers is encouraged in the EU via nationally implemented OSP (European Commission, 2021). OSP targeting BC bi-annually invites women aged 50 to 69 years for BCS through a mammography. Additionally, women aged 45-49 and 70-74 are encouraged to be screened every 2-3 years and 3 years, respectively (European Commission, 2022). Similarly, OSP targeting CRC bi-annually invites individuals aged 50 to 74 for colorectal cancer screening through guaiac faecal occult blood tests (gFOBT) or faecal immunochemical tests (FIT) (Council of the European Union, 2003).

From an intersectional perspective, BCS attendance allows us to examine inequalities while holding a key social dimension constant - sex/gender - whereas CRC screening allows for all social dimensions to interact.

Three European countries were chosen as study contexts given their decentralised health systems: Germany, Spain and Sweden. Their regionalised health systems offer the possibility of exploring regional differences as an additional dimension of intersectional inequalities. Even more, assessing the impact of regional differences on access to care has been considered crucial in decentralised health systems (Ministerio de Sanidad, 2021; Rey del Castillo, 1998).

In Germany, BC was the fifth leading cause of death among women in 2022, with 18,900 deaths (Statistisches Bundesamt, 2024) and in Spain, where cancers were the second cause of death among women in 2021 (45,818 deaths), BC was the most frequent cancer-related death (6,528 deaths) (Estadística, 2024). In Sweden, BCS coverage rates are among the highest in the EU; however, the CRC screening coverage rates are far below the EU average (OECD, 2023d). In fact, in 2020, up to 5,240 people were diagnosed with CRC becoming the third most-diagnosed cancer for women and fourth for men. Furthermore, 1,903 people died of CRC, being the fourth most-mortal cancer in the country (Cancerfonden, 2023). Finally, until 2022, organised screening was implemented only in certain regions of Sweden (Blom et al., 2014), potentially highlighting the equity effects of implementing OSP within the same country.

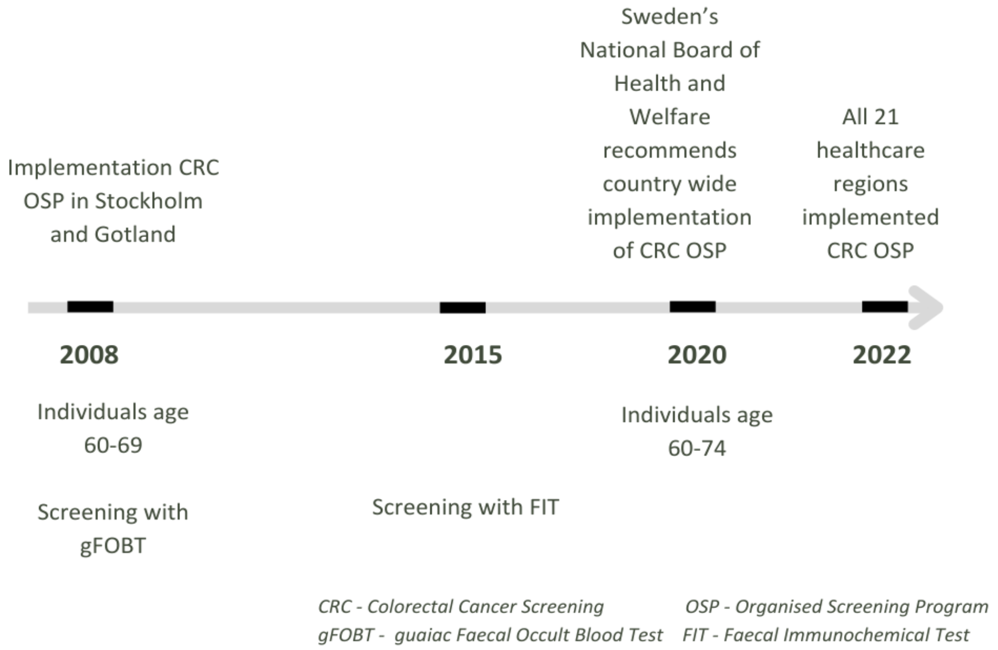
2.2.1 Decentralised health systems in the three case studies

In Germany, many health competencies are allocated at the regional (i.e. federal state) level. Regions administer healthcare policies, supervise municipal public health services, and regulate the implementation of preventive programs such as OSP (Federal Ministry of Health, 2020). The OSP for BC was initiated in 2005, and although the overall framework of the program was standardised across the country, there were differences when the different regions implemented the program. Bremen and Schleswig-Holstein were among the first to implement the program in 2005, whereas Brandenburg and some provinces of Saxony-Anhalt and Baden-Württemberg were the last in 2008 (Kooperationsgemeinschaft Mammographie, 2012). In 2009, the country reached full implementation of the screening program bi-annually, inviting women aged 50 to 69 (Kooperationsgemeinschaft Mammographie, 2014).

In Spain, due to the decentralised nature of the healthcare system, the implementation of OSP for BC and the target populations differ across the regions (i.e. Autonomous Communities). Navarre was the first region to adopt the program in 1990, followed by Catalonia, Castilla-La Mancha, Castile and León, Catalonia, Valencian Community, and Galicia in 1992. In contrast, Ceuta and Melilla did not offer the program until 2005 (Ministerio de Sanidad y Consumo, 1996). Additionally, the targeted population of the program also varies across regions: some implemented a bi-annual screening regimen for women aged 45-69, while others restricted it to women aged 50-69. Decentralisation also entails differences in resource allocation, and as a result, disparities between regions can appear. A striking one is the nearly double concentration of health professionals in certain regions (e.g. Asturias, Navarra, and Castile and León) compared to others (e.g. Ceuta and Melilla) (Ministerio de Sanidad, 2021; Rey del Castillo, 1998).

In Sweden, the healthcare system is divided into 21 healthcare regions. Whereas the national government is responsible for the legislation, the regions are responsible for financing and providing health care for their residents and administrating screening programs (Anell et al., 2012). This decentralisation led to marked differences in the implementation of OSP for CRC in the country (Figure 2). From 2008 to 2020, only the healthcare regions of Stockholm and Gotland, where 25% of the Swedish population lives, implemented OSP for CRC screening, inviting people aged 60 to 69 for gFOBT (Blom et al., 2014) and later in 2015 for FIT (Blom et al., 2019; Mousavinezhad et al., 2016). Consequently, in 2020 the country registered an overall screening coverage rate 22% lower than the EU average (OECD, 2023d), plus stark regional differences between areas with organised (84.90%) and opportunistic (25.81-37.63%)

screening (The Public Health Agency of Sweden, 2019-2020). In response to this situation, Sweden's National Board of Health and Welfare recommended in 2020 the implementation of a nationwide CRC screening program using FIT for individuals aged 60 to 74. By September 2022, all 21 healthcare regions had started implementing the program (Regionala Cancercentrum Samverkant, 2022).



Implementation CRC OSP in Stockholm and Gotland
Sweden's National Board of Health and Welfare recommends country wide implementation of CRC OSP
All 21 healthcare regions implemented CRC OSP
2008
2015
2020
2022
Individuals age 60-69
Individuals age 60-74
Screening with gFOBT
Screening with FIT
CRC - Colorectal Cancer Screening
OSP - Organised Screening Program
gFOBT - guaiac Faecal Occult Blood Test
FIT - Faecal Immunochemical Test

Figure 2. Graphical summary of the implementation of CRC OSP in Sweden

2.3 Breast cancer screening attendance in Germany and gaps in the literature

Since the implementation of OSP in Germany, attendance rates have fluctuated between 43% and 55%, according to data recorded in the screening unit centres (Kooperationsgemeinschaft Mammographie, 2023). Based on self-reported data, in 2020, 65.72% of women aged 50-69 underwent mammography in the last two years, and approximately 10.38% never attended breast cancer screening in their lifetime (Eurostat, 2019). These values are still well below the 70% EU recommendation in 2003 (24) and the latest 90% established by *Europe's Beating Cancer Plan* to be reached by 2025 (European Commission, 2021).

BCS attendance in Germany (Starker, 2017) and worldwide (Mottram et al., 2021) is unequal across eligible populations. Several factors have been linked to BCS attendance, including sociodemographic factors, attitudes, knowledge, beliefs, health behaviours, and accessibility and logistics (Ackerson & Preston, 2009; Crosby, 2018; Schueler et al., 2008). In Germany, Missinne & Bracke (2015) found a positive association between higher income and self-reported BCS attendance but no correlation with education level (Missinne & Bracke, 2015).

Similarly, Heinig et al. (2023) observed no significant link between education and BCS attendance using health insurance claims data (Heinig et al., 2023). However, Lemke et al. (2015) identified a predictive association between higher income and lower education levels and increased BCS attendance (Lemke et al., 2015)

The existing research on sociodemographic inequalities in BCS attendance in Germany is heterogeneous and inconclusive. A comprehensive review of studies assessing these inequalities, particularly since the OSP implementation, is needed to understand the factors influencing screening attendance and inform evidence-based interventions to increase uptake among eligible women. Furthermore, no study has assessed regional sociodemographic inequalities, including regional location in BCS attendance simultaneously, nor from an intersectional perspective. That is, there is no study examining women at higher risk of never attending BCS that goes beyond an additive approach of social dimensions. An empirical study utilising an intersectional approach accounting for regional and sociodemographic dimensions could enhance the understanding of complex social inequalities affecting BCS attendance.

2.4 Breast cancer screening attendance in Spain and gaps in the literature

The overall BCS attendance rate in Spain has consistently risen since the OSP implementation, with 72.6% of eligible women attending their most recent medical appointment in 2020 and 93.13% having undergone BCS at least once in their lifetime (INE, 2021). These BCS attendance rates are, however, not only regionally unevenly distributed but also socio-demographically. Several factors have been associated with higher screening rates, including being married, holding Spanish nationality, being born in Spain, and having higher levels of education and income (Martín-López et al., 2013; Serral et al., 2018; Zamorano-Leon et al., 2020). Nonetheless, no study has comprehensively examined the interplay among different sociodemographic dimensions, including regional inequalities in BCS attendance rates. Additionally, and similar to the case study in Germany, no study has assessed BCS attendance inequalities from an intersectional, non-additive, perspective. Thus, an empirical study employing an intersectional lens with various regional and sociodemographic factors would enhance the understanding of BCS attendance inequalities.

2.5 Colorectal cancer screening attendance in Sweden and gaps in the literature

Sociodemographic inequalities in CRC screening attendance in Sweden have been reported. In a six-year nationwide randomised controlled trial, Strömberg et al. (2022) found that men, individuals from low-income households, those without a university education, unmarried individuals, and those born in a Western country had lower CRC screening attendance rates (Strömberg et al., 2022a). Nevertheless, no study has either assessed patterns of inequalities in lifetime CRC screening attendance in the population aged 60-69 in regions with and without organised screening programs in Sweden or has it intended to identify intersectional subgroups of people at higher risk of never attending CRC screening. That is why an empirical study accounting for regional inequalities based on OSP implementation and sociodemographic factors will facilitate recognising the importance of implementing organised screening programs to reach at-risk individuals.

Chapter 3. Research aim and research objectives

The research aim is to explore the use of decision trees to identify subgroups at higher risk of never attending cancer screening appointments. The following specific research objectives serve the overall aim.

Objective 1: To compare a decision tree-based approach and an evidence-informed approach for identifying intersectional subgroups.

In order to address this research objective, *manuscripts I* and *II* were employed in the context of breast cancer screening attendance in Germany. *Manuscript I* is a scoping review of the published and grey literature on sociodemographic inequalities in BCS attendance in Germany since the implementation of the OSP. It examined the existence and effect sizes of sociodemographic inequalities. The results of *manuscript I* built the evidence-informed approach in *manuscript II. Manuscript II* is an empirical study that compares a decision tree-based approach and an evidence-informed approach for identifying intersectional groups of women aged 50-69 at higher risk of never attending BCS in Germany.

Objective 2: To explore the use of decision trees for reducing intersectional complexity.

In order to tackle this research objective, *manuscript III*, an empirical study contextualised in Sweden assessing inequities in CRC screening attendance, was carried out. This study aims at identifying intersectional groups of people aged 60-69 in regions with and without organised screening programs at higher risk of never attending CRC screening. It does so by utilising decision trees to reduce the complexity of a full intersectional matrix.

Objective 3: To explore decision trees as predictive tools for intersectional subgroup identification.

Manuscript IV addresses this research objective through an empirical study contextualised in Spain that aims to predict which intersectional subgroups of women aged 45/50-69 are at higher risk of never attending breast cancer screening. Here, decision trees were used as predictive tools to predict which women would never attend BCS appointments.

Manuscripts I, II, and *IV* were carried out at the Prevention and Health Promotion research group from the Institute for Public Health and Nursing Research at the University of Bremen (Germany). *Manuscript III* was developed at the Department of Global Health and Epidemiology at the University of Umeå (Sweden). The four manuscripts are available in **Appendix A,** and related scientific presentations are in **Appendix B**.

4. Data and methods

This chapter describes the data and methodology used in each paper constituting the thesis. Table 2 presents an overview of the data sources, methods employed and use of decision trees in the four manuscripts.

Table 2. Overview of the data sources and variables used in the four papers of the dissertation

	Manuscript I	**Manuscript II**	**Manuscript III**	**Manuscript IV**
Country	Germany	Germany	Sweden	Spain
Survey	Several	EHIS 2019 Germany	EHIS 2019 Sweden	EHIS 2019-2020 Spain
Research outcome	BCS life attendance and BCS last appointment attendance	BCS lifetime attendance	CRC screening lifetime attendance	BCS lifetime attendance
Aim	To identify sociodemographic inequalities in BCS attendance among women aged 50-69 years since the OSP implementation	To identify intersectional groups of women aged 50-69 at higher risk of never attending BCS in Germany	To identify intersectional groups at higher risk of never attending CRC screening in Sweden	To identify intersectional groups of women aged 45-69 at higher predicted prevalence of never attending BCS in Spain
Framework of inequalities	PROGRESS-Plus framework	PROGRESS-Plus framework	Conceptual Framework for Action on the Social Determinants of Health	PROGRESS-Plus framework
Analytical strategy	Scoping review - five-step methodological framework proposed by Arksey & O'Malley in 2005	Decision tree-based intersectional groups + regression (compared to evidence-based intersectional groups + regression)	Decision tree for reducing full intersectional matrix + regression	Predictive decision tree + ensemble algorithm
Use of decision tree	None	Variable selection and intersectional subgroup identification	Intersectional matrix complexity reduction	Intersectional subgroup identification and variable importance
Decision tree algorithms	None	CART, CIT, C50	CART, CIT	CART, CIT, C50, CHAID

4.1 Manuscript 1: Scoping review

The scoping review in *manuscript I* was conducted following the PRISMA-ScR guidelines for scoping reviews (Tricco et al., 2018), and the five-step methodological framework proposed by Arksey & O'Malley (2005) (Arksey & O'Malley, 2005). The scoping review protocol was registered with the Centre for Open Science (Pedrós Barnils et al., 2024) and can be found in Supplementary File 1 from *manuscript I.*

The scoping review aimed at identifying sociodemographic inequalities in BCS attendance among women aged 50-69 years since the implementation of the OSP in Germany. For the search strategy, the PCC (Population, Concept and Context) criteria were employed, and the following bibliographic databases were searched for the period from January 2005 to January 2024: Web of Science, Scopus, MEDLINE (via PubMed), PsycINFO (via Ovid), and CINAHL (via EBSCO) (see Supplementary File 2 from *manuscript I* for the search strategy). Following the established guidelines for snowballing, backward snowballing was conducted for the included articles (Wohlin, 2014). Moreover, relevant grey literature was searched on pertinent national public health institutions' websites: Bundesgesundheitsblatt, Bericht zum Krebsgeschehen in Deutschland, etc (see Supplementary File 3 from *manuscript I* for the grey literature search).

After systematically searching all electronic databases, 476 references were imported into Rayyan (Ouzzani et al., 2016) and two authors screened them. Data from 27 included records (11 identified via screening and 16 via snowballing) was charted by three authors in an Excel sheet with a pre-defined table comprising bibliographic information, methods, and results of the studies, prioritising univariable models' information over multivariate when both were available. The records included were critically appraised using the National Institutes of Health Quality Assessment Tool (National Heart Lung and Blood Institute, 2021). Finally, two synthesis approaches were taken based on the available information. On the one hand, the sociodemographic information presented in the 14 national reports was subjected to a narrative synthesis. On the other hand, the heterogeneous effect sizes of the sociodemographic variables provided in 13 peer-reviewed articles were summarised employing harvest plots. These flexible plots, agnostic to the outcomes and measures used, allow simultaneous display of several dimensions (e.g. study design, sample size, etc.) (Ogilvie et al., 2008). Harvest plots were conducted using R (version 4.2.3). Details on the harvest plots building process are available in Supplementary File 6 from *manuscript I.*

4.2 Manuscripts II – IV: Empirical studies on cancer screening attendance

The empirical research of this thesis (*manuscript II* to *manuscript IV*) is based on the cross-sectional European survey European Health Interview Survey (EHIS) in three different countries: Germany (*manuscript II*), Sweden (*manuscript III*) and Spain (*manuscript IV*).

4.2.1 Survey - European Health Interview Survey

In *manuscript II* employed cross-sectional data from the EHIS third wave conducted in Germany in 2019. The survey sample size was 23,001 respondents, corresponding to 21.6% of the invited participants (N= 23,001; 21.6% response rate) (Jennifer Allen et al., 2021). The target population were women aged 50-69 residing in Germany.

In *manuscript III* we used cross-sectional data from the third EHIS wave in Sweden in 2019 for *paper III*. The survey sample size was 9,757 respondents, constituting 32.52% of the invited participants (n=9,757, response rate: 32.52%) (The Public Health Agency of Sweden, 2019-2020). The target population were people aged 60-69 years old residing in Sweden.

In *manuscript IV*, we employed the cross-sectional data from the EHIS third wave conducted in Spain in 2019-2020 with a total survey sample size of 22,072 respondents and a response rate of 59% (N= 22,072; 59% response rate) (INE, 2021). The target population were women aged 45-69 in 7 out of 19 regions (Castilla-La Mancha, Castile and León, Valencian Community, La Rioja, Navarre, Ceuta, and Melilla) and 50-69 years old for the remaining regions.

The EHIS is conducted on a quinquennial basis (sexennial since 2019) and focuses on individuals aged 15 and above residing in private households. Its primary goal is to collect standardised and comparable data across Europe about the population's health status, healthcare services utilisation, and health determinants. The survey is legally binding from the second wave onward and framed within the EC Regulation 1338/2008 (Council of the European Union, 2008) and is currently implemented through the EC Regulation 2018/255 (Council of the European Union, 2008, 2018).

4.2.2 Research outcome

The primary outcome for the three papers was self-reported lifetime cancer screening attendance. For *manuscripts II and IV*, the outcome was lifetime BCS non-attendance. For *manuscript III*, the outcome was lifetime CRC screening non-attendance.

4.2.3 Framework of inequalities and explanatory variables

Two different frameworks were adopted in this thesis to guide the selection of potential dimensions of inequalities. In *manuscripts II* and *IV*, the PROGRESS-Plus framework was used; in *manuscript* III, the Conceptual Framework for Action on the Social Determinants of Health was employed.

PROGRESS-Plus framework

The PROGRESS-Plus framework entails the following characteristics: place of residence, race, ethnicity, culture and language, occupation, gender/sex, education, socioeconomic status, social capital and plus (i.e. other potentially discriminatory factors) (Oliver et al., 2008). Their use to disentangle health inequalities has been extensively discussed (O'Neill et al., 2014) and used (Coetzee et al., 2022). Table 3 presents the PROGRESS-Plus characteristics and the number of categories within each variable employed in *manuscripts II* and *IV* as potential predictors for BCS never-attendance.

Table 3. PROGRESS-Plus characteristics and their categories employed in manuscripts II and IV

PROGRESS-Plus	Manuscript II	Manuscript IV
Place of Residence	Degree of urbanisation: 3 Region (federal states): 17	Size of the municipality: 7 Region (Autonomous Communities): 19
Race, ethnicity, culture, and language	*Proxy variables:* Country of origin: 3 Nationality: 3	*Proxy variables:* Country of origin: 3
Occupation	Working situation: 6	Working situation: 6
Gender/sex	Prerequisite to be female	Prerequisite to be female
Religion	Not reported	Not reported
Education	ISCED-2011 classification: 6	CNED14 classification: 8
Socioeconomic status	Household income: 5	Type of occupation: 6
Social capital	Social network dimensions: 5 Perceived social support: 5 Ease in available help: 5 *Proxy variables:* Marital status: 4 Type of household: 5 Partnership cohabitation: 2	*Proxy variables:* Marital status: 4 Type of household: 5
Plus	Global Activity Limitations Indicator (GALI): 3	Global Activity Limitations Indicator (GALI): 3 Age: 5

Conceptual Framework for Action on the Social Determinants of Health

The Conceptual Framework for Action on the Social Determinants of Health (CSDH) was developed by Solar and Irwin in 2007 for the Commission on Social Determinants of Health of the World Health Organization, and it schematises structural and intermediary determinants that affect equity in health and well-being. Structural determinants include the socioeconomic and political context (e.g. health policies in specific regions) and the socioeconomic position of individuals according to income, education, occupation, gender, race, and ethnicity. These influence intermediary determinants of health that impact health: material circumstances, psychosocial processes, behaviours and biological factors (World Health Organization, 2010). In *manuscript III,* six structural determinants of health were used as social dimensions for building intersectional strata on CRC screening inequalities in Sweden: age (2 categories), gender (proxy sex – 2 categories), migration background (country of origin – 3 categories), education (8 categories), income (5 categories) and region of residence based on screening organisation (2 categories).

4.2.4 Decision tree algorithms

This dissertation employed several decision tree algorithms, which are the most widely applied in public health: Classification and Regression Tree (CART), Conditional Inference Tree (CIT), Chi-squared Automatic Interaction Detection (CHAID) and C5.0. Below, the decision tree methods are introduced, and the specific algorithms employed in this dissertation are methodologically described. Finally, approaches to building decision trees and those applied to the explorative decision trees of this dissertation are explained.

Decision trees are non-parametric tools that can identify homogenous intersectional groups by combining available sociodemographic variables in a dataset based on a set of decision rules. Decision trees consist of nodes, branches, and terminal nodes or leaves. The root node represents all observations. This node is divided based on a splitting criterion that minimises a specific loss function (e.g. misclassification error for classification tasks) at a splitting point (i.e. a value for a continuous variable or a category for a categorical variable), creating tree branches. These branches are recursively split until a stopping rule is reached, forming terminal nodes or leaves. Terminal nodes partition the original data (i.e. the entire data set for explorative decision trees and the training data set for predictive decision trees), with each observation assigned to one leaf. For binary outcomes, observations are predicted as 0 or 1 based on the average predicted value of their respective leaf (Venkatasubramaniam et al., 2017).

While building a decision tree, each algorithm employs different splitting criteria and stopping rules. However, all decision trees can be stopped based on predefined external conditions to the algorithm (e.g. maximum depth of the tree, minimum number of observations per leaf).

Table 4 summarises the specific splitting criteria and stopping rules for every decision tree algorithm used in this dissertation.

Table 4. Splitting criteria and stopping rules of decision trees used in this dissertation

	Splitting criteria	Stopping rule
CART	Gini impurity	Pruning based on complexity parameter
CIT	Chi-squared tests for the significance of splits	The null hypothesis can no longer be rejected
CHAID	Chi-squared tests for independence of each covariate against the outcome	The null hypothesis can no longer be rejected
C5.0	Information gain (Entropy-based)	Pruning based on the binomial confidence limit method

CART bases its splitting decision on the Gini impurity coefficient when involved in a classification task (i.e. binary outcome). The algorithm divides the dataset into two groups to maximise homogeneity within each subgroup and heterogeneity between the subgroups. If not specified, CART does not stop the growing process. Rather, it grows a deep tree and then prunes its branches until the reduction in error resulting from the optimal split is lower than a threshold known as the complexity parameter (Breiman et al., 2017).

The splitting decision in CIT is based on chi-squared tests for the significance of splits. The process of splitting is divided into two-step statistical tests. Initially, univariate statistical associations between all covariates and the outcome are evaluated. If the overall null hypothesis (i.e. no covariates are associated with the outcome) is rejected, the covariate with the strongest association is selected for splitting. Next, the splitting point (a category for categorical variables or a value for a continuous variable) is chosen based on the highest statistical significance. The growth of CIT stops when the overall null hypothesis can no longer be rejected at a predetermined significance level (Hothorn et al., 2006).

CHAID, the first decision tree developed based on statistical significance tests, works only with categorical variables and bases its splitting decision on a chi-square test for independence of each covariate against the outcome. The variable with the highest association is the splitting variable, which can create non-binary splits. CHAID stops growing when no more statistically significant associations emerge, that is, when the null hypothesis can no longer be rejected (Kass, 1980).

C5.0 uses the entropy of the imputed variables to generate splits; nodes are generated based on the data split that produces lower entropy (i.e. higher information gain after the split) (Quinlan, 1993). C5.0 generates a vast tree and prunes it using every branch's binomial confidence limit method. C5.0 employs more complex methods in the growing process than CART, CIT or CHAID. It uses adaptative boosting and winnowing in the growing process. Adaptative boosting is the technique through which a strong learner (i.e. final decision tree) is built from weak learners (i.e. intermediate trees) (Schapire, 1990). After creating a first decision tree, misclassified cases are used to build a second tree (i.e. giving them more attention). Then, misclassifications of a second decision tree are used to construct a third decision tree, and so forth, until it reaches extremely high accuracy or until a pre-determined number of iterations (Freund, 1996). Winnowing entails reducing the dimensionality of the variables in the decision trees by weighting their importance for the growing performance (Littlestone, 1988).

In the process of building any decision tree, cross-validation methods are involved to prevent overfitting the data. These involve dividing the data into multiple subsets (e.g. 5 parts for a 5-fold cross-validation), building the model using all subsets but one (e.g. 4 subsets), and using the left-out for validating the decision tree. This process is repeated k times (e.g. 5 times), and the results of all folds are averaged to obtain the final result.

As mentioned in Chapter 1, there are two applications of decision trees based on their goal: explorative and predictive. For predictive trees, the aim is to build models that perform well in predicting the outcome of unseen observations; therefore, after splitting the data set into training and test data, the parameters of the decision trees are hypertuned on the training dataset and tested on the test data set. However, explorative decision trees aim to examine the data structure and find explainable patterns, for creating subgroups of the entire dataset. Across the literature, explorative decision trees have been built based on a rule of thumb (Mahendran et al., 2022; Mena, Bolte, & Advance Gender study, 2021; Venkatasubramaniam et al., 2017). However, *manuscripts II* and *III* have used optimisation techniques. Following, both approaches are described, and Chapter 6 will discuss their identified advantages and disadvantages.

Applying a rule of thumb for building a decision tree aims at simplifying the process of choosing which is the best tree. In the case of CART, this rule of thumb is to prune the decision tree following the 1-SE rule. After growing a CART without an early stopping criterion, the size of the decision tree is reduced by selecting the least complex tree whose error is within one standard error above the tree with the smallest error (Breiman et al., 2017). Then, the complex parameter that follows this rule is chosen, and the corresponding decision tree is selected. In the case of CIT, created to incorporate statistical language into the splitting process of decision trees, a pre-specified nominal level of the underlying association tests for splitting is defined (e.g. alpha=0.05) as well as the distribution of the test statistic (e.g. Bonferroni, univariate relate to p-values from the asymptotic distribution (adjusted or unadjusted)). CIT then grows until the global null hypothesis cannot be rejected (Hothorn et al., 2006).

An optimisation technique for building a decision tree implies that, instead of choosing the best tree based on a predefined rule, the parameters of the decision tree that composed the splitting criteria are hypertuned optimising the performance measure of our choice (e.g. AUC, sensitivity, specificity, etc). Here, decision trees' parameters are tuned based on trying many combinations of the learner's parameters, and selecting those which performed best in correctly classifying cases rather than a rule of thumb. Here, a 2 x 2 table (i.e. a confusion matrix) is built illustrating the total number of true positives, true negatives cases, false negatives (type II error), and false positives (type I error) when comparing the predicted positive and predicted negative cases to the actual positive and actual negative cases. Then, performance measures such as balanced accuracy, sensitivity and specificity are calculated (Stehman, 1997). This is the classical hypertuning process for predictive modelling; however, in this case, we apply it for explorative goals instead of predicting goals.

Currently, R has different interfaces to hypertuning models' parameters: *caret*, *tidyverse*, and *mlr3*. In *manuscripts II* and *III*, the decision trees have been built based on *mlr3*, which is the most advanced implemented ecosystem for optimisation techniques for building trees in R (Bischl et al., 2024). In this ecosystem, the search for the best parameters can be ordered (i.e. grid search) or random (i.e. random search) and evaluated either in one or two performance measures simultaneously. Building the trees based on hypertuning means that after setting the values of certain (or none) parameters (e.g. alpha=0.05 for CIT to simulate classical statistical significance) we can tune other learner's parameter and evaluate the model's performance giving more importance to the measure we are interested (e.g. AUC, sensitivity, F1 score, etc). In advanced ecosystems such as *mlr3*, even two performance measures can be evaluated simultaneously.

4.2.5 Analytical strategy and use of decision trees

The analytical strategy of *manuscripts II-IV* constitutes the research aim of this thesis: the exploration of the use of decision trees for subgroup identification at higher risk of never attending cancer screening appointments. Three analytical strategies have examined different uses of decision trees for the research aim of this dissertation. Figure 3 represents graphically these analytical strategies.

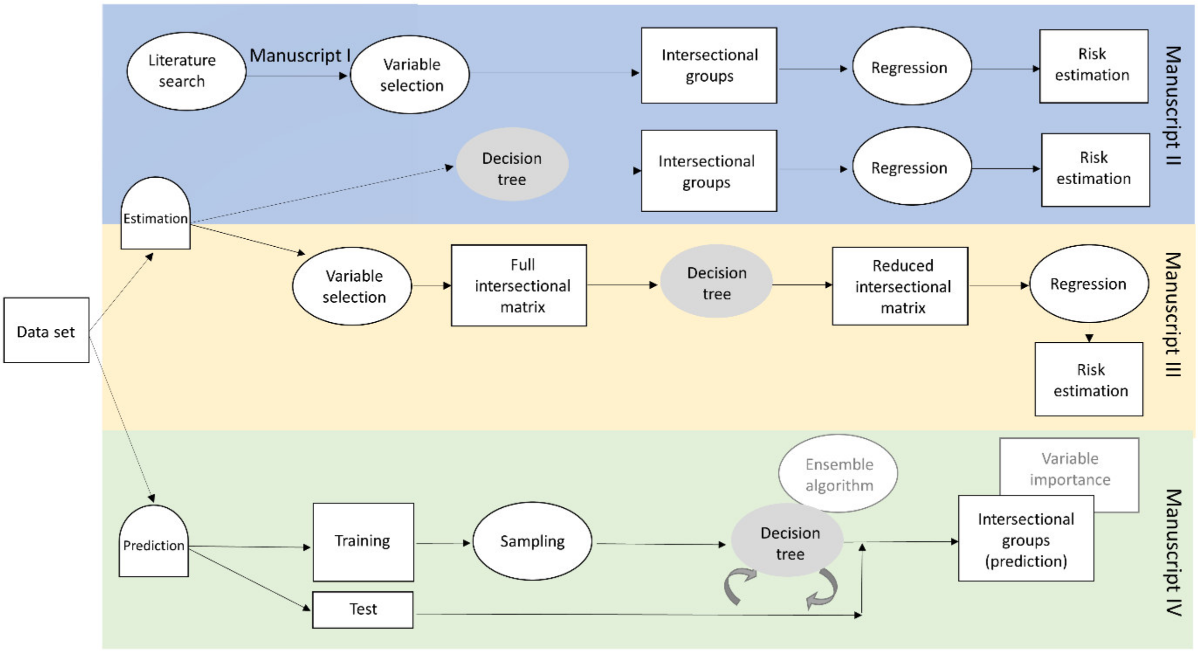


Figure 3. Analytical strategies employing decision trees used in the dissertation

To enhance the clarity of the analytical strategies and results, this subsection is presented in Chapter 5, along with the findings.

4.2.6 Ethical considerations

Ethical approval was not required for these studies as it is a secondary analysis of de-identified data. Spanish data was publicly available and de-identified. Eurostat, the European body for Statistics, granted us access to the German and Swedish data and all authors accessing the data signed the individual confidentiality declaration following EC Regulation number 223/2009 (Council of the European Union, 2009).

Chapter 5. Results

Chapter 5 presents the findings of the different *manuscripts* that constitute the base for answering the research objectives and builds the empirical justification for the discussion developed in Chapter 6.

5.1 Decision trees for intersectional subgroup identification in comparison with an evidence-informed approach

The first research objective of this cumulative dissertation was to compare the use of decision trees for identifying intersectional subgroups of women at risk to an evidence-informed approach, that is, determining the intersectional subgroups of women based on existing evidence. To examine this research objective, first, the existing evidence in the literature on sociodemographic inequalities in BCS attendance in Germany was assessed in *manuscript I* through a scoping review. Second, *manuscript II* compared the use of decision trees for intersectional subgroup identification to an evidence-informed approach.

5.1.1 Manuscript I: Sociodemographic inequalities in BCS attendance following the OSP implementation in Germany

The scoping review conducted in *manuscript I* included 27 relevant records for analysis: 14 national reports and 13 peer-reviewed articles.

Based on the 13 peer-reviewed articles, eight relevant sociodemographic variables were determined and summarised in harvest plots accounting for their effect size direction. After analysing the effects provided for age (six), education (eleven), income (four), migration background (seven), type of district (three), employment status (seven), partnership cohabitation (two), and health insurance (two), the scoping review concluded that older women with lower incomes, women with migration backgrounds, women who live in rural areas and those with statutory health insurance respond more favourably to BCS invitations. However, from a lifetime perspective, these associations only hold for migration background, are reversed for income and urban residency, and are complemented by partner cohabitation.

Based on the 14 national reports, the scoping review stated that women living in the former East German states of Saxony, Mecklenburg-Western Pomerania, Saxony-Anhalt, and Thuringia, as well as in the former West German state of Lower Saxony, had higher BCS attendance rates in the last two years.

5.1.2 Manuscript II: Comparison of evidence-informed approach and decision tree-based approach

Aim, analytical strategy and use of decision trees in manuscript II

Manuscript II aimed to identify intersectional groups of women aged 50-69 at higher risk of never attending BCS in Germany. To do that, it compared two different analytical strategies: evidence-informed regression and decision tree-based regression.

The evidence-informed analytical strategy built a full cross-classification matrix based on the dichotomised social dimensions identified as relevant for BCS attendance estimation on the scoping review in *manuscript I* (i.e. migration background, socioeconomic position (based on income), urbanisation degree, and partner cohabitation). Following this, a multivariate logistic regression was carried out to estimate the odds ratio (OR) of never attending BCS adjusted by age. The DA was estimated through the area under the receiver operating characteristics curve (AUC) with a 95% confidence interval (CI), indicating how well each model discriminates between women attending and women never attending BCS. DA is considered absent or very small when $0.5 \leq AUC \leq 0.6$, moderate when $0.6 < AUC \leq 0.7$, large when $0.7 < AUC \leq 0.8$ and very large $AUC > 0.8$ (Axelsson Fisk et al., 2021).

The decision tree-based regression consisted of two steps. First, an exploratory decision tree was constructed to identify homogeneous subgroups of women at higher risk of never attending BCS in Germany. CART, CIT and C5.0 were trained using the entire sample (N=4,761). Given the (relative) rareness of the outcome in the dataset (10.38% prevalence), stratified cross-validation was performed (i.e. the folds were created to ensure an equal proportion of each outcome category in every fold), and cost weights were applied to distribute the sums of weights equally for cases and non-cases. Decision trees were grown using the *tune* function from the "mlr3tuning" optimisation R packages in R version 4.4.0. This package integrates essential packages for building CART "rpart" (Therneau et al., 2023), CIT "partykit" (Hothorn et al., 2006), and C5.0 "C50" (Max Kuhn, 2023), and it evaluates the best tree based on one or two performance measures. This paper used sensitivity (i.e. enhancing detection of positive cases) and the Area Under the Precision-Recall Curve (i.e. improving overall precision-recall performance for unbalanced datasets) for choosing the optimal decision tree (Saito & Rehmsmeier, 2015).

Second, a multivariate logistic regression was conducted, with the decision tree outcome adjusted for age, to estimate the OR for never attending BCS and the DA of the model.

Manuscript II used decision trees to select relevant exposure variables and form intersectional groups for later use in regressions.

Findings from Manuscript II

Results on the relevant sociodemographic variables identified in *manuscript I* for lifetime BCS attendance shaped the social dimensions used to build the evidence-informed approach. Here, income, migration background, type of district (also named urbanisation degree), and partnership cohabitation were selected from the EHIS survey and dichotomised. Then, the dichotomised categories were cross-classified (i.e. 2 migration background * 2 income * 2 degree of urbanisation * 2 partnership cohabitation), leading to 16 intersectional subgroups.

A variable formed by these sixteen intersectional subgroups was employed as an exposure variable in a logistic regression with high-income women born outside Germany, living in urban areas with a partner as a reference group (i.e. the group expected to have the highest BCS attendance rate based on *manuscript I*) and adjusted by age.

The results indicated that four intersectional subgroups were significantly associated with never attending BCS. Low-income women not born in Germany and living in rural areas with no partner showed the highest odds (OR=9.48 (2.24-40.10), p=0.002). The DA of the full cross-classification model was moderated (AUC=0.6618) and 0.0079 points higher than the main effects model (i.e. multivariate logistic regression with all independent variables – data not shown). This indicates a slightly better DA in classifying women attending or never attending BCS from the cross-classified logistic regression than the main effects model.

The decision tree-based approach identified CART (cp=0.006713025 and maxdepth=4) as the tree with the highest sensitivity (72.47%) and balanced accuracy (61.91%). Specificity was, however, low (51.35%). The decision tree identified household type, marital status, working situation, region and perceived social support as relevant variables. Figure 4 and Table 5 show the final decision tree and the emerged intersectional subgroups.

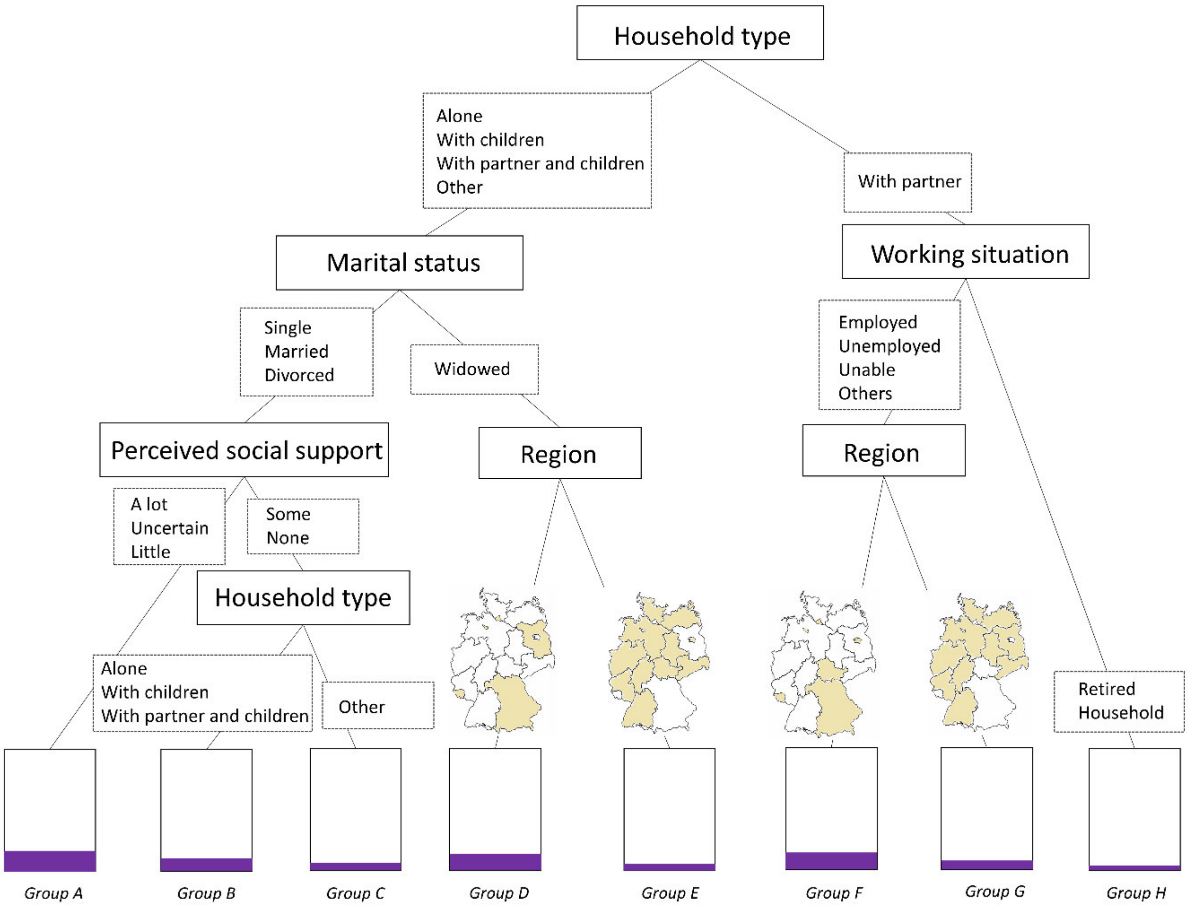


Figure 4. CART on never attending BCS among targeted German women

Table 5. Intersectional subgroups on never attending BCS in Germany based on CART

Tree leaf and label	Intersectional subgroups description	Rank[a]	Size, Prevalence

H	Women living with a partner, retired or doing unpaid household work	1	N=882
			Pr= 0.0454

E	Widowed women living alone, with children, with a partner and children or other arrangements, residing in Baden-Württemberg, Berlin, Hesse, Mecklenburg-Vorpommern, Lower Saxony, North Rhine-Westphalia, Rhineland-Palatinate, Saxony, Saxony-Anhalt, and Schleswig-Holstein or Thuringia	2	N=316
			Pr= 0.0506

C	Single, married or divorced women living in other living arrangements, with some or no perceived social support	3	N=211
			Pr= 0.0616

G	Women living with a partner, who are either employed, unemployed, unable to work, or in other categories, and residing in Baden-Württemberg, Brandenburg, Hesse, Mecklenburg-Vorpommern, Lower Saxony, North Rhine-Westphalia, Rhineland-Palatinate, Saxony, Saxony-Anhalt, or Schleswig-Holstein	4	N= 918
			Pr= 0.0730

B	Single, married or divorced women living alone, with children, with a partner and children, with some or no perceived social support	5	N= 953
			Pr= 0.1301

F	Women living with a partner who are either employed, unemployed, unable to work, or in other working categories and residing in Bavaria, Berlin, Bremen, Hamburg, Saarland or Thuringia	6	N= 472
			Pr= 0.1377

D	Widowed women living alone, with children, with a partner and children or other arrangements, residing in Bavaria, Brandenburg, Bremen, Hamburg, or Saarland	7	N= 136
			Pr= 0.1471

A	Single, married or divorced women living alone, with children, with a partner and children or other arrangements, with little, uncertain or a lot of perceived social support	8	N= 873
			Pr= 0.1707

[a] Ordered by the increasing prevalence of the outcome

Following, a logistic regression with an exposure variable formed by the intersectional subgroups identified by CART adjusted by age was carried out. Subgroup H (i.e. women living with a partner, retired or doing unpaid household work) was the reference category since it depicted the lowest never-attended BCS prevalence.

Four CART intersectional subgroups displayed a statistically significant difference compared to subgroup H, with subgroup D (i.e. widowed women living alone, with children, with a partner and children or other arrangements, residing in Bavaria, Brandenburg, Bremen, Hamburg, or Saarland) showing the highest odds of never attending BCS (OR=3.43 (0.81-1.96); p<0.001). The model's DA was moderate (AUC = 0.6726), showing an improvement of 0.0108 points compared to the evidence-informed regression, therefore, a better classification performance.

5.2 Manuscript III: Decision trees as intersectional complexity reduction tools

Aim, analytical strategy and use of decision trees in manuscript III

The second research objective of this dissertation was to use decision trees as tools for reducing the complexity of full intersectional matrixes. *Manuscript III* did so by examining inequalities in lifetime CRC screening attendance patterns among individuals aged 60-69 in regions with and without organised screening programs in Sweden. These inequalities were investigated by identifying intersectional subgroups at higher risk of never attending CRC screening. This research is framed within the methodological discussions of quantitative intersectionality and postulates using decision trees as a complementary tool for the AIHDA method. Further details on AIHDA can be found elsewhere (Merlo et al., 2023; Mulinari et al., 2015; Wemrell et al., 2021). *Manuscript III* employs decision trees as a prior step for identifying homogenous intersectional groups and reducing the size of the intersectional matrix used when applying AIHDA.

Eight logistic regression models were constructed to step-by-step build the full intersectional matrix. Models 1- 6 estimated the crude main effects of the univariate regression of the six exposure variables (i.e. age, gender, migration background, education, income and region of residence), measuring the relative importance of each exposure variable without considering the other variables. Model 7, a multivariable regression with all main effects, assessed the effects of each variable on the outcome by controlling for all other variables. Finally, model 8 displayed all possible categories' interactions (full cross-classification). The DA of each model was calculated through the AUC with 95% CI.

To build the reduced intersectional matrix, CART and CIT were built using the above-mentioned social dimensions to find the optimal combination of social positions that better classify lifetime CRC screening attendances. The *tune* function from the "mlr3tuning" optimisation package in R version 4.4.0 with CART "rpart" (Therneau et al., 2023) and CIT "partykit" (Hothorn et al., 2006) was employed. The decision trees were tuned with a restricted maximum depth of 7 generations and an alpha of 0.05 (i.e. to simulate the standard 0.05 significance threshold). The decision tree that performed the highest AUC was chosen. The outcome of the decision tree, the leaves, was then employed as the main exposure (i.e. a categorical variable with the best-off intersectional group as the reference category) for a Poisson regression with robust standard errors. The regression estimated the prevalence ratios (PR) of never attending CRC screening. Additionally, the DA of the model was assessed, hence conducting a reduced matrix AIHDA.

Manuscript III used decision trees as tools for reducing the complexity of full intersectional matrixes with high dimensionality of variables and categories. Furthermore, *manuscript III* placed decision trees in the middle of the quantitative intersectional methodological discussion.

Findings from Manuscript III

Models 1-6 (crude models Table 2 in *manuscript III*) displayed significant univariate relationships for country of origin, some education categories, income and region. The sociodemographic variable that showed the highest DA was region of residence (AUC=0.6888), followed by country of origin (AUC=0.5876), education (AUC=0.5633) and income (AUC=0.5419). The main effects regression (adjusted model Table 2 in *manuscript III*) showed significant effects for the region of residence ($OR_{\text{opportunistic}} = 11.09$, $p<0.001$), country of origin ($OR_{\text{Europe-born}} = 0.64$, $p=0.020$; $OR_{\text{no Europe-born}} = 0.47$, $p<0.001$), and two educational groups (OR $_{\text{primary}}$=2.19, p=0.037; OR $_{\text{post-secondary}}$=2.40, p=0.041). Furthermore, the overall DA rose up to 0.7483.

The full cross-classification model (i.e. full intersectional matrix) (model 8, data not shown) presented an AUC of 0.6959, 0.0524 points lower than the main effects model. Initially, this result would suggest no intersectional effects, as an increased number of intersectional subgroups (i.e. categories in the variables) reduced the overall DA. However, this behaviour has a statistical explanation: the size of the full intersectional matrix (400 subgroups), combined with a relatively small sample (n=1,268), led to numerous empty estimates, resulting in a loss of information and accuracy. For more detailed information, Appendix 1 in *manuscript III* illustrates the concave relationship between the increased number of intersectional groups and AUC.

The decision tree with the highest AUC was CIT, with 4 generations and a set alpha of 0.05. The model's inner performance (i.e. calculated on the same training dataset) deployed an AUC of 0.7489, a sensitivity of 92.29%, a specificity of 49.47% and an F-score of 0.7912. Figure 5 and Table 6 illustrate and describe the final decision tree model.

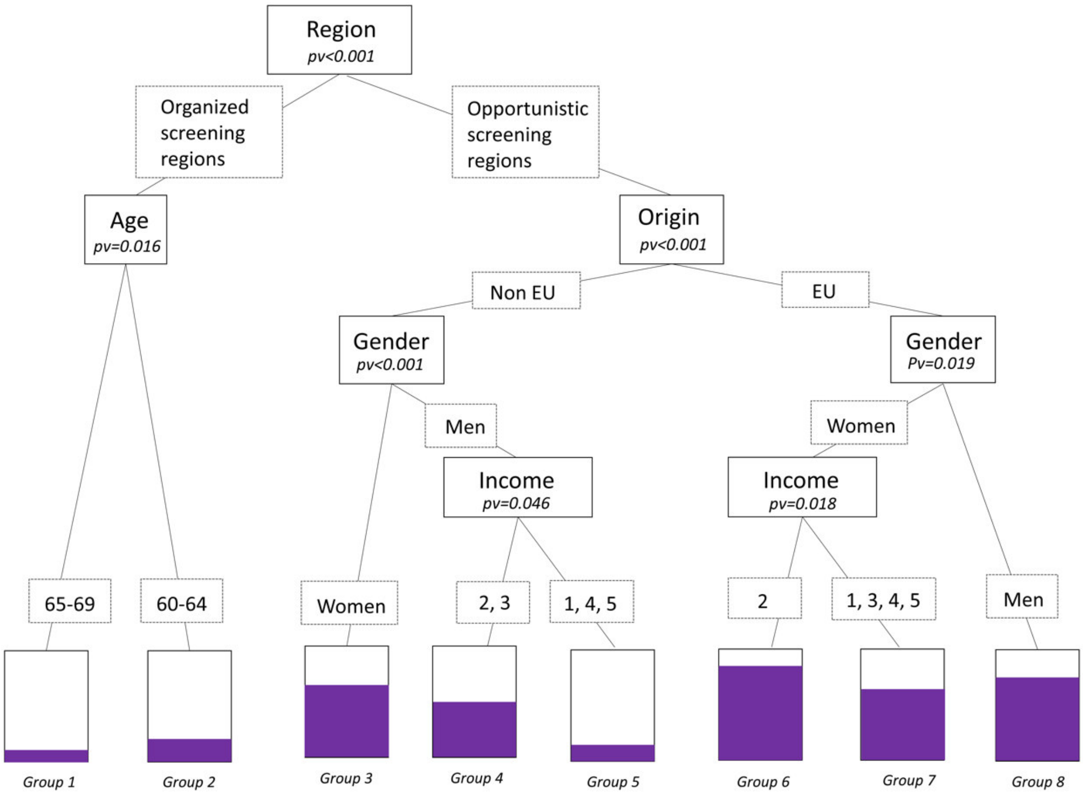


Figure 5. CIT on never attending CRC screening among the targeted Swedish population

Table 6. Intersectional subgroups identified by CIT on never attending CRC screening among the targeted Swedish population

Tree leaf	Intersectional subgroups description	Rank[a]	Size, Prevalence

1	People in organised screening regions aged 65-69	1	N=141
			Pr= 0.0993

5	Non-EU born men in opportunistic screening regions (1st, 4th or 5th quintile)	2	N= 33
			Pr= 0.1515

2	People in organised screening regions aged 60-64	3	N= 155
			Pr= 0.2000

4	Non-EU born men in opportunistic screening regions (2nd or 3rd quintile)	4	N= 20
			Pr= 0.5000

3	Non-EU born women in opportunistic screening regions	5	N=61
			Pr= 0.6393

7	EU-born women in opportunistic screening regions (1st, 3rd, 4th or 5th quintile)	6	N= 369
			Pr= 0.6396

8	EU-born men in opportunistic screening regions	7	N= 443
			Pr= 0.7359

6	EU-born women in opportunistic screening regions (2nd quintile)	8	N=46
			Pr= 0.8478

[a] Ordered by the increasing prevalence of the outcome

CIT identified region of residence, country of origin, gender, age and income as relevant variables in explaining lifetime CRC screening attendance inequalities in Sweden.

Poisson regression with robust standard errors was carried out with the intersectional subgroups identified by CIT (aka reduced intersectional matrix). The intersectional subgroup with the highest lifetime CRC screening attendance prevalence (i.e. people in organised screening regions aged 65-69) was employed as the reference category.

The analysis revealed that women born in the EU living in opportunistic screening regions had the highest risk of never attending CRC screening, with a prevalence ratio eight times higher than the reference intersectional strata (PR=8.54, p <0.001). Closely, EU-born men in opportunistic screening regions showed a seven-fold prevalence ratio (PR=7.41, p <0.001). Following, EU-born women belonging to the 1st, 3rd, 4th or 5th quintile in opportunistic screening regions (PR=6.44, p<0.001), non-EU-born women in opportunistic screening regions (PR=6.44, p<0.001), and non-EU men belonging to the 2nd or 3rd income quintiles in opportunistic screening regions (PR=5.04, p<0.001) depicted five to six times higher prevalence ratio. The last intersectional subgroup showing a statistically significant relationship was people aged 60-64 also residing in organised screening regions with a two-fold prevalence ratio (PR=2.10, p=0.020) compared to the reference group.

The AUC of the Poisson regression with the intersectional subgroups was 0.7489, 0.053 points higher than the full cross-classification (AUC_{M8}=0.6959) and 0.0006 points higher than the main effects logistic regression (AUC_{M7}=0.7483). These values unveil a small intersectional effect of the reduced intersectional matrix compared to the main effects model and largely better DA than the full cross-classification.

5.3 Manuscript IV: Decision trees as predictive tools for intersectional subgroup identification

Aim, analytical strategy and use of decision trees in manuscript IV

The third objective of this dissertation was to explore the use of decision trees as predictive tools for intersectional subgroup identification. *Manuscript IV* did so by identifying subgroups of women aged 45/50 to 69 most at risk of not attending BCS in Spain. In this manuscript, the use of decision trees is as predictive tools, that is, decision trees are built on the train data (80% of the data set) and afterwards tested (i.e. how well does the model classify unseen cases) in the test data (20% of the data set).

Given the unbalanced nature of the outcome (6.87%), balancing sampling techniques (i.e. oversampling) were applied. Then, several decision trees were trained on the oversampled training data set using the R "caret" (Kuhn, 2008) package in R version 4.2.3 for hypertuning the parameters of the following learners: "rpart" (CART) (Therneau et al., 2023), "chaid" (CHAID) (The FoRt Student Project Team, 2009), "partykit" (CIT) (Hothorn et al., 2006), and "C50" (C5.0) (Max Kuhn, 2023).

Further, in order to increase the robustness of the decision tree, decision tree ensembles were performed (Apté & Weiss, 1997). These can be built employing diverse procedures such as bagging (Random Forest (Breiman, 2001a)), boosting (AdaBoost (Freund, 1996)), and bagging and boosting (Extreme Gradient Boosting (Chen & Guestrin, 2016)). This "forest of trees" will then output a robust average predictive importance of the different predictors (e.g. PROGRESS-Plus characteristics). For building the decision tree ensemble, *manuscript IV* run an Extreme Gradient Boosting (XGBoost) using the "xgboost" package in R version 4.2.3. XGBoost calculated the predictive importance of each variable through the information gain criteria. That is, it measures the contribution of each predictor to the improvement in the model's performance at each split.

In *manuscript IV*, decision trees were used as predictive tools to identify intersectional subgroups and built into ensembles to robustly rank the importance of the included predictors.

Findings from Manuscript IV

Several decision trees, CART, CIT, CHAID, and C5.0, were trained on the oversampled training data. Decision tree C5.0 outperformed the others with a balanced accuracy of 81.1%, sensitivity of 80.7%, specificity of 81.5%, and F1-score of 0.374. C5.0 identified the following predictors as relevant in predicting non-attendance to BCS in Spain: region, origin, marital status, age, education, socioeconomic class, origin, and type of household. Figure 6 and Table 7 illustrate and describe the final model.

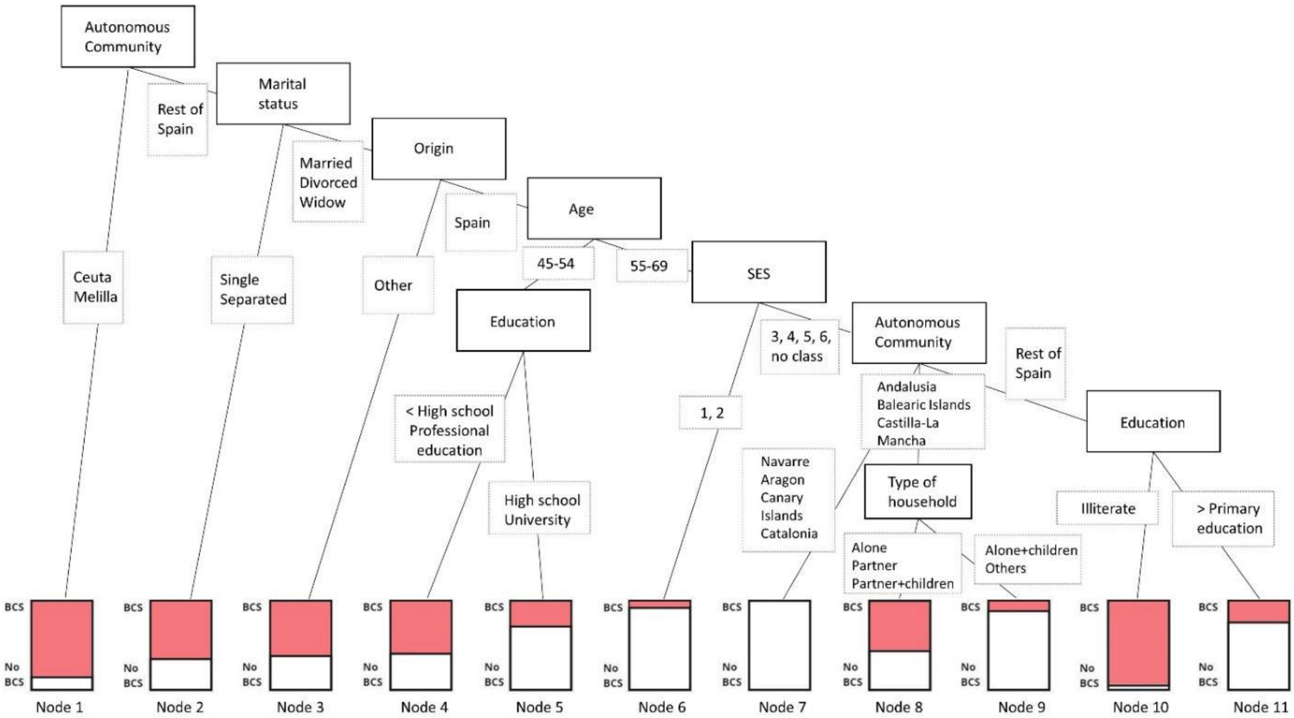


Figure 6. Predictive C5.0 on never attending BCS among targeted Spanish women

Table 7. Intersectional subgroups identified by C5.0 on predicted BCS never attendance among targeted Spanish women

Tree leaf and label	Intersectional subgroup description	Rank[a]
10	Illiterate women living in Asturias, Cantabria, Castile and León, Valencian Community, Extremadura, Galicia, Madrid, Murcia Basque Country or La Rioja, from middle and low social classes, aged 55 or older, born in Spain, married, divorced or widowed	1
1	Women living in Ceuta or Melilla	2
2	Single or separated women not living in Ceuta or Melilla	3
3	Married, divorced, or widowed women born outside Spain and not living in Ceuta or Melilla	4
4	Married, divorced, or widowed women, born outside Spain and not living in Ceuta or Melilla, younger than 55 years old with less than high school education or professional education	5

8	Married, divorced, or widowed women, living alone, with a partner, or with a partner and children, born outside Spain, older than 55 years old, belonging to social class 3-6, living in the regions of Andalusia, Balearic Islands or Castilla-La Mancha	6
5	Married, divorced, or widowed women, born outside Spain and not living in Ceuta or Melilla, younger than 55 years old with more than high school education (except professional education)	7
11	Women living in Asturias, Cantabria, Castile and León, Valencian Community, Extremadura, Galicia, Madrid, Murcia Basque Country or La Rioja, from middle and low social classes, with primary education or greater, aged 55 or older, born in Spain, married, divorced or widowed	8
9	Married, divorced, or widowed women, living alone with children or other household constellations, born outside Spain, older than 55 years old, belonging to social class 3-6, living in the regions of Andalusia, Balearic Islands or Castilla-La Mancha	9
6	Married, divorced, or widowed women, born outside Spain and not living in Ceuta or Melilla, older than 55 years old, belonging to the highest two social class groups	10
7	Married, divorced, or widowed women born outside Spain, older than 55 years old, belonging to social class 3-6, living in the regions of Navarre, Aragon, Canary Islands, or Catalonia	11

[a] Ordered by the decreasing prevalence of the outcome

The XGBoost that performed best output an AUC of 78.80% and was composed of the following parameters: a learning objective of logistic regression for binary classification, a learning rate of 0.1 (low learning rate to be more robust to overfitting), and a maximum number of boosting iterations of 53. The most important predictors for predicting BCS attendance in Spain were region, education, age, and marital status (Figure 7).

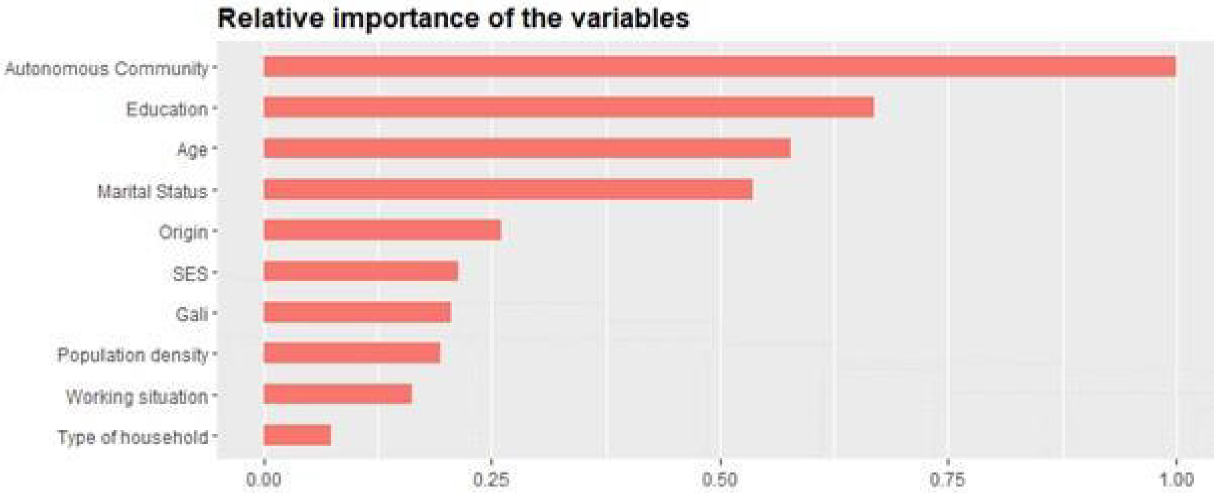


Figure 7. Variables' relative importance in the XGBoost model for predicting BCS never attendance among targeted Spanish women

Chapter 6. Discussion

The aim of this dissertation was to provide evidence on the usability of decision trees as tools for identifying subgroups at higher risk of never attending cancer screening appointments. To this end, three case studies have been conducted, illustrating three relevant analytical approaches in which decision trees facilitate the identification of hitherto unknown risk groups. This chapter first reviews the main empirical findings from the three case studies; second it reviews and discusses the methodological findings that contribute to the research objectives and the overall research aim; third, it underlines the contributions of this dissertation for research, practice and policy; fourth it points out the strengths and limitations of the thesis; and, fifth it develops a short excurses on the concept of generalisability.

6.1 Empirical findings of the three European cases

This dissertation explored inequalities in accessing cancer screening appointments in three European countries: Germany, Sweden and Spain. Following, the relevant intersectional empirical findings for each case study are presented.

6.1.1 Inequalities in breast cancer screening attendance in Germany

Inequalities in BCS attendance in Germany were investigated in *manuscripts I* and *II*. The scoping review (*manuscript I*) identified sociodemographic inequalities in the last two years' BCS attendance and lifetime attendance. Older women with lower incomes, women with migration backgrounds, women who live in rural areas and women with statutory health insurance respond more favourably to BCS invitations. Also, women living in the former East German states of Saxony, Mecklenburg-Western Pomerania, Saxony-Anhalt, and Thuringia, as well as in the former West German state of Lower Saxony, showed higher BCS attendance rates in the last two years. From a lifetime perspective, women with migration backgrounds, women with high incomes, women who live in urban areas and women who cohabitate with their partners showed higher attendance rates.

The empirical research (*manuscript II*) compared evidence-informed and decision tree-based approaches to identify intersectional subgroups of women aged 50-69 at higher risk of never attending BCS in Germany. The evidence-informed approach identified low-income women not born in Germany, residing in rural areas and not cohabitating with their partner as those at the highest risk of never attending BCS. In contrast, the decision tree-based approach determined women living alone, with children, with a partner and children, or in other arrangements, residing in Bavaria, Brandenburg, Bremen, Hamburg, or Saarland as those at higher risk of never attending BCS.

Both approaches recognised household composition as a determining factor. Indeed, the first split of the decision tree in the decision tree-based approach was based on living with a partner or any other form of cohabitation. Furthermore, for the evidence-based approach, the logistic regression revealed that the intersectional subgroup with the highest and lowest risk deferred only on their cohabitation status. Here, an intersectional hypothesis is illustrated: the contingency of inequalities (Wolfson et al., 2017). That is, the discrimination faced by individuals in a particular social position is influenced by their interactions with other social positions. By adopting an intersectional approach, cohabitation with a partner could be defined as a determining social dimension for low-income women not born in Germany and living in rural areas on their likelihood of attending BCS. Several authors have also recognised the importance of partnership cohabitation and breast cancer screening attendance (Hanske et al., 2016; Missinne et al., 2013).

In the decision tree-based approach, inequalities in BCS attendance within regions were identified. Women residing in Bavaria, Bremen, Hamburg, and Saarland depicted a higher risk of never attending BCS than women in their same intersectional position from all other regions. Inequalities between the former East and West German states in preventive behaviour compliance have been previously identified. Großmann (2023) suggested that the previously centralised healthcare and prevention systems in former Eastern states positively influenced women's perception of participation as a public responsibility (Großmann et al., 2023).

The adoption of an intersectional framework, and especially the use of decision trees, in the examination of inequalities in BCS in Germany, makes a precise and nuanced identification of at-risk populations possible, easing the adoption of targeted preventive interventions (Delgado-Gallegos et al., 2023; Eagle et al., 2022).

6.1.2 Inequalities in breast cancer screening attendance in Spain

Manuscript IV explored inequalities in BCS attendance in Spain among women aged 45/50-69 through the implementation of predictive decision trees and ensemble algorithms. The results of the ensemble algorithm indicated that region, followed by education, age, and marital status, were the most important variables for predicting BCS attendance. The intersectional subgroup with the highest predicted prevalence of never attending BCS was illiterate women living in Asturias, Cantabria, Castile and León, Valencian Community, Extremadura, Galicia, Madrid, Murcia Basque Country or La Rioja, from middle and low social classes, aged 55 or older, born in Spain, married, divorced, or widowed. This subgroup of women, although in the intersection of several protective factors, has very low levels of education, reinforcing the importance, on the one hand, of education for BCS attendance (Pons-Rodriguez et al., 2020), and on the other hand, on the intersection with the region of residence. Indeed, during the last decade, the regions of Asturias, Castile and León, Extremadura, Madrid, and Murcia did not develop any intervention to decrease inequalities in BCS attendance (Molina-Barceló et al., 2021). Our analysis strengthens the importance of developing interventions to ensure access to preventive screenings for all.

The second intersectional subgroup with the lowest BCS attendance rates were women living in Ceuta and Melilla. These regions were the last regions to implement OSP (Carmona-Torres et al., 2018) and they have the lowest density of health professionals (2.5/1,000 inhabitants) in Spain (Ministerio de Sanidad, 2021).

6.1.3 Inequalities in colorectal cancer screening attendance in Sweden

Inequalities in CRC screening attendance among people aged 60-69 in Sweden were explored in *manuscript III*. The results showed that sociodemographic inequalities prevailed more in opportunistic screening regions than in organised screening regions. Only age differences were identified in regions with ongoing OSP: people aged 60-64 had twice the risk of never attending CRC screening than those aged 65-69 (reference group). Besides, this difference could, at least partly, be explained by the number of screening invitations received over their lifetime. Similar age differences in lifetime cancer screening attendance in Sweden are found among other organised programs (Lagerlund et al., 2021). This additional evidence reinforces that organised CRC screening can contribute to equal access for all.

The findings also demonstrated that in regions with opportunistic screening, individuals born in the EU exhibited the lowest CRC screening attendance rates. Indeed, the three intersectional subgroups at the highest risk of never attending screenings included men and women born in the EU. Specifically, EU-born women in the second income quintile exhibited the highest risk of never attending CRC, which was eight times greater than the reference group. While these results contrast several studies examining CRC screening attendance in other high-income countries (Pallesen et al., 2021; van de Schootbrugge-Vandermeer et al., 2023), they are consistent with findings from other research conducted in Sweden (Strömberg et al., 2022a; Strömberg et al., 2022b).

The recent implementation of CRC OSPs in all 21 healthcare regions is encouraging for reducing long-term CRC inequalities (Regionala Cancercentrum Samverkant, 2022). However, close attention should be directed to at-risk intersectional subgroups, including EU-born men and women and non-EU-born women.

6.1.4 Regional inequalities across the three case studies

The three empirical studies were the first in their respective contexts to assess inequalities at the intersections of individual- and regional- levels of inequality. These are inequalities that would not have emerged else non-linear methods such as decision trees would have been applied.

In Spain, *manuscript IV* revealed strong inequalities in BCS attendance across regions. Ceuta and Melilla showed the lowest attendance rates, and several examples of individual-regional intersections emerged. The most relevant is the following: whereas women in higher social classes tend to attend BCS more than women in low and middle social classes (Mottram et al., 2021), our analysis showed that women born in Spain, married, divorced, or widowed, belonging to middle or low social classes living in Navarre, Aragon, Canary Islands, and Catalonia had a lower predicted prevalence of never attending to BCS than those belonging to high social classes living in any region but Ceuta or Melilla. This entails that Navarre, Aragon, Canary Islands, and Catalonia are strongly protective factors for BCS attendance, potentially outweighing social dimensions such as social class. It is noteworthy to mention that, except for the Canary Islands, these regions are, on the one hand, among the wealthiest in the country (INE, 2023) and, on the other hand, developed programs to fight inequalities in BCS attendance (Molina-Barceló et al., 2021). Again, this reinforces the responsibility of public institutions to offer everyone the same opportunities to access BCS.

In Germany, the decision-tree approach from *manuscript II* identified the region of residence as an important variable. Specifically, the decision tree split twice in the third node by region of residence, supporting the necessity of exploring individual- and regional- intersections. In both splits, women residing in Bavaria, Bremen, Hamburg, and Saarland - specifically women living with a partner who are either employed, unemployed, unable to work, or in other working categories and widowed women living alone, with children, with a partner and children or other arrangement - depicted a higher risk of never attending BCS than women in that intersectional position from all other regions.

Finally, in Sweden, *manuscript III* underlined stark inequalities in CRC screening attendance based on residing in a region with organised or opportunistic screening. Indeed, only age-related inequalities were identified in regions with implemented OSPs, whereas inequalities in the intersections of gender, migration background and income were identified in regions with opportunistic screening. Examining the intersection of individual and regional variables within decentralised health systems significantly bolsters the understanding of inequalities in cancer screening attendance.

6.1.5 Migration background: unexpected findings

A further commonality across the case studies is the minimal or inverse impact of migration background on cancer screening uptake. Contrary to other studies conducted in the EU, which emphasised the significant role of migration background in predicting BCS attendance (Ding et al., 2021; Rondet et al., 2014), or CRC screening attendance (Pallesen et al., 2021; van de Schootbrugge-Vandermeer et al., 2023), the four manuscripts of this thesis revealed distinct findings.

In Germany, the scoping review identified having a migration background as a protective factor for attending BCS screening. Nonetheless, the evidence-based approach of *manuscript II* identified low-income women not born in Germany, residing in rural areas cohabitating with their partner as those at the lowest risk of never attending BCS, but low-income women not born in Germany, residing in rural areas not cohabitating with their partner as the highest risk of never attending BCS. As described above, for women not born in Germany with low income and residing in rural areas, partnership cohabitation plays a pivotal role. The decision tree-based approach did not recognise migration background as a relevant social dimension.

In Spain, the ensemble algorithm did not identify migration background as a relevant predictor for predicting BCS attendance and the decision tree did not split based on this predictor.

In Sweden, the association between migration background and CRC screening attendance exhibited an inverse relationship. For people residing in regions with opportunistic screening, those not born in the EU had lower risks of never attending CRC screening than those born in the EU.

Unexpected findings in cancer screening attendance, like those presented above, would not have emerged without adopting a conceptual framework such as intersectionality and a non-linear and non-parametric methodological approach such as decision trees.

6.2 Methodological implications on the use of decision trees for subgroup identification

The methodological implications of this dissertation extend beyond the specific contributions of each research objective, building the ground for debate on the overall research aim. First, the methodological implications of each research objective are discussed, second, the way these implications contribute to the broad thesis aim is clarified, third, the strengths and limitations of the use of decision trees for subgroup identification are presented; and, fourth, the advantages and disadvantages of using optimisation techniques for building explorative decision trees are discussed.

As a brief reminder, the research objectives of this thesis were the following: 1) to compare a decision tree-based approach and an evidence-informed approach for identifying intersectional subgroups, 2) to explore the use of decision trees for reducing intersectional complexity, and 3) to explore decision trees as predictive tools for intersectional subgroup identification.

6.2.1 Decision trees compared to an evidence-informed approach

Evidence-informed approaches synthesise existing research to identify relevant social dimensions to be included in an analysis. This process makes an inherently normative decision on which social dimensions to explore. Then, the dimensions of inequality examined will be based on previous research and will contribute to the existing body of literature. Here, there is a risk of overrepresenting certain social dimensions, resulting in the stigmatisation of specific groups and the neglect of others (North & Fiske, 2014; Turan et al., 2019). Furthermore, recommendations for interventions are limited to the included social dimensions, raising concerns about potential biases in variable selection and subsequent recommendations. In *manuscript II*, the evidence-informed approach identified low-income women not born in Germany, residing in rural areas not cohabitating with their partner as the highest risk of never attending BCS. However, the decision tree-based approach did not identify migration background as a relevant variable, but the region of residence.

In contrast, decision tree-based approaches use statistical algorithms to inductively identify patterns within datasets (Venkatasubramaniam et al., 2017). This method offers the advantage of revealing previously unexplored combinations of social dimensions, enabling more targeted interventions. For instance, the decision tree in *manuscript II* helped identify regional disparities among specific intersectional subgroups of women that the evidence-informed approach or traditional linear statistical methods would have missed.

6.2.2 Decision trees as intersectional complexity reduction tool

One of the most recurrent problems in quantitative intersectionality is the complexity of higher-dimensional interactions. Here, large matrices of intersectional subgroups (i.e. full cross-classification) with small sample sizes present statistical problems, such as small statistical power of prediction, loss of information and decrease of DA of the model (for more detailed information, Appendix 1 of *manuscript III* illustrates the concave relationship between increased number of intersectional groups and AUC).

To address full cross-classification matrices with small cell sizes, sociodemographic variables are often dichotomised (Bauer, 2014; McCall, 2005). This approach necessitates a rigid categorisation, which, intersectionally speaking, drifts the inter-categorical approach even more apart from the anti-categorical approach. Moreover, the dichotomisation imposes a linear relationship between the categories (1 vs 0, more vs less), depriving the analysis of nuanced and detailed content, such as the ability to include middle-income positions (McCall, 2005). In these circumstances, the quantitative exploration of intersectional complexities limits the theoretical and empirical depth of the analysis.

Decision trees can help to reduce the complexity of large intersectional matrices, without using rigid categorisations and basing the process and choice of model on its DA. Using all available categories from the included variables (i.e. social dimensions) allows the algorithm to detect non-linear effects (e.g. quadratic or 1-0-1 patterns) where, for instance, low and high categories show effects, but mid-range not. *Manuscript III* illustrates a good example of non-linear effects for men not born in the EU living in opportunistic screening regions: those belonging to the two highest or the lowest income quintile had no statistically significant difference in attending CRC screening compared to the reference group. However, those belonging to the middle-income quintiles revealed a five-fold risk compared to the reference group. This nuanced analysis is only possible when including all categories available from the predictors and allowing for hierarchical non-linear interactions of the categories.

Furthermore, decision trees perform well with small sample sizes even if many covariates are included in the model (Henrard et al., 2015). In contrast, in multiple regression analysis (both linear and logistic), a minimum number of observations is necessary in order to have enough statistical power to compute estimations, especially as more covariates are incorporated into the model (Mason & Perreault, 1991). Indeed, smaller sample sizes can lead to an increased risk of type I error, that is, not detecting a significant relationship of certain covariates with the outcome given the underpowered statistical inferences (Speed, 1994).

Additionally, the outcome of decision trees, which produces fewer intersectional groups than a full intersectional matrix, when used as an exposure variable in a regression analysis, enhances the statistical power and improves the interpretability of regression results. When left with a large full intersectional matrix, a common concern among quantitative intersectional scholars is to decide which effect sizes to report and which not. Reducing the number of intersectional subgroups can thus enhance the explainability of the model.

6.2.3 Decision trees as predictive tools for intersectional subgroup identification

Using decision trees as predictive tools implies different emphases and goals than when employed as explorative tools. Whereas explorative/inference approaches focus on understanding the relationships between variables and the outcome, predictive approaches prioritise building models that accurately predict outcomes on unseen data by optimising performance measures like accuracy, sensitivity, or specificity (Flaxman & Vos, 2018). This means that the models will be built to optimise the prediction of unknown data rather than to be explainable and understandable (Bzdok et al., 2018).

A good example is the decision tree in *manuscript IV.* Since the main goal was to achieve high performance, pre-processing steps were taken (i.e. oversampling) on the training data. Then, in the hypertuning process, no restrictive criteria were applied so that the tree could optimise predictive performance at all costs. A decision tree was built with very high performance, yet its size and leaves composition (i.e. different dimensions of intersectional interactions) are more difficult to interpret. Here, again, the goal of the decision tree is relevant, and if this is a high predictive power, sacrificing interpretability might be accepted. To make predictions more robust and decrease the risk of overfitting, decision trees are often built into ensembles (e.g. random forest, extreme gradient boosting). These algorithms provide information on the importance of the included predictors for predicting the outcome but not on the structure of the decision trees (i.e. no intersectional subgroups) (Loyola-Gonzalez, 2019).

6.2.4 Advantages and disadvantages of decision trees for subgroup identification

Decision trees have accurately identified at-risk populations based on sociodemographic factors in numerous public health studies (Batterham et al., 2009; Battista et al., 2023; Freitas et al., 2012; Seeley et al., 2009), including those in this dissertation.

Decision trees provide a non-parametric assumption about the distribution of the outcome or the explanatory variables and no assumption of linearity regarding the relationships between predictors and the outcome (Venkatasubramaniam et al., 2017). This contrasts with linear regressions, which assume a normal distribution of the errors and each exploratory variable to be linearly associated with the outcome, or logistic regression, which assumes each explanatory variable to be linearly associated with the log odds of the outcome variable (Henrard et al., 2015).

Furthermore, unlike linear or logistic regression, where including numerous potential predictors may result in multicollinearity, decision trees enable the incorporation of multiple variables (i.e. PROGRESS-Plus, structural determinants of CSDH or any set of pre-selected variables in the dataset). Decision trees are not affected by multicollinearity, given their hierarchical splitting methods. That is, the decision tree independently selects the most relevant predictor based on a specific splitting rule (i.e. different for every algorithm) for every current split. Consequently, highly correlated predictors do not generate instability or bias like linear or logistic regression models (Henrard et al., 2015).

Therefore, decision trees do not limit the number of explanatory variables that can be examined. On the contrary, they make a data-driven selection of predictors. Each algorithm decides which variables have a greater role (e.g. stronger statistical association for CIT, larger gini impurity reduction for CART, and larger entropy reduction for C5.0) in the process of building the classification model (Novaković, 2016). However, Lemon (2003) cautioned against "data dredging" by including all possible variables without careful consideration. They advocated for a thorough approach to selecting predictors to avoid spurious results (Lemon et al., 2003).

Decision trees also have limitations that affect their ability to identify intersectional subgroups. First, although their hierarchical nature is an advantage regarding multicollinearity, it can bring instability to the model since the initial split significantly conditions the following (Apté & Weiss, 1997). In other words, if two variables have a similar impact on the outcome (i.e. for CIT) or contribute equally to reducing the model's entropy (e.g. for CART and C5.0), there may be competition over which of these variables to select, and this choice will significantly impact the final model (Venkatasubramaniam et al., 2017).

Secondly, decision trees are highly sensitive to data quality when noise or outliers are present. Consequently, they can create splits that capture these anomalies instead of the true underlying patterns, reducing the generalisability of the model (Battista et al., 2023). This phenomenon is called overfitting the data, and it hinders the bias-variance trade-off (i.e. models that are neither overly complex (high variance) nor overly simple (high bias)) by developing models with low bias and high variance (Bramer, 2007). It is worth noting that linear and logistic regression models, due to their strong assumptions, also unbalance the bias-variance trade-off, but in the opposite way: showing high bias and low variance (Dietterich, 1995).

Decision trees handle data overfitting through several strategies such as pruning (CART initially grows a larger tree and then prunes it back based on the complexity parameter (Breiman et al., 2017), and C5.0 prunes every branch based on the binomial confidence limit method (Quinlan, 1993)), limiting the depth of the tree (CIT limits the growth of the tree based on statistical significant measures (Hothorn et al., 2006)) or applying cross-validation techniques (e.g. k-fold cross-validation). For predictive decision trees, validating the model developed on unseen data prevents performance overestimation and model overfitting (Lones, 2021). Furthermore, decision trees ensembles reduce overfitting by aggregating the predictions of many decision trees (i.e. bagging – e.g. random forest (Breiman, 2001a)) or by applying more complicated methods such as sequentially building trees that correct the errors of the previous ones, optimising the loss function by gradient descent (i.e. gradient boosting– e.g. XGBoost (Chen & Guestrin, 2016)) (Breiman, 1996).

6.3.5. Advantages and disadvantages of using optimisation techniques for building explorative decision trees

Heuristic approaches, with a rule of thumb, are designed not to overfit the data but still be close to having the minimum possible error. Consequently, in algorithms such as CART, trees with the smallest possible error are not chosen, but the next simpler one whose error is within one standard error above the tree with the smallest error. This measure ensures then low error and a certain level of generalisation. Yet, this pre-set rule does not ensure the optimal classification of cases or overall accuracy. Using parameter hypertuning for building explorative decision trees allows us to optimise accuracy or other performance measures. However, it also has its own drawbacks. Below, the advantages and disadvantages of optimisation techniques for building explorative decision trees, that is, using predictive tools for non-predictive tasks, are presented.

The main advantage is that the decision tree is grown based on the optimising criteria we care for. For example, in the case of identifying subgroups of women who never attended BCS in Germany (*manuscript II*), we are more interested in building a model that optimises the identification of those who never attended BCS (sensitivity) rather than those who attend BCS (specificity). Therefore, we will build a model that optimises sensitivity. To illustrate, if we want to have a certain number of intersectional subgroups (e.g. between 6-10), we can set a maximum depth of the tree and ask to find those subgroups optimised for detecting true positive cases. Or, if we are interested in the general AUC of the model, we can ask the model to build subgroups that optimally balance sensitivity and specificity. Furthermore, in the *mlr3*, models can be built considering the optimisation of two criteria simultaneously (e.g. AUC and sensitivity).

For the explorative purposes of decision trees in *manuscripts II* and *III*, where decision trees are built as a previous step for regression analysis, we foresee potential in using this approach. However, it is important to mention the limitations on the generalisability and the over-positive nature of the results that need to be reported. Evaluating the performance measures of the decision tree in the same dataset where the decision tree has been built would be a no-go in machine learning. This outputs over-optimistic values, which do not refer to unseen data; hence, the decision tree results are not generalisable to unknown data. More on generalisability in section 6.6 of this Chapter.

6.3 Implications of the findings for public health research

Through conceptual and methodological analysis, this dissertation has proposed several analytical strategies for using decision trees to identify at-risk intersectional subgroups. The adoption of the intersectional lens, as already presented in Chapter 1, facilitates answering the classical research question "Who is at higher risk of an outcome?" not only by outlining the individual effects of each included social dimension but also by accounting for the effects of being at the intersection of several disadvantaged social positions. Then, using the analytical strategies proposed in this dissertation (Figure 3), a door is opened to ask new research questions to address health inequalities in public health.

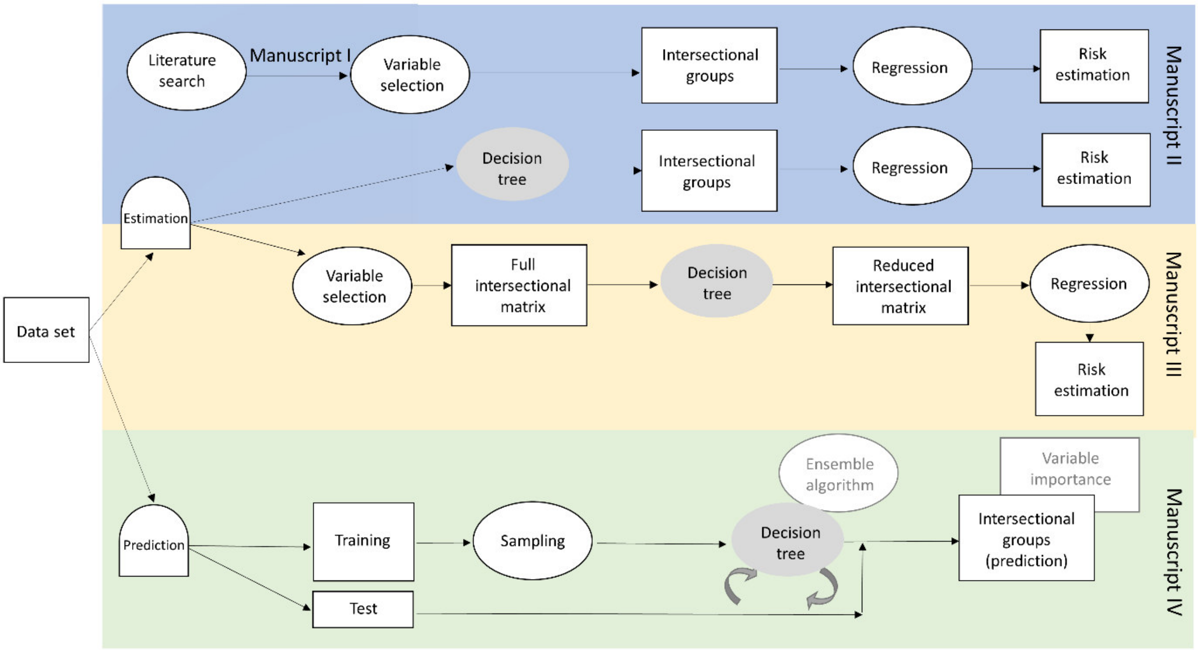


Figure 3. Analytical strategies employing decision trees used in the dissertation

If we aim at estimating who is at higher risk of an outcome (intersectionally or not) through a regression model, we have to decide upon which variables (i.e. social dimensions) to include. Often an evidence-informed decision is taken, where social dimensions are selected based on existing evidence in the literature. In this case, and regarding the research outcome of this thesis, the research question we are answering is: *Who is at higher risk of never attending cancer screening based on certain already known social dimensions?*

In the absence of adequate theoretical or empirical foundations, or if no social dimensions are to be excluded a priori, data-driven methods, such as decision trees, can be employed to select the relevant variables. Then, a different research question is answered: *Who is at higher risk of never attending cancer screening based on currently unknown dimensions revealed through data-driven approaches?*

In *manuscript II,* the risk of never attending BCS in Germany was estimated through an evidence-informed and data-driven (decision tree-based) approach. On the one hand, the evidence-informed approach answered the first question by systematically reviewing the available literature (*manuscript I*) and identifying the social dimensions (and their direction) more relevant for assessing BCS attendance in Germany. Then, these social dimensions were employed to create an intersectional matrix combining the dichotomised variables. Finally, a logistic regression was performed on the intersectional matrix to estimate the effect sizes of the evidence-informed regression. On the other hand, the decision tree-based approach answered the second question. It did so by identifying intersectional subgroups through a decision tree built with available PROGRESS-Plus characteristics. These identified subgroups were then used as exposure variables for logistic regression to estimate the risk of never attending BCS.

Manuscript III also answered the question *Who is at higher risk of never attending cancer screening based on currently unknown dimensions revealed through data-driven approaches?* Although decision trees were used to reduce the complexity of a full intersectional matrix, this reduction implied a selection of relevant variables for constructing the reduced intersectional matrix (i.e. final intersectional subgroups).

If we shift our focus from explorative to predictive models, decision trees enable us to address practical questions relevant to medical decision-making. Whereas these models do not estimate inferences, they can assist practitioners in anticipating whether a patient will enrol in preventive actions. In other words, employing decision trees within a predictive manner can help answer questions like: *Will the next person I meet in my practice attend cancer screening?*

Manuscript IV developed a predictive model using decision trees and decision tree ensembles to accurately predict whether a woman in the target age group in Spain will undergo BCS based on her individual social dimensions and region of residence.

The research on health inequalities in public health can benefit from widening the range of asked research questions and applying novel methods to understand these inequalities and the underlying mechanisms.

6.4 Implications for practice and policy

The ability of decision trees to identify at-risk groups is particularly valuable for public health practitioners facing resource constraints, as it allows them to effectively target prevention and intervention efforts. Moreover, decision trees' results are easy to understand for non-technical audiences, facilitating knowledge translation with diverse stakeholders (Battista et al., 2023).

The studies in this dissertation highlight the importance of implementing organised population screening programs, where people at higher risk of developing cancer are systematically invited for early detection. In countries with a decentralised health system, financial and human resources must be fairly allocated so that the accessibility to cancer screening services does not depend on the program's implementation in the region of residence.

In Sweden, all healthcare regions have recently implemented OSP for CRC screening. Monitoring of inequalities in the newly organised screening regions is essential, and special attention should be given to high-risk groups, such as men and women born in the EU, as well as women not born in the EU.

In Spain, special attention must be paid to regions where OSP for BCS was belatedly implemented, such as Ceuta and Melilla. Furthermore, incentives are needed to increase the number of healthcare professionals in these regions, as their density is nearly half that of wealthier regions. Finally, all regions should implement interventions to reduce social inequalities in BCS, focusing on the identified high-risk groups.

In Germany, household composition was identified as a key factor influencing BCS attendance, with cohabitation with a partner acting as a protective factor. Several studies have built upon the idea that cohabitation with a partner is particularly relevant for increasing preventive health behaviours (Gram et al., 2021; Han et al., 2019; Hanske et al., 2016; Manjer et al., 2015). Indeed, intimate relationships are key to increasing an individual's social support, and this, for example, enhances individual self-efficacy, which may help overcome barriers (e.g. fear) to cancer screening (Gallant, 2013; Von Wagner et al., 2011). In the absence of such a living arrangement, other measures to increase levels of social support among women could be developed, especially for women not cohabitating with their partners residing in Bavaria, Bremen, Hamburg, and Saarland.

6.5 Strengths and limitations of this dissertation

This dissertation has several strengths and also limitations.

The exploration of explorative and predictive decision trees helped to shed light on diverse applicability through the above-presented analytical strategies. The present dissertation is the first academic work that extensively explores the manifold use of decision trees for subgroup identification and comprehensively debates its utilisation.

Furthermore, as discussed in section 6.2.2, applying decision trees for full matrix reduction, aka complexity reduction, is a notable contribution to the ongoing intersectionality discussions and the need for dealing with high-dimensional interactions (Bauer et al., 2021; Mahendran et al., 2022).

In addition, the structure and results of decision trees are easy for a wider audience to understand, thereby improving scientific and practical knowledge transfer (Battista et al., 2023). For example, the decision tree in *manuscript III,* illustrating inequalities in CRC screening attendance in Sweden, clearly visualises sociodemographic inequalities in regions with and without OSP. Indeed, public health practitioners could benefit from using decision trees to communicate inequalities expertise and develop tailored interventions.

The limitations of this dissertation range from intrinsic shortcomings of the data used both for the scoping review and the empirical papers to the (intersectional) interpretation of the results. The scoping review (*manuscript I*) combines very heterogeneous studies, which might have led to mixed findings. The empirical studies (*manuscripts II-IV*) employed the EHIS survey, which features a cross-sectional design; hence, we cannot establish causality. Furthermore, attendance to breast (*manuscript II* and *manuscript IV*) or colorectal (*manuscript III*) cancer screening was a self-reported outcome potentially inducing recall bias that over- or under-estimates attendance. In addition, the response rate for the three surveys ranged from 21.6% *(manuscript II*) to 59.0% (*manuscript IV*), underscoring the necessity for caution when drawing conclusions.

This dissertation aimed to advance the study of individual and regional–level inequalities in accessing cancer screening in three case studies. To this end, the region of residence of participants was considered in the models. However, the decision trees did not account for the hierarchical nature of the data, whereby participants were clustered within regions; rather, regions were handled as individual-level variables. More complex decision tree algorithms, such as RE-EM (168) and M-CART (169), have been developed to account for multilevel data structures and could be examined in future research.

In quantitative intersectionality research, several authors have pointed out the need to interpret the results of the studies in the light of social injustice and equity, pointing out present and historical systems of oppression and discrimination (Bauer et al., 2021; Bowleg, 2012). The present dissertation has concentrated on methodological discussions rather than examining the existing hegemonic structures that might have caused the highlighted inequalities. If we want to examine the whole intersectional research cycle and interpret the results considering the contextual systems of oppression, we would need information about the underlying mechanisms of non-attendance at cancer screening. Only by investigating the reasons for non-attendance can we discuss how systems of oppression and discrimination affect people at the intersection of multiple disadvantaged positions.

6.6 Excurses on generalisability

Since the present dissertation lies at the edge of two different methodological disciplines, estimative and predictive models, and the employed methodologies, decision trees and regression analysis, and their interpretations are also used in both domains, I considered it relevant to write an excurses on the concept of generalisability.

The concept of generalisability of the results of the models in this dissertation can cause confusion or misunderstandings. On the one hand, in predictive models, when a classification model has an accuracy of 70% (i.e. tested on the test data set), this percentage informs how well the model will classify new unseen data. In this sense, the model is generalisable to unknown data. This would be the case of the decision tree and the decision tree ensemble presented in *manuscript IV*.

In the case of explorative decision trees, since the model is not tested on unseen data but on the same training data, the performance measures do not entail generalisability to unknown data; they just refer to the data used for training the model. Therefore, we do not know how well that subgrouping of data would perform outside the survey observations.

Following this same line of thought, an evidence-informed decision for building intersectional groups is only as good as the evidence is based on. That means that if the surveys used in other studies are not representative of the entire population, the intersectional subgroups which are evidence-inform are also not. In the case of *manuscript II*, the evidence-informed approach was based on a scoping review (*manuscript I*) that included heterogeneous data collection methods (e.g. self-reported, claims data (i.e. health insurance) and register-based data (mammography centres)) and study designs (e.g. cross-sectional and cohort studies) (more information available in Supplementary File 4 from *manuscript I*). Therefore, one could question herself what is better for the later use for calculating generalisable estimations: intersectional subgroups based entirely on one survey data or intersectional subgroups based on heterogeneous studies not always carried out on representative sampling. This question presents a hard challenge: the evidence-informed approach is limited by the available data potentially not good enough for ones own own data, and the decision trees approach is solely tailored to your data, potentially overrepresenting it.

On the other hand, in estimative models (e.g. regression analysis), estimations of effect sizes are considered generalisable based on the data quality. If the data is drawn from a representative sample (i.e. national), results are generalisable to that specific country. That is the case, especially if the survey data is solely based on random sampling, which probably is true only in an ideal world. In reality, population sampling is not completely random, and not everyone invited to participate in a survey does so (i.e. response rate). To deal with this potential lack of representativity of certain population groups, sampling weights are calculated; these are, however, solely based on certain social dimensions. For example, in the German third wave of EHIS, sampling weights are calculated based on sex, education, age, structure of the regions, and structure between urban and rural areas (Lange et al., 2017). Since the last four variables are included in the analysis, sampling weights could have led to multicollinearity and biased standard error estimation (sensitivity analyses provided in Appendix A in *manuscript III*). In most epidemiological studies, whether sampling weights are applied or not, estimation of the effect sizes are interpreted as if they would be generalisable to the population from which the representative national survey is taken. This is how the results from *manuscript II* and *manuscript III* are interpreted, transferring the learned estimation of the effect sizes to Germany and Sweden, respectively.

To sum up, the performance measures presented from the decision tree and the decision tree ensemble in *manuscript IV* are generalisable to the entire population in Spain since this performance was measured on unseen data. However, the performance measures of the identified subgroups from the decision trees in *manuscripts II* and *III* refer solely to the dataset's observations. Nevertheless, the risk estimates from the regression analyses presented in these two manuscripts are generalisable to their pertinent national population through the representative sampling of the national surveys.

Chapter 7. Future research outlook

In this chapter, some research possibilities that emerged from the different analytical strategies employed throughout the dissertation are suggested.

A promising tool for building predictive decision trees is the *mlr3* ecosystem in R presented in this dissertation. This model optimisation tool offers powerful and complex approaches to building predictive models, from which predictive decision trees could benefit. These processes entail the construction of graphs with pipelines that incorporate pre-processing steps and non-sequential processes (Binder & Pfisterer, 2024). Pre-processing steps are any changes performed to the data before it is used to build a model, such as oversampling to deal with imbalanced datasets, as performed in *manuscript IV* (Thomas, 2024). Once the data is pre-processed, the non-sequential process allows combining several learners differently. Branching is the process of forking potential learners and choosing the one performing best in building the model. Stacking (Zenko et al., 2001) combines the predictions (i.e. feature union) from several individual learners (i.e. level 0 models) and uses them as parameters for a subsequent model (i.e. level 1 model) (Binder et al., 2024). This powerful ensemble method can significantly improve prediction outcomes and could be a compelling approach to improve the predictive performance of predictive decision trees (Wolpert, 1992).

Explorative decision trees performed in *manuscripts II* and *III* open the door to exploring the data in multifaceted ways which could be adapted to the specific needs of public health practitioners. For example, if a specific social dimension is of interest (e.g. educational level), stratified decision trees could be developed to see how other social dimensions hierarchically unfold for each educational group. A step even further could be to explore the hierarchical interaction of solution-linked variables for each educational group instead of social dimensions. These variables, essentially classical mediators in public health, are defined as variables capturing the underlying process of marginalisation through societal and contextual elements that reflect unequal power relations (O'Campo & Dunn, 2012). Solution-linked variables include main earner status, burden due to care, perceived social support, etc. Stratified decision trees based on social positions (e.g. educational groups) or even intersectional groups (e.g. gender-education intersections) that explore the hierarchical relation between solution-linked variables could aid in understanding non-linear relationships in processes of marginalisation for each social position/intersectional group. Interestingly, this approach would bring us closer to adopting an anti-categorical intersectional approach.

Following a more standard approach for examining underlying mechanisms, another future research step could be exploring the mechanisms underlying inequalities in cancer screening through mediation analysis. Bauer & Scheim (2019) proposed applying a moderated mediation to account for the heterogeneous effects on the health outcome across the intersectional groups when testing for potential mechanisms explaining the observed inequalities (Bauer & Scheim, 2019).

Additionally, qualitative interviews could be carried out to deepen insight into the reasons for low cancer screening attendance rates for certain intersectional groups. The interviews could focus on gaining insights on the experienced barriers and challenges faced by intersectional groups at risk of not attending cancer screening, to guide developing targeted interventions that address their unique experiences.

In future research, improving the quality of the collected data is important. To perform analyses that more closely represent the experiences of an entire population, it is necessary to enhance the available categorisation of existing social dimensions. For example, in *manuscript III,* the variable sex was used as a proxy for gender. Given that gender is a social construct and sex is a biological construct (Krieger, 2003), it only makes sense to assess unfair inequalities in CRC screening attendance based on the social construct. However, inferring female - women and male - men is a stark postulation that oversees anyone on the non-binary or trans spectrum (aside from totally disregarding intersex people) (Fausto-Sterling, 2008). The dichotomisation of sex/gender not only neglects the existence of some but also simplifies the experience of the entire category. Here, it is assumed that everyone in the same category shares some common characteristics (sex - e.g. genitalia and hormones; gender – e.g. societal norms and expectations) and interventions are developed accordingly. To overcome this simplistic view of sex/gender, many authors have proposed alternative operationalisations and theoretical frames (for a review of these publications, see (Bolte et al., 2021)). Bolte et al. (2021) proposed an overarching conceptual framework to operationalise the sex/gender dimension, incorporating several relevant conceptual criteria such as multidimensionality, variety, embodiment, and intersectionality (Bolte et al., 2021).

In the present dissertation, the empirical research carried out in three different countries brought to light the urgent need for research on the taxonomy of intersectional groups and intersectional processes. Although the processes of discrimination suffered by certain intersectional groups are specific to their context, it is necessary to find common grounds to discuss discrimination for a specific intersectional group across contexts. Here, it is crucial to recognise that systems of oppression vary across contexts, leading to different levels of discrimination within the same intersectional group (e.g. experiences of heterosexism might be different for lesbian women in Sweden than in Saudi Arabia) (Evans, 2019). Therefore, first, we need to acknowledge the interaction between intersectional groups and context; second, we need to routinely and systematically gather information on diversity and discrimination experiences across contexts (as proposed by (Stadler et al., 2023)); and third, we need to understand and interpret the meaning of the dimensions of discrimination within each specific context.

Last but not least, a further research suggestion would be to perform periodic health reporting adopting an intersectional perspective besides collecting enough data on relevant and multidimensional social dimensions. Here, methods such as AIHDA (as proposed in (Merlo et al., 2023)) or explorative decision trees could aid the repeated examination of intersectional dimensions of inequality.

Chapter 8. Conclusion

This dissertation explored intersectional inequalities in accessing cancer screening appointments in three European countries: Germany, Sweden and Spain. The three empirical studies were the first in their respective contexts to assess inequalities at the intersections of individual and regional inequality levels. Moreover, the three case studies identified strong regional and unexpected migration background effects. In Spain, regions played a significant role in predicting BCS attendance, reflecting the OSP implementation timeline and the economic inequalities between regions. In fact, the region of residence has acted as a protective factor when intersecting with other variables and as a risk factor for the most economically disadvantaged regions. In Germany, cohabitating with a partner was a protective factor for BCS attendance, and women residing in specific regions were at higher risk of never attending BCS than women in the same intersectional group from other regions. In Sweden, regions played a role so that only age-related inequalities in CRC screening were identified for those with implemented organised screening programs. In contrast, in regions with opportunistic screening, inequalities in the intersections of gender, migration background and income were identified.

The research conducted in the scheme of this dissertation highlighted the importance of adopting an intersectional lens when assessing those who do not attend cancer screening. Furthermore, practical efforts to reach these intersectional subgroups are of utmost importance. In the case of decentralised health systems, resources must be fairly allocated so that the accessibility to cancer screening services does not depend on the program's implementation in the region of residence.

Methodologically, the present dissertation has meaningfully contributed to advancing the identification of intersectional subgroups at higher risk of never attending cancer screening appointments. First, it has proposed a method to reduce complex intersectional matrices, often highlighted as a statistical constraint for applying methods such as AIHDA. Moreover, it has employed the same measure of individual heterogeneity as the already established AIHDA: the AUC. Second, decision trees have not only increased the statistical power of the analysed intersectional matrix in regression analysis but also improved the interpretability of the results. Third, since decision trees perform well with small sample sizes, no rigid categorisation (e.g. dichotomisation) is necessary to reduce the number of analysed intersectional groups. Consequently, non-linear interactions of categories across intersectional groups can arise (e.g.

0, 1, 0), and nuanced and unexpected hierarchical interactions emerge. Fourth, this dissertation has technically proposed and discussed optimisation techniques from predictive modelling for growing explorative decision trees.

Consequently, this dissertation has proposed different analytical strategies using decision trees within the framework of intersectionality that opened a door to ask new research questions to address health inequalities in public health. When using explorative decision trees as a prior step to identifying intersectional subgroups to use them as exposure variables in a regression analysis, the following question is answered: *Who is at higher risk of never attending cancer screening based on currently unknown dimensions revealed through data-driven approaches?* Applying predictive decision trees allows us to ask questions anticipating a future event: *Will the next person I meet in my practice attend cancer screening?*

On a practical level, this dissertation's implications include the evidence of decision trees' ability to identify at-risk groups by assessing a comprehensive set of individual and regional indicators - particularly valuable for public health practitioners. Furthermore, the results of decision trees are easy to understand for non-technical audiences, facilitating knowledge translation with diverse stakeholders.

In a nutshell, the present dissertation has made significant contributions to understanding inequalities in cancer screening in three different European countries to the methodological field of quantitative intersectionality, and it has proposed new avenues for questioning relevant public health questions, offering a new methodological tool for public health practitioners.

References

Ackerson, K., & Preston, S. D. (2009). A decision theory perspective on why women do or do not decide to have cancer screening: Systematic review [Review]. *Journal of Advanced Nursing*, *65*(6), 1130-1140. https://doi.org/10.1111/j.1365-2648.2009.04981.x

Anell, A., Glenngård, A. H., Merkur, S., & Organization, W. H. (2012). Sweden: Health system review.

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems*, *13*(2), 197-210. https://doi.org/https://doi.org/10.1016/S0167-739X(97)00021-6

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19-32. https://doi.org/10.1080/1364557032000119616

Armaroli, P., Villain, P., Suonio, E., Almonte, M., Anttila, A., Atkin, W. S., Dean, P. B., de Koning, H. J., Dillner, L., Herrero, R., Kuipers, E. J., Lansdorp-Vogelaar, I., Minozzi, S., Paci, E., Regula, J., Törnberg, S., & Segnan, N. (2015). European Code against Cancer, 4th Edition: Cancer screening. *Cancer Epidemiol*, *39 Suppl 1*, S139-152. https://doi.org/10.1016/j.canep.2015.10.021

Axelsson Fisk, S., Lindström, M., Perez-Vicente, R., & Merlo, J. (2021). Understanding the complexity of socioeconomic disparities in smoking prevalence in Sweden: a cross-sectional study applying intersectionality theory. *BMJ Open*, *11*(2), e042323. https://doi.org/10.1136/bmjopen-2020-042323

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, *12*(4), 387-415.

Banerjee, M., Reynolds, E., Andersson, H., & Nallamothu, B. (2019). Tree-Based Analysis: A Practical Approach to Create Clinical Decision-Making Tools. *Circulation: Cardiovascular Quality and Outcomes*, *12*. https://doi.org/10.1161/CIRCOUTCOMES.118.004879

Batterham, P. J., Christensen, H., & Mackinnon, A. J. (2009). Modifiable risk factors predicting major depressive disorder at four year follow-up: a decision tree approach. *BMC Psychiatry*, *9*(1), 75. https://doi.org/10.1186/1471-244X-9-75

Battista, K., Diao, L., Patte, K. A., Dubin, J. A., & Leatherdale, S. T. (2023). Examining the use of decision trees in population health surveillance research: an application to youth mental health survey data in the COMPASS study. *Health Promot Chronic Dis Prev Can*, *43*(2), 73-86. https://doi.org/10.24095/hpcdp.43.2.03 (Utilisation des arbres décisionnels dans la recherche en surveillance de la santé de la population : application aux données d'enquête sur la santé mentale des jeunes de l'étude COMPASS.)

Bauer, G. R. (2014). Incorporating intersectionality theory into population health research methodology: challenges and the potential to advance health equity. *Soc Sci Med*, *110*, 10-17. https://doi.org/10.1016/j.socscimed.2014.03.022

Bauer, G. R., Churchill, S. M., Mahendran, M., Walwyn, C., Lizotte, D., & Villa-Rueda, A. A. (2021). Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM Popul Health*, *14*, 100798. https://doi.org/10.1016/j.ssmph.2021.100798

Bauer, G. R., Mahendran, M., Walwyn, C., & Shokoohi, M. (2022). Latent variable and clustering methods in intersectionality research: systematic review of methods applications. *Social Psychiatry and Psychiatric Epidemiology*, *57*(2), 221-237. https://doi.org/10.1007/s00127-021-02195-6

Bauer, G. R., & Scheim, A. I. (2019). Methods for analytic intercategorical intersectionality in quantitative research: Discrimination as a mediator of health inequalities. *Social Science & Medicine*, *226*, 236-245. https://doi.org/https://doi.org/10.1016/j.socscimed.2018.12.015

Behrens, G., Gredner, T., Stock, C., Leitzmann, M. F., Brenner, H., & Mons, U. (2018). Cancers due to excess weight, low physical activity, and unhealthy diet: estimation of the attributable cancer burden in Germany. *Deutsches Ärzteblatt International*, *115*(35-36), 578.

Binder, M., & Pfisterer, F. (2024). Sequential Pipelines. In B. Bischl, R. Sonabend, L. Kotthoff, & M. Lang (Eds.), *Applied Machine Learning Using mlr3 in R*. CRC Press. https://mlr3book.mlr-org.com/sequential_pipelines.html

Binder, M., Pfisterer, F., Becker, M., & Wright, M. N. (2024). Non-sequential Pipelines and Tuning. In B. Bischl, R. Sonabend, L. Kotthoff, & M. Lang (Eds.), *Applied Machine Learning Using mlr3 in R*. CRC Press. https://mlr3book.mlr-org.com/non-sequential_pipelines_and_tuning.html

Bischl, B., Sonabend, R., Kotthoff, L., & Lang, M. (2024). *Applied machine learning using mlr3 in R*. CRC Press.

Block Jr, R., Golder, M., & Golder, S. N. (2023). Evaluating claims of intersectionality. *The Journal of Politics*, *85*(3), 795-811.

Blom, J., Kilpeläinen, S., Hultcrantz, R., & Törnberg, S. (2014). Five-year experience of organized colorectal cancer screening in a Swedish population – increased compliance with age, female gender, and subsequent screening round. *Journal of Medical Screening*, *21*(3), 144-150. https://doi.org/10.1177/0969141314545555

Blom, J., Löwbeer, C., Elfström, K. M., Sventelius, M., Öhman, D., Saraste, D., & Törnberg, S. (2019). Gender-specific cut-offs in colorectal cancer screening with FIT: Increased compliance and equal positivity rate. *Journal of Medical Screening*, *26*(2), 92-97. https://doi.org/10.1177/0969141318804843

Bolte, G., Jacke, K., Groth, K., Kraus, U., Dandolo, L., Fiedel, L., Debiak, M., Kolossa-Gehring, M., Schneider, A., & Palm, K. (2021). Integrating Sex/Gender into Environmental Health Research: Development of a Conceptual Framework. *International Journal of Environmental Research and Public Health*, *18*(22), 12118. https://www.mdpi.com/1660-4601/18/22/12118

Bowleg, L. (2008). When Black + Lesbian + Woman ≠ Black Lesbian Woman: The Methodological Challenges of Qualitative and Quantitative Intersectionality Research. *Sex Roles*, *59*(5-6), 312-325. https://doi.org/10.1007/s11199-008-9400-z

Bowleg, L. (2012). The problem with the phrase women and minorities: intersectionality-an important theoretical framework for public health. *Am J Public Health*, *102*(7), 1267-1273. https://doi.org/10.2105/AJPH.2012.300750

Bramer, M. (2007). Avoiding overfitting of decision trees. *Principles of data mining*, 119-134.

Braveman, P., & Gottlieb, L. (2014). The Social Determinants of Health: It's Time to Consider the Causes of the Causes. *Public Health Reports®*, *129*(1_suppl2), 19-31. https://doi.org/10.1177/00333549141291s206

Braveman, P., & Gruskin, S. (2003). Defining equity in health. *Journal of Epidemiology & Community Health*, *57*(4), 254-258.

Breiman, L. (1996). Technical Note: Some Properties of Splitting Criteria. *Machine Learning*, *24*(1), 41-47. https://doi.org/10.1023/A:1018094028462

Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199-231, 133. https://doi.org/10.1214/ss/1009213726

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification And Regression Trees. https://doi.org/10.1201/9781315139470

Brown, K. F., Rumgay, H., Dunlop, C., Ryan, M., Quartly, F., Cox, A., Deas, A., Elliss-Brookes, L., Gavin, A., & Hounsome, L. (2018). The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. *British journal of cancer*, *118*(8), 1130-1141.

Brown, N. E., Caballero, G., & Gershon, S. A. (2021). Intersectionality in Political Science. *Political Science*.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233-234. https://doi.org/10.1038/nmeth.4642

Cairney, J., Veldhuizen, S., Vigod, S., Streiner, D. L., Wade, T. J., & Kurdyak, P. (2014). Exploring the social determinants of mental health service use using intersectionality theory and CART analysis. *J Epidemiol Community Health*, *68*(2), 145-150. https://doi.org/10.1136/jech-2013-203120

Cancerfonden. (2023). *Statistik om cancer*. Svensk Insamlings Kontroll. https://www.cancerfonden.se/om-cancer/statistik

Cardoso, R., Hoffmeister, M., & Brenner, H. (2023). Breast cancer screening programmes and self-reported mammography use in <scp>European</scp> countries. *International Journal of Cancer*, *152*(12), 2512-2527. https://doi.org/10.1002/ijc.34494

Carmona-Torres, J. M., Cobo-Cuenca, A. I., Martín-Espinosa, N. M., Piriz-Campos, R. M., Laredo-Aguilera, J. A., & Rodríguez-Borrego, M. A. (2018). Prevalence in the performance of mammographies in Spain: Analysis by Communities 2006-2014 and influencing factors [Article]. *Atencion Primaria*, *50*(4), 228-237. https://doi.org/10.1016/j.aprim.2017.03.007

Caserta, M. S., Lund, D. A., Utz, R. L., & Tabler, J. L. (2016). “One size doesn’t fit all”—partners in hospice care, an individualized approach to bereavement intervention. *Omega-Journal of Death and Dying*, *73*(2), 107-125.

Chen, T., & Guestrin, C. (2016, 2016). *XGBoost: A Scalable Tree Boosting System* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining, San Francisco, California. https://dx.doi.org/10.1145/2939672.2939785

Coetzee, M., Clifford, A. M., Jordaan, J. D., & Louw, Q. A. (2022). Global profile of individuals undergoing total knee replacement through the PROGRESS-PLUS equity lens: Protocol for a systematic review. *South African Journal of Physiotherapy*, *78*(1), 1649. https://doi.org/doi:10.4102/sajp.v78i1.1649

Collective, C. R. (1977). *'A Black Feminist Statement'*. na.

Collins, P. H., & Guo, R. Y. (2021). Reflections on Class and Social Inequality: Sociology and Intersectionality in Dialogue. *Handbook of Classical Sociological Theory*.

Council of the European Union recommendation of 2 December 2003 on cancer screening., 2003 12 (2003).

Regulation (EC) No 1338/2008 of the European Parliament and of the Council of 16 December 2008 on Community statistics on public health and health and safety at work, (2008). http://data.europa.eu/eli/reg/2008/1338/oj

Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics, (2009). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32009R0223

Commission Regulation (EU) 2018/255 of 19 February 2018 implementing Regulation (EC) No 1338/2008 of the European Parliament and of the Council as regards statistics based on the European Health Interview Survey (EHIS), (2018). http://data.europa.eu/eli/reg/2018/255/oj

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics In (pp. 57-80). Routledge. https://doi.org/10.4324/9780429500480-5

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*, 1241.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Crosby, R. (2018). *Predictors of uptake of screening mammography* University of Warwick]. Warwick. https://wrap.warwick.ac.uk/132583/

Cuadraz, G. H., & Uttal, L. (1999). Intersectionality and in-depth interviews: Methodological strategies for analyzing race, class, and gender. *Race, Gender & Class*, 156-186.

Davis, A. Y. (1983). *Women, race & class*. Vintage.

Delgado-Gallegos, J. L., Aviles-Rodriguez, G., Padilla-Rivas, G. R., De Los Angeles Cosio-Leon, M., Franco-Villareal, H., Nieto-Hipolito, J. I., de Dios Sanchez Lopez, J., Zuniga-Violante, E., Islas, J. F., & Romo-Cardenas, G. S. (2023). Application of C5.0 Algorithm for the Assessment of Perceived Stress in Healthcare Professionals Attending COVID-19. *Brain Sci*, *13*(3). https://doi.org/10.3390/brainsci13030513

Dietterricht, T. G. (1995). Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree.

Ding, L., Jidkova, S., Greuter, M. J. W., Van Herck, K., Goossens, M., De Schutter, H., Martens, P., Van Hal, G., & de Bock, G. H. (2021). The Role of Socio-Demographic Factors in the Coverage of Breast Cancer Screening: Insights From a Quantile Regression Analysis. *Front Public Health*, *9*, 648278. https://doi.org/10.3389/fpubh.2021.648278

Durand, M.-A., Carpenter, L., Dolan, H., Bravo, P., Mann, M., Bunn, F., & Elwyn, G. (2014). Do Interventions Designed to Support Shared Decision-Making Reduce Health Inequalities? A Systematic Review and Meta-Analysis. *PLoS One*, *9*(4), e94670. https://doi.org/10.1371/journal.pone.0094670

Eagle, S. R., Brent, D., Covassin, T., Elbin, R. J., Wallace, J., Ortega, J., Pan, R., Anto-Ocrah, M., Okonkwo, D. O., Collins, M. W., & Kontos, A. P. (2022). Exploration of Race and Ethnicity, Sex, Sport-Related Concussion, Depression History, and Suicide Attempts in US Youth. *JAMA Netw Open*, *5*(7), e2219934. https://doi.org/10.1001/jamanetworkopen.2022.19934

ECIR. (2024). Uncovering Inequalities Colorectal Cancer Screening in Europe. In: European Comission.

ECIS. (2020). *European Cancer Information System*. https://ecis.jrc.ec.europa.eu/

Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., & Amirabadizadeh, A. (2018). A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *Journal of research in health sciences*, *18*(2), 412.

Estadística, I. N. d. (2024). *Estadística de Defunciones según la Causa de Muerte* (Nota de prensa, Issue. https://www.ine.es/dyngs/Prensa/es/pEDCM2023.htm

European Commission. (2021). *Communication from the Commission to the European Parliament and the Council. Europe's Beating Cancer Plan*. Brussels Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021DC0044

European Commission. (2022). *Screening ages and frequencies*. https://healthcare-quality.jrc.ec.europa.eu/ecibc/european-breast-cancer-guidelines/screening-ages-and-frequencies

European Commission. (2024). *European Cancer Inequalities Registry*. https://cancer-inequalities.jrc.ec.europa.eu/

Eurostat. (2019). *European Health Interview Survey Germany*.

Eurostat. (2024). *General mortality*. https://doi.org/https://doi.org/10.2908/HLTH_CD_ARO

Evans, C. R. (2019). Reintegrating contexts into quantitative intersectional analyses of health inequalities. *Health & Place*, *60*, 102214. https://doi.org/https://doi.org/10.1016/j.healthplace.2019.102214

Fausto-Sterling, A. (2008). *Sexing the body: Gender politics and the construction of sexuality*. Basic books.

Federal Ministry of Health. (2020). *The German healthcare system*.

Flaxman, A. D., & Vos, T. (2018). Machine learning in population health: Opportunities and threats. *PLoS Med*, *15*(11), e1002702. https://doi.org/10.1371/journal.pmed.1002702

Frederick, A., & Shifrer, D. (2018). Race and Disability: From Analogy to Intersectionality. *Sociology of Race and Ethnicity*, *5*, 200 - 214.

Freitas, C., Tura, L. F., Costa, N., & Duarte, J. (2012). A population-based breast cancer screening programme: conducting a comprehensive survey to explore adherence determinants. *Eur J Cancer Care (Engl)*, *21*(3), 349-359. https://doi.org/10.1111/j.1365-2354.2011.01305.x

Fremon, D. K. (2000). The Jim Crow laws and racism in American history. *Berkeley Heights, NJ*.

Freund, Y. (1996). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139.

Fu, R., Shi, J., Chaiton, M., Leventhal, A. M., Unger, J. B., & Barrington-Trimis, J. L. (2022). A Machine Learning Approach to Identify Predictors of Frequent Vaping and Vulnerable Californian Youth Subgroups. *Nicotine Tob Res*, *24*(7), 1028-1036. https://doi.org/10.1093/ntr/ntab257

Gallant, M. P. (2013). Social networks, social support, and health-related behavior. *The Oxford handbook of health communication, behavior change, and treatment adherence*, *16*, 305-322.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning Variation in Multilevel Models. *Understanding Statistics*, *1*(4), 223-231. https://doi.org/10.1207/S15328031US0104_02

Grace, D. (2014). Intersectionality-informed mixed methods research: A primer. *Health Sociology Review*, *19*(4), 478-490.

Gram, M. A., Therkildsen, C., Clarke, R. B., Andersen, K. K., Mørch, L. S., & Tybjerg, A. J. (2021). The influence of marital status and partner concordance on participation in colorectal cancer screening. *European Journal of Public Health*, *31*(2), 340-346.

Großmann, L. M., Napierala, H., & Herrmann, W. J. (2023). Differences in breast and cervical cancer screening between West and East Germany: a secondary analysis of a german nationwide health survey [Article]. *BMC Public Health*, *23*(1), Article 1931. https://doi.org/10.1186/s12889-023-16849-4

Han, S. H., Kim, K., & Burr, J. A. (2019). Social support and preventive healthcare behaviors among couples in later life. *The Gerontologist*, *59*(6), 1162-1170.

Hanske, J., Meyer, C. P., Sammon, J. D., Choueiri, T. K., Menon, M., Lipsitz, S. R., Noldus, J., Nguyen, P. L., Sun, M., & Trinh, Q.-D. (2016). The influence of marital status on the use of breast, cervical, and colorectal cancer screening. *Preventive Medicine*, *89*, 140-145.

Harris, A. C., & Bartlow, S. (2015). Intersectionality: Race, Gender, Sexuality, and Class.

Heinig, M., Schäfer, W., Langner, I., Zeeb, H., & Haug, U. (2023). German mammography screening program: adherence, characteristics of (non-)participants and utilization of non-screening mammography—a longitudinal analysis [Article]. *BMC Public Health*, *23*(1), Article 1678. https://doi.org/10.1186/s12889-023-16589-5

Henrard, S., Speybroeck, N., & Hermans, C. (2015). Classification and regression tree analysis vs. multivariable linear and logistic regression methods as statistical tools for studying haemophilia. *Haemophilia*, *21*(6), 715-722. https://doi.org/10.1111/hae.12778

Hooks, B. (1952). *Ain't I a woman: Black women and feminism*. South End Press

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674. https://doi.org/10.1198/106186006x133933

INE. (2021). *Encuesta europea de salud en España* Version 1). https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176784&menu=resultados&idp=1254735573175#!tabs-1254736195745

INE. (2023). *Contabilidad Regional de España* https://ine.es/prensa/cre_2022.pdf

Jennifer Allen, Sabine Born, Stefan Damerow, Ronny Kuhnert, Johannes Lemcke, Anja Müller, Tim Weihrauch, & Wetzstein, M. (2021). German Health Update (GEDA 2019/2020-EHIS) –Background and methodology. *Journal of Health Monitoring*, *6*(3). https://doi.org/10.25646/8559

Kass, G. V. (1980). An exploratory techinque for investigating large quantities of categorical data. *Journal of the Royal Statistical Society.*, *29*(2), 119-127. https://www.jstor.org/stable/2986296

Kooperationsgemeinschaft Mammographie. (2012). *Evaluationsbericht 2008-2009. Ergebnisse des Mammographie-Screening-Programms in Deutschland*. https://fachservice.mammo-programm.de/download/evaluationsberichte/Evaluationsbericht_2008-2009.pdf

Kooperationsgemeinschaft Mammographie. (2014). *Evaluationsbericht 2010. Ergebnisse des Mammographie-Screening-Programms in Deutschland*. https://fachservice.mammo-programm.de/download/evaluationsberichte/Evaluationsbericht-2010.pdf

Kooperationsgemeinschaft Mammographie. (2023). *Jahresbericht Evaluation 2021. Deutsches Mammographie-Screening-Programm*. https://fachservice.mammo-programm.de/download/evaluationsberichte/Eval-2021-Webversion.pdf

Krieger, N. (2003). Genders, sexes, and health: what are the connections—and why does it matter? *International Journal of Epidemiology*, *32*(4), 652-657. https://doi.org/10.1093/ije/dyg156

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *5*(28), 1-26. https://doi.org/10.18637/jss.v028.i05

Lagerlund, M., Åkesson, A., & Zackrisson, S. (2021). Population-based mammography screening attendance in Sweden 2017–2018: A cross-sectional register study to assess the impact of sociodemographic factors. *The Breast*, *59*, 16-26. https://doi.org/https://doi.org/10.1016/j.breast.2021.05.011

Lange, C., Finger, J. D., Allen, J., Born, S., Hoebel, J., Kuhnert, R., Müters, S., Thelen, J., Schmich, P., Varga, M., von der Lippe, E., Wetzstein, M., & Ziese, T. (2017). Implementation of the European health interview survey (EHIS) into the German health update (GEDA). *Arch Public Health*, *75*, 40. https://doi.org/10.1186/s13690-017-0208-6

Lemke, D., Berkemeyer, S., Mattauch, V., Heidinger, O., Pebesma, E., & Hense, H.-W. (2015). Small-area spatio-temporal analyses of participation rates in the mammography screening program in the city of Dortmund (NW Germany). *BMC Public Health*, *15*(1). https://doi.org/10.1186/s12889-015-2520-9

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Annals of Behavioral Medicine*, *26*(3), 172-181. https://doi.org/10.1207/s15324796abm2603_02

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, *2*(4), 285-318. https://doi.org/10.1023/a:1022869011914

LMU. (2022). *I2ML - ML Basics - What is Machine Learning?* Course Creator.

Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497*.

Lorenc, T., Petticrew, M., Welch, V., & Tugwell, P. (2013). What types of interventions generate inequalities? Evidence from systematic reviews. *Journal of Epidemiology and Community Health*, *67*(2), 190-193. https://doi.org/10.1136/jech-2012-201257

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, *7*, 154096-154113.

Mahendran, M., Lizotte, D., & Bauer, G. R. (2022). Quantitative methods for descriptive intersectional analysis with binary health outcomes. *SSM Popul Health*, *17*, 101032. https://doi.org/10.1016/j.ssmph.2022.101032

Manjer, Å. R., Emilsson, U. M., & Zackrisson, S. (2015). Non-attendance in mammography screening and women's social network: a cohort study on the influence of family composition, social support, attitudes and cancer in close relations. *World Journal of Surgical Oncology*, *13*, 1-7.

Martín-López, R., Jiménez-García, R., Lopez-de-Andres, A., Hernández-Barrera, V., Jiménez-Trujillo, I., Gil-de-Miguel, A., & Carrasco-Garrido, P. (2013). Inequalities in uptake of breast cancer screening in Spain: Analysis of a cross-sectional national survey [Article]. *Public Health*, *127*(9), 822-827. https://doi.org/10.1016/j.puhe.2013.03.006

Mason, C. H., & Perreault, W. D. (1991). Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research*, *28*(3), 268-280. https://doi.org/10.2307/3172863

Max Kuhn, S. W., Mark Culp, Nathan Coulter, Ross Quinlan. (2023). *C50: C5.0 Decision Trees and Rule-Based Models*. https://rdrr.io/cran/C50/

McCall, L. (2005). The Complexity of Intersectionality. *Signs*, *30*(3), 1771-1800. https://doi.org/10.1086/426800

Mena, E., Bolte, G., & Advance Gender study, g. (2021). Classification tree analysis for an intersectionality-informed identification of population groups with non-daily vegetable intake. *BMC Public Health*, *21*(1), 2007. https://doi.org/10.1186/s12889-021-12043-6

Mena, E., Bolte, G., & AdvanceGender Study, G. (2021). CART-analysis embedded in social theory: A case study comparing quantitative data analysis strategies for intersectionality-based public health monitoring within and beyond the binaries. *SSM Popul Health*, *13*, 100722. https://doi.org/10.1016/j.ssmph.2020.100722

Merlo, J. (2018). Multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) within an intersectional framework. *Soc Sci Med*, *203*, 74-80. https://doi.org/10.1016/j.socscimed.2017.12.026

Merlo, J., Mulinari, S., Wemrell, M., Subramanian, S. V., & Hedblad, B. (2017). The tyranny of the averages and the indiscriminate use of risk factors in public health: The case of coronary heart disease. *SSM - Population Health*, *3*, 684-698. https://doi.org/https://doi.org/10.1016/j.ssmph.2017.08.005

Merlo, J., Öberg, J., Khalaf, K., Perez-Vicente, R., & Leckie, G. (2023). Geographical and sociodemographic differences in statin dispensation after acute myocardial infarction in Sweden: a register-based prospective cohort study applying analysis of individual heterogeneity and discriminatory accuracy (AIHDA) for basic comparisons of healthcare quality. *BMJ Open*, *13*(9), e063117. https://doi.org/10.1136/bmjopen-2022-063117

Miles, A., Cockburn, J., Smith, R. A., & Wardle, J. (2004). A perspective from countries using organized screening programs. *Cancer*, *101*(S5), 1201-1213. https://doi.org/10.1002/cncr.20505

Ministerio de Sanidad. (2021). *Número de profesionales de la medicina que trabajan en el Sistema Nacional de Salud (SNS) en Atención Primaria, Atención Hospitalaria, Servicios de urgencias y emergencias (112/061) y Especialistas en formación según comunidad autónoma*. Retrieved 22.11.2023 from https://www.sanidad.gob.es/estadEstudios/sanidadDatos/tablas/tabla13.htm

Ministerio de Sanidad y Consumo. (1996). *Cribado poblacional de cancer de mama en España*. https://ingesa.sanidad.gob.es/ciudadanos/suSalud/mujer/docs/inform13.pdf

Missinne, S., & Bracke, P. (2015). A cross-national comparative study on the influence of individual life course factors on mammography screening. *Health Policy*, *119*(6), 709-719. https://doi.org/https://doi.org/10.1016/j.healthpol.2015.04.002

Missinne, S., Colman, E., & Bracke, P. (2013). Spousal influence on mammography screening: A life course perspective. *Social Science & Medicine*, *98*, 63-70. https://doi.org/https://doi.org/10.1016/j.socscimed.2013.08.024

Molina-Barceló, A., Moreno Salas, J., Peiró-Pérez, R., Arroyo, G., Ibáñez Cabanell, J., Vanaclocha Espí, M., Binefa, G., García, M., & Salas Trejo, D. (2021). Inequalities in access to cancer screening programmes in Spain and how to reduce them: data from 2013 and 2020. *Rev Esp Salud Publica*, *95*. (Desigualdades de acceso a los programas de cribado del cáncer en España y cómo reducirlas: datos de 2013 y 2020.)

Mottram, R., Knerr, W. L., Gallacher, D., Fraser, H., Al-Khudairy, L., Ayorinde, A., Williamson, S., Nduka, C., Uthman, O. A., Johnson, S., Tsertsvadze, A., Stinton, C., Taylor-Phillips, S., & Clarke, A. (2021). Factors associated with attendance at screening for breast cancer: a systematic review and meta-analysis. *BMJ Open*, *11*(11), e046660. https://doi.org/10.1136/bmjopen-2020-046660

Mousavinezhad, M., Majdzadeh, R., Akbari Sari, A., Delavari, A., & Mohtasham, F. (2016). The effectiveness of FOBT vs. FIT: A meta-analysis on colorectal cancer screening test. *Med J Islam Repub Iran*, *30*, 366.

Mulinari, S., Bredström, A., & Merlo, J. (2015). Questioning the discriminatory accuracy of broad migrant categories in public health: self-rated health in Sweden. *European Journal of Public Health*, *25*(6), 911-917. https://doi.org/10.1093/eurpub/ckv099

National Heart Lung and Blood Institute. (2021). *Quality asessment tool for observational and cohort and cross-sectional studies* https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools

North, M. S., & Fiske, S. T. (2014). Social Categories Create and Reflect Inequality: Psychological and Sociological Insights. In J. T. Cheng, J. L. Tracy, & C. Anderson (Eds.), *The Psychology of Social Status* (pp. 243-265). Springer New York. https://doi.org/10.1007/978-1-4939-0867-7_12

Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of operations research*, *21*(1).

O'Neill, J., Tabish, H., Welch, V., Petticrew, M., Pottie, K., Clarke, M., Evans, T., Pardo Pardo, J., Waters, E., White, H., & Tugwell, P. (2014). Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. *J Clin Epidemiol*, *67*(1), 56-64. https://doi.org/10.1016/j.jclinepi.2013.08.005

O'Campo, P., & Dunn, J. R. (2012). Rethinking social epidemiology : towards a science of change.

OECD. (2023a). *Breast cancer screening (mammography), survey data and programme data*. http://stats.oecd.org/wbos/fileview2.aspx?IDFile=eb5acd7d-2445-401a-b624-62fcdad85091

OECD. (2023b). *Cervical cancer screening, survey data and programme data*. http://stats.oecd.org/wbos/fileview2.aspx?IDFile=c0bc67c1-500a-4167-8eda-b2d7c4631dbc

OECD. (2023c). *Colorectal cancer screening, survey data and programme data*. http://stats.oecd.org/wbos/fileview2.aspx?IDFile=55ec72e2-9b2b-4e12-bd02-490849d0d770

OECD. (2023d). EU Country Cancer Profile: Sweden 2023. In (Vol. EU Country Cancer Profiles). Paris: OECD Publishing.

Ogilvie, D., Fayter, D., Petticrew, M., Sowden, A., Thomas, S., Whitehead, M., & Worthy, G. (2008). The harvest plot: A method for synthesising evidence about the differential effects of interventions. *BMC Medical Research Methodology*, *8*(1), 8. https://doi.org/10.1186/1471-2288-8-8

Oliver, S., Kavanagh, J., Caird, J., Lorenc, T., Oliver, K., Harden, A., Thomas, J., Greaves, A., & Oakley, A. (2008). Health promotion, inequalities and young people's health: a systematic review of research.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, *5*, 1-10.

Pallesen, A. V. J., Herrstedt, J., Westendorp, R. G. J., Mortensen, L. H., & Kristiansen, M. (2021). Differential effects of colorectal cancer screening across sociodemographic groups in Denmark: a register-based study. *Acta Oncologica*, *60*(3), 323-332. https://doi.org/10.1080/0284186X.2020.1869829

Pedros Barnils, N., Eurenius, E., & Gustafsson, P. E. (2020). Self-rated health inequalities in the intersection of gender, social class and regional development in Spain: exploring contributions of material and psychosocial factors. *Int J Equity Health*, *19*(1), 85. https://doi.org/10.1186/s12939-020-01202-7

Pedrós Barnils, N., Härtling, V., Singh, H., Haug, U., & Schüz, B. (2024). A scoping review of sociodemographic inequalities on the uptake of breast cancer screening among targeted women in Germany since the implementation of the Organized Screening Program. osf.io/x79tq

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker [Article]. *American Journal of Epidemiology*, *159*(9), 882-890. https://doi.org/10.1093/aje/kwh101

Persaud, N., Sabir, A., Woods, H., Sayani, A., Agarwal, A., Chowdhury, M., de Leon-Demare, K., Ibezi, S., Jan, S. H., Katz, A., LaFortune, F. D., Lewis, M., McFarlane, T., Oberai, A., Oladele, Y., Onyekwelu, O., Peters, L., Wong, P., & Lofters, A. (2023). Preventive care recommendations to promote health equity. *Cmaj*, *195*(37), E1250-e1273. https://doi.org/10.1503/cmaj.230237

Pons-Rodriguez, A., Martinez-Alonso, M., Perestelo-Perez, L., Garcia, M., Sala, M., Rue, M., en nombre del grupo, I., & El grupo InforMa esta formado, p. (2020). Informed choice in breast cancer screening: the role of education. *Gac Sanit*, *35*(3), 243-249. https://doi.org/10.1016/j.gaceta.2020.01.002 (Eleccion informada en el cribado del cancer de mama: el papel del nivel educativo.)

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers Inc*.

Regionala Cancercentrum Samverkant. (2022). *Tjock- och ändtarmscancer- screening. Nationellt vårdprogram*. https://kunskapsbanken.cancercentrum.se/globalassets/kunskapsbanken/tjock--och-andtarmscancerscreening/nationellt-vardprogram-tjock-och-andtarmscancerscreening.pdf

Rey del Castillo, J. (1998). *Descentralización de los servicios sanitarios. Aspectos generales y análisis del caso espanol* (Vol. Monografias EASP: 23). Copartgraf S.C.A.

Rondet, C., Lapostolle, A., Soler, M., Grillo, F., Parizot, I., & Chauvin, P. (2014). Are Immigrants and Nationals Born to Immigrants at Higher Risk for Delayed or No Lifetime Breast and Cervical Cancer Screening? The Results from a Population-Based Survey in Paris Metropolitan Area in 2010. *PLoS One*, *9*(1), e87046. https://doi.org/10.1371/journal.pone.0087046

Rose, G. (1993). *The Strategy of Preventive Medicine*. Oxford University Press. https://doi.org/10.1093/oso/9780192624864.001.0001

Sahoo, A. K., Pradhan, C., & Das, H. (2020). Performance Evaluation of Different Machine Learning Methods and Deep-Learning Based Convolutional Neural Network for

Health Decision Making. In M. Rout, J. K. Rout, & H. Das (Eds.), *Nature Inspired Computing for Data Science* (pp. 201-212). Springer International Publishing. https://doi.org/10.1007/978-3-030-33820-6_8

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Schapire, R. E. (1990). The Strength of Weak Learnability. *Machine Learning*(5), 197-122.

Schueler, K. M., Chu, P. W., & Smith-Bindman, R. (2008). Factors Associated with Mammography Utilization: A Systematic Quantitative Review of the Literature. *Journal of Women's Health*, *17*(9), 1477-1498. https://doi.org/10.1089/jwh.2007.0603

Seeley, J. R., Stice, E., & Rohde, P. (2009). Screening for depression prevention: identifying adolescent girls at high risk for future depression. *Journal of abnormal psychology*, *118*(1), 161.

Serral, G., Borrell, C., & Puigpinos, I. R. R. (2018). [Socioeconomic inequalities in mammography screening in Spanish women aged 45 to 69]. *Gac Sanit*, *32*(1), 61-67. https://doi.org/10.1016/j.gaceta.2016.12.010 (Desigualdades socioeconomicas en el control mamografico en mujeres espanolas de 45 a 69 anos de edad.)

Soerjomataram, I., Shield, K., Marant-Micallef, C., Vignat, J., Hill, C., Rogel, A., Menvielle, G., Dossus, L., Ormsby, J.-N., & Rehm, J. (2018). Cancers related to lifestyle and environmental factors in France in 2015. *European Journal of Cancer*, *105*, 103-113.

Solar, O., & Irwin, A. (2010). *A conceptual framework for action on the social determinants of health*.

Speed, R. (1994). Regression type techniques and small samples: A guide to good practice. *Journal of Marketing Management*, *10*(1-3), 89-104. https://doi.org/10.1080/0267257X.1994.9964262

Stadler, G., Chesaniuk, M., Haering, S., Roseman, J., Straßburger, V. M., Martina, S., Aisha-Nusrat, A., Maisha, A., Kasia, B., Theda, B., Pichit, B., Marc, D., Sally, D. M., Ruth, D., Ilona, E., Marina, F., Paul, G., Denis, G., Ulrike, G., . . . Mine, W. (2023). Diversified innovations in the health sciences: Proposal for a Diversity Minimal Item Set (DiMIS). *Sustainable Chemistry and Pharmacy*, *33*, 101072. https://doi.org/https://doi.org/10.1016/j.scp.2023.101072

Starker, A. K., Klaus; Kuhner, Ronny. (2017). Early detection of breast cancer: the utilization of mammography in Germany. *Journal of Health Monitoring*, *2*(4). https://doi.org/10.17886/rki-gbe-2017-125

Statistisches Bundesamt. (2024). *Causes of death*. Retrieved 07.04.2024 from https://www.destatis.de/EN/Themes/Society-Environment/Health/Causes-Death/_node.html#sprg267092

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77-89. https://doi.org/https://doi.org/10.1016/S0034-4257(97)00083-7

Strömberg, U., Bonander, C., Westerberg, M., Levin, L., Metcalfe, C., Steele, R., Holmberg, L., Forsberg, A., & Hultcrantz, R. (2022a). Colorectal cancer screening with fecal immunochemical testing or primary colonoscopy: An analysis of health equity based on a randomised trial. *eClinicalMedicine*, *47*, 101398. https://doi.org/10.1016/j.eclinm.2022.101398

Strömberg, U., Bonander, C., Westerberg, M., Levin, L. Å., Metcalfe, C., Steele, R., Holmberg, L., Forsberg, A., & Hultcrantz, R. (2022b). Colorectal cancer screening with fecal immunochemical testing or primary colonoscopy: An analysis of health equity based on a randomised trial [Article]. *eClinicalMedicine*, *47*, Article 101398. https://doi.org/10.1016/j.eclinm.2022.101398

Tabery, J. (2011). Commentary: Hogben vs the Tyranny of Averages. *International Journal of Epidemiology*, *40*(6), 1454-1458. https://doi.org/10.1093/ije/dyr027

The FoRt Student Project Team, H., T. (2009). *CHAID: CHi-squared automated interaction detection. R package version 0.1-2.* https://rdrr.io/rforge/CHAID/

The Public Health Agency of Sweden. (2019-2020). *European Health Interview Survey (EHIS)* https://www.folkhalsomyndigheten.se/folkhalsorapportering-statistik/om-vara-datainsamlingar/european-health-interview-survey-ehis/

Therneau, T. M., Atkinson, B., & Ripley, B. D. (2023). *rpart: Recursive Partitioning and Regression Trees*. Retrieved 10.09.2022 from https://cran.r-project.org/web/packages/rpart/index.html

Thomas, J. (2024). Preprocessing. In B. Bischl, R. Sonabend, L. Kotthoff, & M. Lang (Eds.), *Applied Machine Learning Using mlr3 in R*. CRC Press. https://mlr3book.mlr-org.com/preprocessing.html

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., . . . Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR):

Checklist and Explanation. *Annals of Internal Medicine*, *169*(7), 467-473. https://doi.org/10.7326/m18-0850

Turan, J. M., Elafros, M. A., Logie, C. H., Banik, S., Turan, B., Crockett, K. B., Pescosolido, B., & Murray, S. M. (2019). Challenges and opportunities in examining and addressing intersectional stigma and health. *BMC Medicine*, *17*(1), 7. https://doi.org/10.1186/s12916-018-1246-9

van de Schootbrugge-Vandermeer, H. J., Lansdorp-Vogelaar, I., de Jonge, L., van Vuuren, A. J., Dekker, E., Spaander, M. C. W., Ramakers, C. R. B., Nagtegaal, I. D., van Kemenade, F. J., van Leerdam, M. E., & Toes-Zoutendijk, E. (2023). Socio-demographic and cultural factors related to non-participation in the Dutch colorectal cancer screening programme. *Eur J Cancer*, *190*, 112942. https://doi.org/10.1016/j.ejca.2023.112942

Venkatasubramaniam, A., Wolfson, J., Mitchell, N., Barnes, T., Jaka, M., & French, S. (2017). Decision trees in epidemiological research. *Emerging Themes in Epidemiology*, *14*(1). https://doi.org/10.1186/s12982-017-0064-4

Von Wagner, C., Good, A., Whitaker, K. L., & Wardle, J. (2011). Psychosocial determinants of socioeconomic inequalities in cancer screening participation: a conceptual framework. *Epidemiologic reviews*, *33*(1), 135-147.

Wemrell, M., Karlsson, N., Perez Vicente, R., & Merlo, J. (2021). An intersectional analysis providing more precise information on inequities in self-rated health. *International Journal for Equity in Health*, *20*(1), 54. https://doi.org/10.1186/s12939-020-01368-0

Whitehead, M. (1992). The concepts and principles of equity and health. *International journal of health services : planning, administration, evaluation*, *22 3*, 429-445.

Wild, C. P., Espina, C., Bauld, L., Bonanni, B., Brenner, H., Brown, K., Dillner, J., Forman, D., Kampman, E., Nilbert, M. C., Steindorf, K., Storm, H. H., Vineis, P., Baumann, M., & Schüz, J. (2019). Cancer Prevention Europe. *Molecular Oncology*, *13*, 528 - 534.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. Proceedings of the 18th international conference on evaluation and assessment in software engineering,

Wolfson, M., Gribble, S., & Beall, R. (2017). Exploring Contingent Inequalities: Building the Theoretical Health Inequality Model. In A. Grow & J. Van Bavel (Eds.), *Agent-Based Modelling in Population Studies: Concepts, Methods, and Applications* (pp. 487-513). Springer International Publishing. https://doi.org/10.1007/978-3-319-32283-4_17

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241-259. https://doi.org/https://doi.org/10.1016/S0893-6080(05)80023-1

World Health Organization. (2010). *A conceptual framework for action on the social determinants of health*. World Health Organization.

Yep, G. A. (2002). From Homophobia and Heterosexism to Heteronormativity. *Journal of Lesbian Studies*, *6*, 163 - 176.

Zamorano-Leon, J. J., López-de-Andres, A., Álvarez-González, A., Astasio-Arbiza, P., López-Farré, A. J., de-Miguel-Diez, J., & Jiménez-García, R. (2020). Reduction from 2011 to 2017 in adherence to breast cancer screening and non-improvement in the uptake of cervical cancer screening among women living in Spain [Article]. *Maturitas*, *135*, 27-33. https://doi.org/10.1016/j.maturitas.2020.02.007

Zenko, B., Todorovski, L., & Dzeroski, S. (2001). A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods. Proceedings 2001 IEEE international conference on data mining,

Appendices

A. Individual publications

B. Overview of presentation of the work from the present dissertation at scientific contexts

C. Declaration of originality

A Individual publications

The following four manuscripts and their appendix are the core of this dissertation. Below I provide the bibliographic information of the manuscripts together with the contributions of every author.

Manuscript 1: Pedrós Barnils N, Härtling V, Singh H, Schüz B. Sociodemographic inequalities in breast cancer screening attendance in Germany following the implementation of an Organized Screening Program: Scoping Review. BMC Public Health. 2024;24, 2211. https://doi.org/10.1186/s12889-024-19673-6

NP: Conceptualization, Data curation, Formal Analysis, Visualization, Writing - original draft. VH: Data curation, Formal Analysis. HS: Conceptualisation, Formal Analysis. BS: Conceptualisation, Supervision, Writing - review and editing. NP led the revision of the manuscript based on the reviewers' feedback. She compiled the responses and revised the final manuscript.

Manuscript 2: Pedrós Barnils, N., Schüz, B. (2025). Identifying intersectional groups at risk for missing breast cancer screening: Comparing regression- and decision tree-based approaches, SSM - Population Health, Volume 29, 2025, 101736, ISSN 2352-8273, https://doi.org/10.1016/j.ssmph.2024.10173

NP and BS conceived and designed the study. NP acquired the data, contributed methodologically, carried out and interpreted the final analysis, produced the visualisations, and prepared the primary manuscript. BS supervised, contributed methodologically, and critically revised the manuscript. NP led the revision of the manuscript based on the reviewers' feedback. She compiled the responses and revised the final manuscript.

Manuscript 3: Pedrós Barnils, N., Gustafsson, Per E. (2025). Intersectional inequities in colorectal cancer screening attendance in Sweden: using decision trees for intersectional matrix reduction, Social Science & Medicine, 117583, ISSN 0277-9536, https://doi.org/10.1016/j.socscimed.2024.117583

NP and PEG conceived and designed the study. NP acquired the data, acquired funding, contributed methodologically, carried out and interpreted the final analysis, produced the visualisations, and prepared the primary manuscript. PEG acquired funding, supervised, contributed methodologically, critically revised the manuscript. NP led the revision of the manuscript based on the reviewers' feedback. She compiled the responses and revised the final manuscript.

Manuscript 4: Pedrós Barnils N, Schüz B. Intersectional analysis of inequalities in self-reported breast cancer screening attendance using supervised machine learning and PROGRESS-Plus framework. Frontiers in Public Health. 2024;11. https://doi.org/10.3389/fpubh.2023.1332277

NP: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. BS: Conceptualization, Supervision, Validation, Writing – review & editing. NP led the revision of the manuscript based on the reviewers' feedback. She compiled the responses and revised the final manuscript.

B Presentations of work from the present dissertation in scientific contexts

European Public Health Conference, 2022

Núria Pedrós Barnils and Benjamin Schüz. The "grey" digital divide in older adults during COVID-19 in Germany: Who is most at risk? 2022. European Journal of Public Health, Volume 32, Issue Supplement_3, ckac130.059, https://doi.org/10.1093/eurpub/ckac130.0591

Poster presentation

World Public Health Conference, 2023

Núria Pedrós Barnils and Benjamin Schüz. Inequities in breast cancer screening utilisation in Spain - Using decision trees to identify intersections. 2023. Population Medicine, 5 (Supplement), A1476. https://doi.org/10.18332/popmed/163755

Poster presentation

AI in Health, U Bremen Research Alliance, 2023

Núria Pedrós Barnils and Benjamin Schüz. Using machine learning to identify social intersections in healthcare. 2023.

Oral presentation

Presentation at a colloquium at the Health Psychology Department of the University of Mannheim, 2023

Núria Pedrós Barnils and Benjamin Schüz. Quantitative intersectionality and Machine Learning – opportunities for health inequities research. 2023.

Colloquium

Presentation at the Seminar Series “Theme Equity in Health” from the Medical Faculty at Umeå University (Sweden), 2023

Núria Pedrós Barnils, Antonio Moreno Llamas, Cynthia Anticona, Per Gustafsson. Investigating intricate inequities – different approaches to intersectional inequities in health and healthcare. 2023.

Colloquium

European Society for Health and Medical Sociology 2024

Núria Pedrós Barnils and Benjamin Schüz. Using decision trees to identify intersectional risk groups for never attending breast cancer screening in Germany. 2024.

Oral presentation

European Health Psychology Society Conference, 2024

Núria Pedrós Barnils and Benjamin Schüz. Quantitative methods to understand disadvantage in health research – breast cancer screening attendance in Germany. 2024

Oral presentation

European Public Health Conference, 2024

Núria Pedrós Barnils and Benjamin Schüz. Facilitating subgroup identification: the use of decision trees in breast cancer screening uptake. 2024. European Journal of Public Health, Volume 34, Issue Supplement_3, ckae144.583, https://doi.org/10.1093/eurpub/ckae144.583

Oral presentation

Núria Pedrós Barnils, Victoria Härtling, Himal Singh, Ulrike Haug, and Benjamin Schüz, A scoping review on sociodemographic inequalities in breast cancer screening attendance in Germany. 2024. European Journal of Public Health, Volume 34, Issue Supplement_3, ckae144.918, https://doi.org/10.1093/eurpub/ckae144.918

Poster presentation

C Declaration of originality

I hereby declare, that …

- … it is my original work, and it is conducted without unauthorised assistance.
- … only the referenced sources and tools were used for this dissertation.
- … I made due references to all published or unpublished work either quoted or used as the basis for ideas.

I permit, this dissertation to be checked for plagiarism using appropriate software.

Bremen, 20.11.2024

____

Núria Pedrós Barnils