TU WIEN Informatics

Methods and applications for the secondary use of claims data from the Austrian health insurance system

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Medizinische Informatik

eingereicht von

Florian Endel, BSc
Matrikelnummer 0548659

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.-Prof. Dipl.-Ing. Dr. Georg Duftschmid
Mitwirkung: Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser
Assoc. Prof. Priv.-Doz. Dr.med.univ. Alexander Niessner, MSc.

Wien, 15. Februar 2021

Florian Endel

Georg Duftschmid

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Informatics

Methods and applications for the secondary use of claims data from the Austrian health insurance system

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Medical Informatics

by

Florian Endel, BSc
Registration Number 0548659

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof. Dipl.-Ing. Dr. Georg Duftschmid
Assistance: Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser
Assoc. Prof. Priv.-Doz. Dr.med.univ. Alexander Niessner, MSc.

Vienna, 15th February, 2021

Florian Endel Georg Duftschmid

Erklärung zur Verfassung der Arbeit

Florian Endel, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 15. Februar 2021

Florian Endel

Abstract

Background Previous studies suggest that a significantly increased risk to suffer from myocardial infarction can be observed for parents in comparison with adults without children. However, definitive evidence is lacking and insufficient to adapt clinical practice guidelines on that basis. Furthermore, actual cases seem to be rare and backing data is not available.

Objectives The main objective of this work is to gather evidence on the relative risk of myocardial infarction in parents compared with couples without children in a retrospective, observational cohort study.

Methods Reimbursement data from the health care system are routinely collected by Austrian social insurance institutions for administrative and accounting purposes. *GAP-DRG*, a linked research database holding data from the Austrian health and social insurance system covering several years is utilized to determine the number of potential cases.

Genealogical information is crucial but lacking from the data source. As a result, parentship is not encoded in the available administrative data. Therefore, a method for indirect deduction of this personal information is developed and implemented. Furthermore, information about individual comorbidities and the social-economic status are deduced.

Identified cohorts are documented in detail and various statistical procedures are applied. Such methods include univariate statistics and cross-tabulations, decision trees, and multivariate regression models. Resampling, balancing, and propensity score matching are used to achieve more accurate estimates and robustness.

Results In summary, no additional evidence was found to support the initial claim.

In addition, unprecedented developments such as the disclosure of previously unknown data quality problems in the administrative data source, coinsurance networks and genealogical information, the handling of socioeconomic information, and morbidity scores are reported in detail.

Kurzfassung

Kontext Vorangehende Studien deuten darauf hin, dass für Eltern im Vergleich zu Erwachsenen ohne Kinder ein signifikant erhöhtes Risiko einen Myokardinfarkt zu erleiden zu beobachten ist. Die vorhandenen Hinweise reichen allerdings nicht aus um z.B. klinische Leitfäden anzupassen. Darüber hinaus ist die Anzahl der betroffenen Personen gering und es scheinen kaum Daten zu diesem Thema verfügbar zu sein.

Fragestellung Das primäre Ziel dieser Arbeit ist das Sammeln zusätzlicher Evidenz über das relative Risiko eines Myokardinfarkts bei Eltern im Vergleich zu Paaren ohne Kinder.

Methoden Daten zur Inanspruchnahme von Gesundheitsdienstleistungen werden von österreichischen Sozialversicherungsträgern zur Abrechnung und Verwaltung erhoben und gespeichert.

Genealogische Informationen sind in dieser Datenquelle nicht enthalten. Die Verwandtschaftsverhältnisse wie Eltern- und Partnerschaft werden daher indirekt über die Mitversicherungen abgeleitet. Darüber hinaus werden Informationen über individuelle Komorbiditäten und den sozioökonomischen Status integriert.

Für die klinische Fragestellung werden die identifizierten Kohorten detailliert dokumentiert und unterschiedliche statistische Verfahren eingesetzt. Ausgehend von univariaten Teststatistiken und Kreuztabellen werden Entscheidungsbäume und multivariate Regressionsmodelle angewandt. Resampling, Balancing und Propensity Score Matching sollen dabei genauere Schätzungen und mehr Robustheit ermöglichen.

Ergebnisse Beim Vergleich der Gruppen konnte keine Evidenz für ein signifikant erhöhtes Risiko von Eltern entdeckt werden.

Zusätzlich wird ausführlich über neue Erkenntnisse wie das Aufdecken von Datenqualitätsproblemen, Netzwerke von Mitversicherungen und genealogischen Informationen, der Umgang mit sozioökonomischen Informationen und abgeleiteten Morbiditätsbewertungen berichtet.

Contents

CHAPTER 1

Introduction

1.1 Motivation and problem statement

Preceding findings suggest that parenthood is a risk-modifying factor for mortality in midlife [Einiö et al., 2015]. The relative cardiovascular risk profile of parents is also suspected to be worse compared with a similar cohort of couples without children. Experts suggest that this effect may be influenced by parents neglecting a health-promoting lifestyle due to lack of time and a shifted life focus. However, clear evidence is lacking and insufficient to adjust clinical practice guidelines or establish targeted clinical trials. In addition, actual cases appear to be rare and firm data are not available.

Therefore, this retrospective, observational cohort study analyzes parenthood as a risk-modifying factor for myocardial infarction in young adults. Additional evidence is collected and the applicability of secondary-use reimbursement data to this problem is examined. Routinely collected linked administrative claims data from the Austrian health care system represent the primary data source.

It is extremely important to treat and interpret the results obtained with great care. Due to the nature of the data and the methods used, it is not possible to determine causation. The results can only support or weaken initial assumptions about influencing factors and are intended to lay the foundation for new hypotheses and possible further analysis.

1.2 Outline

Beginning with an introduction and information on the objectives of the project, the study protocol is presented in full, including a translation, interpretation, and discussion of its application to the available data and the structure of the Austrian health care system.

Due to the nature of secondary-use administrative data, their complex structure, heterogeneous origins, and large amount of information, exploratory data analysis and detailed documentation of compared cohorts will be conducted. Data quality issues and potentially inadequate or confounding structures will be thoroughly investigated, and unexpected content will be documented.

Since the affiliation of individuals (i.e., adult couples, spouses, and parenthood) is not directly coded in the research database, networks of co-insurance are extracted, visualized, and analyzed. This development is an essential part and achievement of this project, as it has not been done before. Various aspects of the Austrian social insurance system and reimbursement conventions need to be considered in detail to derive the relationships of individuals. The extracted networks allow to tighten the definition of the cohorts and provide necessary information on the cohort assignment (parents or spouses without children) of the observed population.

Following the cohort definition, additional information on socioeconomic status and comorbidities is collected. Since socioeconomic status is not known for all selected individuals, it is imputed based on their individual relationship network. Comorbidities are extracted from hospital episodes and predicted diagnoses based on the $ATC \rightarrow ICD$ project [Filzmoser et al., 2009]. Several common morbidity scores are applied and compared to summarize the collected diagnoses. This completes the data collection, cohort selection, and variable preparation.

Great care is taken in data analysis and statistical modeling. Cohorts are compared using a variety of commonly used methods, while shortcomings in the available information are addressed. Matching at the individual level to reduce bias is also applied to provide further insight. Methods, results, and interpretations are presented in a nuanced manner, and advantages and disadvantages of the chosen approaches are discussed.

REporting of studies Conducted using Observational Routinely collected health Data (RECORD) [Benchimol et al., 2015, Nicholls et al., 2015], a specialized reporting guideline based on *Strengthening the Reporting of Observational Studies in Epidemiology* (STROBE) [von Elm et al., 2007, Vandenbroucke et al., 2014] is followed to comply with current standards.

1.3 Study protocol

A study protocol is provided by Alexander Niessner, an advisor of this thesis. It mainly outlines the compared cohorts and provides a minimum set of statistical analysis to be applied.

Two cohorts are defined in the study protocol: adult couples without children (i.e., *spouses*) and adult couples with children (i.e., *parents*). Based on relationships derived from co-insurance described in section 3.1 on page 33, the cohorts are meant to be classified and extracted by rules defined ex ante. The original instructions are cited, translated, interpreted, and comment in the following list.

Population Patients from the GAP-DRG database who were insured in Austria and received medical service in 2006 and 2007. The study focuses on couples who co-insure each other and, if they are parents, their children as well. Only persons between the ages of 30 and 60 are included.

German original [1] PatientInnen der Datenbank GAP-DRG (= PatientInnen, die in AUT zwischen 2006 und 2007 krankenversichert waren und eine medizinische Leistung in Anspruch genommen haben). Der Fokus liegt auf Paaren, die sich und gegebenenfalls auch ein/mehrere Kinder versichern, der Altersbereich liegt zwischen 30 und 60 Jahren

Interpretation The definition of the study population is a blend of the entire population available in GAP-DRG and the standardized research population. This specific subset includes only patients who have claimed at least one service and do not have data quality issues such as missing or discrepant personal information.

Application While the entire population of GAP-DRG is utilized to gather information on co-insurance, the final study cohorts will be limited to the standardized research population. Couples where the year of birth of at least one partner is unknown and consequently the difference in age cannot be determined also have to be omitted. Additionally, persons deceased before 2006 or missing information on gender are removed. Univariate and multivariate details on couples and individuals affected are provided in chapter 3.1 on page 33.

Intervention *Parents* are co-insured with at least one child, defined by age and difference in age. Co-insured children are younger than 28 years (i.e. ≤ 27) and at least 18 years younger than their parent (i.e. "age of the insured person" - "age of the co-insured person" ≥ 18). Only parents with a known spouse due to co-insurance of the adult partners or a relationship of both with at least one common child are included.

German original Eltern, i.e., PatientInnen mit einem oder mehreren mitversicherten PatientInnen, von denen aufgrund ihres Alters und dem Altersunterschied angenommen werden kann, dass es sich um Kinder handelt. Annahme z.B. "Alter(Versicherter) – Alter(Mitversicherter) >= 18 AND Alter(Mitversicherter) <= 27". Es werden nur Eltern betrachtet, bei denen auch der Partner durch eine Mitversicherung erkannt wird (und gegebenenfalls auch das/die Kinder mitversichert).

Interpretation Children and parents are clearly defined in the study protocol. As no upper size of a family is defined, cases where multiple adults are co-insured with the same child or with each other have also to be included.

[1] The German text is an exact citation from the original study protocol.

Application The provided definition is applied strictly and additional information on the size (e.g., number of parents per child, number of spouses per adult) and combination of genders are utilized as quality indicators. As there are children with parents younger than 30, a parent does not automatically have to be included in this cohort even though all other parameters are fitting.

Control *childless adults*: Because children are often co-insured with only one parent and as a result the second parent might be misclassified as a childless person, couples without children are identified as the control group. Childless couples are persons in a relationship (identified by co-insurance) with a difference in age of 17 years or less, where both (all) partners are not in relationship with children.

German original Kinderlose Erwachsene => Da vielfach Kinder nur bei einem Elternteil als Mitversicherte aufscheinen und der 2. Elternteil dann irrtümlich als kinderlos interpretiert würde, Identifikation von kinderlosen Paaren als Kontrollgruppe. Diese werden definiert durch eine andere mitversicherte Person, von der aufgrund ihres Alters angenommen werden kann, dass es sich um einen Partner handelt (der Altersunterschied der Partner beträgt maximal 17 Jahre: "Alter(Versicherter) – 17 Jahre <= Alter(Partner) <= Alter(Versicherter) + 17 Jahre"). Wenn beim Versicherten dann kein mitversichertes Kind aufscheint, würde man das Paar als kinderlos interpretieren.

Interpretation Spouses with and without children are defined in the same way. Their difference in age is less than 18 years (i.e., $\leq$ 17 years). In both cohorts, only adults aged between 30 and 60 will be included. The association with children can, on the one hand, be utilized to define a couple in case a child is co-insured with both partners, and on the other hand, be *inherited* from one adult to another. As a result, there are more possibilities to identify parents in comparison with childless couples.

Application First, all adult couples are identified due to co-insurance or common children. Second, each couple is labeled as *parent* or *childless*. Special attention is paid to larger family networks where several adults are co-insured with each other.

Some clarification or minor extensions, respectively, must be included to fill in undefined cases.

First, it is possible that more complex relationships exist in addition to the classic family structure of two (heterosexual) parents with children. While such networks may be plausible on a smaller scale where only a few adults and children are involved, there are larger (>100 or even >1,000 participants) groups of individuals who are connected. These constellations cannot be interpreted as a joint family consisting of parents with children and could be a result of flawed data.

Second, since the members of both cohorts are between 30 and 60 years old and the definition of couples and children is based on age difference, there are mixed cases where one partner is a member of the cohort but the second is not. For example, an adult 25-year-old man is co-insured with a 2-year-old child (age difference ≥ 18) and with a 37-year-old woman (age difference ≤ 17). To correctly identify this family, the man must be included for cohort extraction, but must be removed from the final cohort because of the minimum age of 30. Of course, several other mixed cases are possible and likely.

Furthermore, the outcome, i.e., myocardial infarctions, is defined. Additional covariates as classes for stratification as well as the matching of the defined cohorts is described briefly:

Outcome The rate of myocardial infarction of the group *intervention* (i.e., parents) is compared to the rate in cohort *control* (i.e., adult spouses without children). Myocardial infarction is identified using diagnoses from the inpatient sector[2] covering ICD-10[3] codes *I21* "Acute myocardial infarction" and *I22* "Subsequent myocardial infarction", including all sub-codes of the ICD-10 hierarchy.

German original Myokardinfarktrate bei Interventionsgruppe (Eltern) im Vergleich zur Control-Gruppe (kinderlose Erwachsene). Hierbei wird nach MBDS-Diagnosen mit den ICD10-Codes "I21.- Akuter Myokardinfarkt" sowie "I22.- Rezidivierender Myokardinfarkt" der PatientInnen gesucht.

Interpretation The main objective is a comparison of the rates of myocardial infarctions between cohort *intervention* (parents) and the *control* group (adult couples without children). Myocardial infarctions are deduced from inpatient data, i.e., the linked data source *MBDS* of the GAP-DRG database. Diagnoses from ICD-10 chapter I21[4] and I22[5] define an event of interest. There is no distinction whether main or additional diagnoses from the Austrian DRG system are expected.

Application Record linkage between inpatient and outpatient data [Endel et al., 2012, Endel et al., 2011] is already integrated into the *GAP-DRG* database. Therefore, gathering the requested diagnoses for a defined cohort is a straightforward procedure. Detailed analysis concerning the distribution of main and additional diagnoses is required. Furthermore, comparing all relevant diagnoses in the database in comparison to the events occurring in the selected population is expected to give an impression of the generalizable, coverage and possibly even bias of this study.

[2] Concerning GAP-DRG, data about hospital spells originate from the *MBDS* dataset, which is often used as a synonym for the inpatient sector in general. *MBDS* is an abbreviation for *Minimal Basic DataSet*, which is a defined data structure for Austrian hospitals to report inpatient spells.

[3] ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems

[4] ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction

[5] Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction

Stratification involves the segmentation into the defined cohorts and, in addition, grouping by age (e.g., 5 year groups) and sex. Optionally, the number of children and arbitrarily grouped age of the youngest child can be integrated.

German original Hypothese (1): PatientInnen werden unterteilt in Eltern und kinderlose Erwachsene, nach dem Geschlecht und ihrer Altersgruppen (z.B. 5 Jahresintervalle). Weitere Fragestellung (2): die Eltern werden zusätzlich nach der Anzahl ihrer Kinder und dem Alter des jüngsten Kindes (0-5, 6-10, 11-15, 16-18, 19-27 Jahre entsprechend des Ausbildungssystems) unterteilt.

Interpretation This section of the study protocol defines the main covariates of this study, i.e., age and sex. Additional information concerning the number of children and age of the youngest child can be integrated optionally. These variables are expected to be utilized as dimensions of a cross table or exogenous variables in a statistical model.

Application Age and gender is suspected to be available for most persons. It is also possible to limit the selected cohorts to solely contain persons where this personal information is available. It turns out that the number of children and, as a result, age of the youngest child is not directly useful because there are networks of relationships where a rather high number of children are present. Although only a small fraction of all individuals in cohort *intervention* are affected, no clear boundary for the maximum number of children is defined. Furthermore, preliminary analysis showed that the overall results do not change significantly when covariates describing the number and age of children are included for parents. Therefore, this issue has been discussed with the supervisor and removed from the analysis.

Matching of the cohorts according the individual socioeconomic status has to be applied.

German original Gruppen-Matching (= Zusammensetzung aller Kontrollen entsprechend der Zusammensetzung der Fälle) nach dem soziökonomischen Status.

Interpretation The cohorts are expected to be matched by the individual socioeconomic status. No details concerning the matching method, distance function, stratification of the matched cohorts and replacement, i.e., whether individuals are expected to be replicated, are provided.

Application Propensity score matching is selected by the author due to its flexibility and wide application in similar scenarios. In addition to matching by the requested variable, the cohorts are stratified by sex and age and are also matched by a grouped (co-) morbidity score for comparison. Both matching with and without replacement are applied.

Additionally, suggestions for cross-tabulation and the presentation of results are provided, but not replicated here.

While the provided study protocol defines constrictions specifying the age difference of spouses and between parents and their children, several other technical as well as system related parameters are not defined and will therefore be decided based on the exploratory data analysis. Furthermore, arrangements to enhance the data quality and minimize distorting effects and misclassification are implemented.

1.4 Expected limitations

Gathering new evidence for or against a presumed causal correlation of a higher risk of suffering an early myocardial infarction for parents compared with couples without children is the main objective of this study.

Genealogical information is the main distinguishing feature of these cohorts. Since this information is not directly available in the research database, family relationships must be inferred through co-insurance of relatives. This limitation has a strong impact on the study design, the definition of the cohorts, and the quality and interpretation of the results. Therefore, the quality of the supplemented genealogical information is examined extensively, as described in chapter 3.1 on page 33 for the exploration of the underlying coinsurance networks and in chapter 3.4 on page 110 for the results of the data quality analysis. This rather restrictive approach of distinguishing between parents and spouses without children is expected to have a high specificity[6] but a weaker sensitivity[7].

The detection of a myocardial infarction is expected to be reliable, since such severe and outstanding events are usually treated, diagnosed and thus reliably documented in a hospital. The coded diagnoses are therefore directly available. 2.2% of all Austrians reported having suffered a myocardial infarction in their entire lifetime according to ATHIS 2006/07 [Klimont et al., 2007], and only a small proportion were younger than 60 years at the time of the event. It is expected that there would be few cases of identifiable parents aged between 30 and 60 years suffering a myocardial infarction during the observed period. Therefore, the number of outcome events and, in particular, parents with a documented event is likely to be very small.

1.5 Objectives of this thesis

In summary, the main objectives of this thesis are:

- to gather information on the effect of parenthood on the risk of myocardial infarction
- to explore and discuss the applicability of using secondary data for the research question

[6] correct identification of patients without the outcome, in this case, without children: $\frac{true\,negative}{negative} = \frac{identified\,controls}{all\,spouses\,without\,children}$

[7] correct identification of patients with the outcome, in this case, parents: $\frac{true\,positive}{positive} = \frac{identified\,parents}{all\,parents}$

- to derive genealogical information from co-insurance networks

CHAPTER 2

Methods

Routinely collected, linked administrative claims data from the Austrian health care system are used secondarily as the data source of the analysis. Therefore, it is defined as a retrospective observational study. An ecological study design [Porta, 2014, p.89] is chosen to provide an overview when preparing the data set and data quality assessment. Subsequently, a cohort study design [Porta, 2014, p. 50] is applied, where the status *parent* is defined as *exposure* or *intervention*, respectively. Due to the observational nature of the data source and the study design itself, only evidence of correlated occurrence of events and individual characteristics can be reliably identified, but not causal effects.

There are two major sections in this study that consist of several topics:

1. data preparation
   a) data extraction from the data sources, exploratory and descriptive analysis
   b) co-insurance networks and calculation of genealogical information
   c) cohort extraction, and additional variables: socioeconomic status, comorbidity, and myocardial infarction
2. statistical analysis
   a) logistic regression
   b) decision trees
   c) gradient boosting machines and handling class imbalance
   d) propensity score matching

2.0.1 Tools used

Originally, it was intended to transform the research database *GAP-DRG*, consisting of administrative reimbursement data of the Austrian social insurance, into a Clinical Research Data Warehouse (CRDW). Subsequently, data exploration and data extraction ought to be carried out with this novel tool.

The CRDW should be based on the open source software *i2b2*, an acronym for "Informatics for Integrating Biology & the Bedside" [Murphy et al., 2006, Murphy et al., 2010].

Current technologies such as software containers and common ETL procedures should be used. The data analysis for the cohort study should then be carried out within this new platform.

Most parts of the implementation of *i2b2* and the transformation of *GAP-DRG* were successfully accomplished. The use of software containers like *docker* enabled the deployment of *i2b2* and its notoriously complex installation procedure on the highly secured servers of the *GAP-DRG* database. The construction of metadata repositories, the so-called *ontology* for the Austrian reimbursement system and German terms, and the transformation of large areas of the research database into i2b2's star(-like) schema were completed. Performance optimizations of the underlying database software PostgreSQL, e.g. the implementation of a columnar storage engine *cstore fdw*, enabled meaningful complex queries with the integrated interface. The results of this work were published separately and awarded [Endel and Duftschmid, 2016].

Nevertheless, the intended data extraction and analysis of the clinical research question could not be performed with *i2b2*. The integrated interface did not provide the required functionality mainly due to the inherent complexity of the cohort definition. Missing information not directly available in this administrative data collection, such as the required genealogical information, could not be generated and explored within the CRDW. Even the implementation of extensions to i2b2 such as the Integrated Data Repository Toolkit (IDRT) [Bauer et al., 2015] and an *R engine cell* [Segagni et al., 2011, Weinlich, B. et al., 2014] could not sufficiently alleviate the limitation. Similar experiences about advantages and disadvantages regarding *i2b2* are reported by other research groups, e.g., [Deshmukh et al., 2009, Ganslandt et al., 2011, Johnson et al., 2014].

As a result, the entire topic pertaining to i2b2 was removed from this thesis. Instead of i2b2, prevailing tools like SQL-queries and the statistical computing environment R were applied to answer the clinical research question.

2.1 Data sources

The primary data source used is GAP-DRG[1], a research database consisting of pseudonymized claims data from Austrian social insurance institutions. This collection of routinely col-

[1] GAP-DRG is an abbreviation of one of the first larger research projects based on this data collection called "General Approach for Patient-oriented Ambulant DRGs"

lected administrative data is linked for almost the entire Austrian population[2] for the years 2006 and 2007. Furthermore, data from the second largest regional insurance institution *NÖGKK* (Niederösterreichische Gebietskrankenkasse) are integrated for the years 2008 to 2011.[3]

The main source of the GAP-DRG research database is the *FoKo* data warehouse, in which reimbursement information from all 19 Austrian social insurance institutions is collected but not linked. This main source provides information on the outpatient sector as well as filled prescriptions and sick leave, accompanied by master data on all insured persons, health care providers, and their reimbursement and coding systems. Limited information on inpatient hospital episodes, but no data on the large sector of outpatient care provided by hospitals, is integrated.

This information has been linked to the *MBDS* dataset on all hospital episodes for the same period [Endel et al., 2011, Endel et al., 2012]. Although MBDS contains all inpatient episodes for all Austrian hospitals, only discharges from hospitals funded by the Austrian *LKF*-system[4] are included.

Additionally, fundamental personal information (i.e., from the data collection called *zentrale Partnerverwaltung, ZPV)*, derived socioeconomic status[5], and several other sources of information as well as various metadata (e.g., spatial data and diagnostic schemes) are integrated. Due to the lack of coded diagnoses for the majority of the population by, e.g., the outpatient sector, they are substituted with predictions based on filled prescriptions on an individual level derived by the project *ATC → ICD* [Filzmoser et al., 2009].

Data quality assessments [Endel, 2014] are conducted during the transformation, linkage, and loading phase of the data collection. As a result, defined standard populations (i.e., the research population *Forschungspopulation*) are defined based on insights from these explorations and experts' opinions.

In summary, GAP-DRG is a rather outdated, but well understood, thoroughly documented, and easily prepared collection of claims data. The large amount of cleaned information, a well-defined data model, integrated extensions such as ZPV and *ATC → ICD*, and straightforward availability are key advantages. As a result, GAP-DRG appears to be a valid and appropriate source for this study.

[2] About 3% of Austria's population is insured by other institution based on various legislation (e.g., municipalities, religious orders, unemployment service) and are not included in GAP-DRG.

[3] In detail, reimbursement information from the largest regional insurance institution *WGKK* (Wiener Gebietskrankenkasse) is also included for persons insured by the NÖGKK but claiming services in Austria's capital Vienna.

[4] "LKF" is an abbreviation of the DRG system instituted in Austria. In addition to publicly funded hospitals reimbursed by the LKF system, there are also other specialized or private hospitals which are not included in this study.

[5] `https://www.sozialversicherung.at/cdscontent/?contentid=10007.844151` (last visited: 2020-07-16)

2.2 Data preparation and data wrangling

In general, data preparation, also known as *data wrangling*, is the most complex and time-consuming part of most data-driven (research) projects [Endel and Piringer, 2015]. The process of data wrangling, including data extraction, linkage, transformation, exploration, and quality assessment, is therefore an important part of this study and must be performed with great care. In [Haug et al., 2011], the authors address the cost of poor data quality, its negative impact on reliable results, and the effort required to correct data quality problems. Visualization and visual analytics are one of the most important tools to explore and evaluate data and its quality, according to [Kandel et al., 2011]. Therefore, data wrangling, exploration, and quality assessment are essential components of this study and are discussed in detail in the Results section 3 on page 33.

All available data should be used. The selection of the final cohort should only be done according to the rules of the study protocol to omit selection and allocation bias [Sedgwick, 2013]. In case there are good reasons, based on the documented and discussed properties of the analyzed data, additional selection criteria might be applied during the analysis, as pointed out in [Wilkinson et al., 2016]. Concerning data originating from GAP-DRG, especially the pre-defined standard populations (i.e., the research population *Forschungspopulation*) and appearance, distribution and influence of missing information will be focused. As a result, the final cohorts will include persons with missing personal properties like age and gender until data quality assessment provides a substantial rationale to exclude such cases.

Data preparation can be split in three major steps:

First, database queries are developed to extract relevant information and metadata from the data source. The resulting data is thoroughly examined and documented, with a focus on co-insurance networks and data quality.

Second, genealogical information is extracted using co-insurance networks. This novel contribution relies heavily on the specifics of the Austrian insurance system and the available administrative data. Based on this data source, such networks have not been built and analyzed before. Therefore, every single step, every variable, and every unexpected or unknown structure has to be investigated and documented in detail.

Third, based on the previously added genealogical information, cohorts are extracted according to the study protocol. Next, they are enriched with additional personal characteristics such as individual socioeconomic status, comorbidity scores, and outcome criteria, which are discussed in detail. Finally, an assessment of the data quality of the selected variables is performed to validate the extracted information and check its consistency with publicly available official statistics.

As a result, a well-documented and understandable data set with novel variables is prepared and ready for further analysis.

2.2.1 Exploratory data analysis and data visualization

Exploratory data analysis (EDA), as prominently promoted by Tukey, e.g., [Tukey, 1977], is the process of analyzing (new) data to gain new insights and formulate hypotheses. Common graphical approaches and cross-tabulations are created for each variable, and selected bivariate and multivariate combinations of these are analyzed. The resulting tables and graphs will be carefully selected and discussed in detail. This most important procedure is repeated for the original data and, based on these results, for the final cohort.

Many decisions about cohort selection details, filtering criteria, and presented characteristics of the available data are based on exploratory data analysis. A selection of key steps and results is presented, documenting the final selection of variables, methods, and interpretation. This data exploration and discussion of identified features, potential problems, and new findings combined with thorough interpretation are the major outcomes of this study.

Visualization is a major tool in EDA. Utilizing R's potent plotting library like ggplot2 [Wickham, 2009a, Wickham, 2011] based on the grammar of graphics [Wickham, 2010] and *tableplots*, introduced in [Tennekes et al., 2013], newly discovered structures, data quality issues, and conclusions are presented and discussed in the results section.

Plotting multivariate or higher dimensional data can be implemented using univariate profiles, parallel linked plots of marginal distributions as provided by *tableplot*, and using the faceting functionality of ggplot2, which can be applied to almost any graphical object within the same framework. While the univariate profile is important for a first impression, multidimensional concurrency can only be explored using more sophisticated methods.

To explore the underlying data of coinsurance networks, the number of relationships for each combination of ages is relevant, resulting in an *age-age matrix* for relationships. There are several ways to explore this matrix, partitioned by personal covariates such as gender and state of insurance. As an example, a two-dimensional density is calculated according to [Venables and Ripley, 2002a] and presented as a 3-dimensional shape and contour plot.

Indeed, it is quite difficult to extract essential information from a rendered 3D surface. Moreover, the methods used are not too flexible and already required manual interventions. As an alternative, the age-age matrix of relationships can be displayed directly as a heatmap or further aggregated for better readability.

Hexagonal binning is chosen for summarizing the data according to [Carr et al., 1987]. For the tessellation of a plane, hexagons provide the maximum number of sides and resemble a circle rather than a square. Varying the number of bins gives a finer or coarser representation. A single square per value (i.e., 2d binning) would also be appropriate for the data at hand. The hexagonal structure and size of single hexagons were chosen to slightly blur the effects introduced by the data source and to focus on the larger structures. Although hexagonal bins are often colored, only grayscale is used to highlight structures on a continuous scale (i.e., number of relationships per bin or derivatives).

Several common visualization techniques such as network plots, integration of zoomed details, box- and violinplots, and bar charts are used in combination with careful application of color coding, transformed axes, and smoothing. Most importantly, all plots presented are interpreted and discussed in detail to provide a comprehensible rationale for the data, results, and final conclusions.

2.2.2 Data quality assessment

Data quality assessment (DQA) is an important topic in health care and medical research [Stausberg et al., 2015].

DQA is strongly related to EDA. The main differences are the goal of the process and the classification of methods. In this study, EDA is used to gather information about data, become familiar with unexpected structures, and document them in detail. In contrast, DQA is applied after much of the data wrangling process to gain further insight into the final data set(s). It is also used to make an informed decision about additional selections, filters, and exclusion of individuals based on observed and documented evidence. Specifically, individuals with missing information on age or sex are not excluded by default to rule out selection bias, as the information may not be missing at random, and to get most records from the source database into the final study cohorts. Given the thoroughly discussed characteristics and data quality issues that are most common in the cohort with missing personal information, the affected individuals may eventually be removed from the statistical analysis rather than imputing missing values, for example.

Another aspect of data quality assessment in this study is the comparison with official statistics regarding family size, number of children and partners. Since the genealogical information is retrieved and not directly measured, and since the data source does not fully match the Austrian population documented by the national statistical office, it cannot be expected that, for example, the number of children per family matches perfectly, but a comparison of the distributions and especially of the extreme values are important quality indicators.

In addition, checking the variables for plausibility helps to exclude gross errors that could distort the result. For example, some algorithms for calculating multimorbidity scores have errors that lead to nonsensical (negative) values that should not be possible. This behavior is not intentional but has been documented according to publications. A similar problem regarding socioeconomic status data and previously unknown data errors regarding year of birth were discovered and documented in the results section.

In summary, DQA is known to be an important part of the data wrangling process. Selected results are presented and discussed in chapter 3.4.

2.2.3 Co-insurance network and genealogical information

Co-insured children and spouses are distinguished based on their age relative to the age of the insured. To avoid misclassification with the childless control group (a child can be co-insured with only one of the two parents, making the other seemingly childless), we focus only on co-insured couples. Using a cohort study design, couples in which one spouse co-insures the other spouse and one or more children will be the exposed group. Couples in which one spouse co-insures the other spouse but no co-insured child is documented will form the control group.

First, all potential relationships between two individuals based on co-insurance are collected from multiple sources. Co-insurance information can be derived directly from the insurance carrier's master data or indirectly from outpatient contacts[6], inpatient contacts, and filled prescriptions. Implicit co-insurances are derived from claims data when the person ID of the patient claiming a service is not the same as the ID of the person whose insurance covers the cost. In addition, for each partner of a (potential) couple, the number of available co-insurance references and other personal data are collected.

Co-insured patients are also referred to as *dependent* in distinction to compulsorily *insured* persons. A connection between two persons is referred to as an indication or *hint* of an actual relationship. Each pair consists of two individuals, an insured and a dependent person. Naturally, a single individual may occupy both roles and be in a relationship (in terms of co-insurance) with several other individuals, resulting in a complex network of individuals.

Therefore, each person can be understood as a node in a directed network of co-insurances. Each network in the co-insurance dataset has at least two nodes, a compulsorily insured person and a dependent person. Multiple relationships per person (e.g., multiple persons dependent on the same compulsorily insured person or a child co-insured by both parents) are possible and common. Directional edges are defined that describe the direction of dependence and weights of various measures such as age difference or amount of evidence. In addition, the analysis of samples of these networks gives an idea of possible quality issues, typical structures, unexpected events, and interpretation of results. Although the final cohort extraction requires a maximum *distance*[7] of two, the analysis of much more complex networks confirms reliable data and allows (random) sampling controls.

An algorithm for recursive exploration of these networks is developed. The recursive method is capable of identifying (directed) *cycles*[8] (e.g., from two individuals that also depend on each other in both directions to cycles of arbitrary distance), constraining the

[6] Instead of single services claimed, only information related to health insurance vouchers is utilized. These vouchers, called *Krankenschein* in German, are stored in the table *leistungsdaten_vp* in GAP-DRG or *Satzart 10* in FoKo respectively.

[7] *Distance* as a measure in graph theory describing the minimum number of edges (i.e., co-insurances) or the shortest path between two nodes (i.e., persons).

[8] *Cycle* as a property defined in graph theory is a path starting and ending in the same node. In directed graphs, the direction of the vertices has to be followed.

maximum explored distance, and running in parallel on multiple CPU cores. Its output is a flat data structure prepared for further analysis and visualization.

Finally, the required genealogical information is extracted from this dataset using a customized algorithm. The entire procedure is performed twice, once for networks derived from (almost) the entire co-insurance dataset, and once for a cleaned version where individuals with unknown information are excluded. While the list of potential study participants is strictly limited, their relationship networks are less restricted.

Starting with the entire coinsurance network, the data is narrowed down to individuals with a maximum of 20 relationships[9]. This full cohort is the basis for the following steps. Next, a subset of networks is defined that is restricted to individuals with known gender and year of birth, referred to as the no-NA cohort. This reduces the number of potentially included individuals by 61.598 individuals from 2.003.707 persons in the full cohort to 1.942.109 in the adjusted subset.

For each potentially included person, the individual relationship network is extracted up to a distance of 2. Next, all associated nodes in these networks are classified as partner, child, or neither, according to the study protocol. Finally, the resulting data for each individual are summarized such that each observation (i.e., row in the dataset) corresponds to a single individual.

If a person is in a relationship with someone who can be classified as a partner, they are referred to as a *control*. If a person not only has a partner but is also associated with a child, the term intervention is applied. Last, any case in which no partner can be found, regardless of parenthood, is excluded.

Official reports from Statistics Austria[10] and the *Austrian Institute for Family Studies* (i.e., reports "Familie in Zahlen", promoted by the federal ministry for family and youth)[11] commonly aggregate families with 3 or more children into a single group. These reports show that there are few occurrences of families with a higher number of children. Interpreting the statistics on divorces and second marriages from the same sources, similar interpretations can be derived for the joint number of spouses per person. However, a clear limit on the maximum number of children and partners cannot be derived from these reports.

Furthermore, figures 3.69 on page 117, 3.68, 3.71 and 3.70 on page 120 show that the number of extreme outliers concerning the number of children and partners per person decreases in the second cohort where the co-insurance networks are reduced to persons without missing information on age and gender. Although the total number of children

[9] Altogether, 234 persons and 19.876 relationships are removed by this filter.

[10] Reports based on the microcensus 2015, e.g., "Familien nach Familientyp, Zahl der Kinder und Bundesländern - Jahresdurchschnitt 2015", http://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/haushalte_familien_lebensformen/familien/index.html

[11] "Familie in Zahlen 2018": https://backend.univie.ac.at/fileadmin/user_upload/p_oif/FiZ/fiz_2018.pdf

and partners above the commonly expected limits (e.g., more than two co-insured spouses and more than four children) can be identified as outliers based on these graphs, no clear thresholds can be obtained. Additionally, the small group of persons with a larger number of partners or children still has the potential to represent actual cases.[12] As a result, no further filters are applied, and no persons are excluded due to the number of partners or children.

2.2.4 Socio-Economic Status (SES)

A measure for the socioeconomic status (SES) is used in this study, which was derived by the Main Association of Austrian Social Security Institutions from the corresponding index variables in GAP-DRG and is called sozioökonomischer Status (SÖS) within GAP-DRG.[13]

The data provided are linkable but not integrated into the core schema of the GAP-DRG database. They consist essentially of three variables, a past socioeconomic status indicator, current status (according to the main dataset), and their average as a summary. One of the first two variables may be missing a priori, but the summary is expected to be filled in completely. If both variables describing the past and the present are missing, a person will not be included in the SÖS dataset as a whole.

The entire spectrum of possible SES values lies continuously between 1 and 4, with a higher value corresponding to a worse status. It can therefore be read as an index for the *social burden of disease*[14]. As a rule of thumb, a difference of about 0,2 in SES scores can be interpreted as a relevant disparity.

Not only is the summarizing third variable unexpectedly missing, but the pattern of missing values is particularly noteworthy. In each case, one of the first two values is unknown, and their mean value is also missing. This could be an error introduced by the special way databases usually handle missing values. For the present project, the true average of the provided SES, called variable *soes*, is calculated and used in the following steps.

In addition, because of its origin, SES is not available for the entire population, but only for a nonrandom subset. Altogether, it is missing for 142.048 $(7,31\%)$ persons of the *cohort no-NA*[15]. To compensate for this, the average SES per individual is calculated based on all values in each individual's relationship network (with a maximum distance of 2), called variable *soes_mean*. This interpolated SES is missing for only 2.083 $(0,11\%)$ persons of the *cohort no-NA*[16].

[12] For example, self-employed farmers and small family enterprises where several family members, possibly from multiple generations, show tight relationships and are co-insured with each other in different constellations over time.

[13] Details concerning *SÖS* are provided by the *Main Association of Austrian Social Security Institutions*: https://www.sozialversicherung.at/cdscontent/?contentid=10007.844151.

[14] *soziale Gesundheitsbelastung*

[15] 142.048 $(7,16\%)$ of *cohort full*

[16] 2.042 $(0,1\%)$ of *cohort full*

2.2.5 Comorbidity and morbidity score

Multimorbid conditions are known to alter risk factors for disease (e.g., [Starfield, 2006]). Thus, in case comorbidity is not evenly (randomly) distributed between compared cohorts, it is a potential source of bias. To control for this likely propensity, which is presumably correlated with age, sex, cohort assignment, and outcome, information on additional diagnoses is collected and pooled.

After a brief introduction to multimorbidity in general, this chapter presents the selection of the data source and three different approaches to calculate a univariate score representing the individual burden of disease. Following the description of socioeconomic status, rationales and estimates are discussed according to the available data and the scope of the study, although no clear, universal conclusion can be drawn in the context of this analysis.

Poorer health, more expensive and complex treatment in combination with inferior results is generally linked to patients in a multimorbid state. Various terms such as burden of morbidity, patient complexity, and multimorbidity describe similar conditions but are not clearly and consistently defined [Valderas et al., 2009].

In this project, the term *morbidity* is used to describe the (co-)occurrence of one or several varying or identical diagnoses encoded according to the *International Classification of Diseases*, 9th revision (ICD-9) or 10th revision (ICD-10) per person during 2 to 6 years. Hence, there is no defined primary medical condition with which other diagnoses co-occur and no general knowledge about known diagnoses which have not been recorded directly or indirectly during the observation period, but single or multiple morbidities per person [Jakovljević and Ostojić, 2013, van den Akker et al., 1996]. Therefore, the simultaneous presence of diseases without any assumption about causation, chronological order, or interdependence is gathered and summarized.

The *condition* of a patient can be described by various details, ranging from terms and concepts depictable by classification systems, up to properties which can be measured or estimated like the SES and personal attributes to non-health-related environmental factors, called a patient's individual *complexity* [Safford et al., 2007]. Furthermore, there are various documentation and classification systems, such as the *International Classification of Diseases* (ICD), the *Diagnostic and Statistical Manual of Mental Disorders* (DSM), or the *International Classification of Primary Care* (ICPC). In addition to medical diagnoses, there are unclassified diseases, disorders, conditions, illnesses, or health problems which are not diagnosed, encoded, or represented in these systems as such.[Valderas et al., 2009] The WHO defines health as “a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity”[17] which evidently reaches beyond a mere collection of recorded and encoded diagnoses. Differentiating the source of records,

[17] Preamble to the Constitution of WHO as adopted by the International Health Conference, New York, 19 June - 22 July 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of WHO, no. 2, p. 100) and entered into force on 7 April 1948. The definition has not been amended since 1948.

the nature of the conditions as well as the shortcomings of the available data is therefore critical to enable the interpretation of results in the right context. [Valderas et al., 2009]

In this study, ICD-9 diagnoses derived by the $ATC \rightarrow ICD$ project [Filzmoser et al., 2009] are utilized. In contrast to more reliable and precise diagnoses from inpatient data, this data source provides a much broader coverage of the entire population due to its origin from reimbursed prescriptions. All diagnoses available in this dataset are extracted from the GAP-DRG database, resulting in a list of three-digit ICD-9 diagnoses per person in the study cohort.

As a result, there are no diagnoses for persons without a known medication. *Known medication* refers to the table *rezeptdaten* in GAP-DRG. It holds reimbursements of prescriptions covered by social insurance institutions without privately paid prescriptions and medications sold at a lower price than the current prescription charge, except for persons disengaged from this fee (and other types of co-payment). The prescription charges in 2006 and 2007 have been € $4,6$ and € $4,7$ respectively.[18]

Mixing different sources of diagnoses (i.e., $ATC \rightarrow ICD$, inpatient data, and sick leaves) is not feasible due to different coding systems (ICD-9 and ICD-10), increased complexity of the interpretation and the interdependence of these data sources.

Due to missing comprehension of the actual medical incidence or the historical succession of events in general, the outcome *myocardial infarction* is not necessarily occurring after the observed predictors as, e.g., comorbidity. As a result, concurrent medication and hence diagnoses have to be reduced by diagnoses directly related, i.e., from the same ICD-9 chapter, to the outcome because their occurrence might not be independent but the cause of the event itself. Furthermore, statistical analysis has been performed with and without information on multimorbidity to reduce potential distortions caused by unclear sequence of events and observation periods.

The resulting lists of diagnoses are still difficult to include directly into the analysis due to their varying length and miscellaneous concurrence of diseases. Therefore, a summarizing score is calculated based on commonly used comorbidity indices for administrative data. Comorbidity measures or comorbidity indices are most commonly developed to predict the risk of death from comorbid disease(s) after hospital discharge. Nevertheless, they are also utilized in studies dealing with long-term outpatient mortality and morbidity. According to [Sharabiani et al., 2012], the most common scores are the *Charlson index* [Charlson et al., 1987], its adaptation developed by Deyo R. A. [Deyo et al., 1992], and the Elixhauser comorbidity measure [Elixhauser et al., 1998]. For this study, the original Charlson index and the more recent update and validation from Quan et al. [Quan et al., 2011] as well as a point system based on the Elixhauser comorbidity measure developed by van Walraven et al. [van Walraven et al., 2009] (as used by the US AHRQ) are implemented.

[18] Prescription charges listed by the official representation of pharmacies in Austrian (österreichische Apotherkerkammer): `https://www.apotheker.or.at/Internet/OEAK/NewsPresse.nsf/webPages/D2FCB57ED4671F75C1256F2C005D30F3`

These scores are examined and described. Finally, one score or derivative will be included in the data analysis and for matching of the cohorts, as suggested by the study protocol.

The selection of the calculated scores is based on a systematic review [Sharabiani et al., 2012] and the implementations provided by [Wasey, 2016, McCormick and Joseph, 2016].

In general, these scoring systems group the single ICD diagnoses into different groups, e.g., so-called Charlson comorbidities. These groups are defined in the according publications and do not cover the entire spectrum of available codes. Hence, updates like developed by Quan et al. [Quan et al., 2011] adapt the system to better reflect morbidity and mortality in a more recent population. For each of these groups of diagnoses, scores are attributed. To measure the wellbeing of a patient or the severeness of a multimorbid condition, all groups of diagnoses can be used separately or accumulated to a single score.

Alternative approaches: updated scoring algorithms

Two diverging alternative possibilities for obtaining information about multimorbidity have been tested and finally discarded during this study.

More recent methods to score and summarize ICD-10 diagnoses like [Quan et al., 2005], describing ICD-10 versions of the Charlson (with Deyo's coding algorithm) and Elixhauser comorbidities, as well as [Sundararajan et al., 2004], another translation of Deyo's version of the Charlson comorbidity score, have been evaluated. Sundararajan et al. rate both to be rather equal with a slightly better performance of the second one in [Sundararajan et al., 2007]. Furthermore, a new ICD-10 version of the *Multipurpose Australian Comorbidity Scoring System* (MACSS) [Toson et al., 2016], which outperforms both Charlson and Elixhauser comorbidities according to the authors' measurements, has been applied to the data available, utilizing coding tables provided online as supplementary files to the publication.

Two main obstacles have been met for all three alternative approaches, leading to their final discard. Both the availability of ICD-10 diagnoses in GAP-DRG and the final scoring of the derived comorbidities remained unsatisfactory. A similar scoring as for multimorbidity derived from ICD-9 diagnoses might be appropriate, although an increased performance of models incorporating individual comorbidities in comparison to weighted scores is mentioned by [Toson et al., 2016]. More significantly, ICD-10 diagnoses are only available from hospital discharge information in GAP-DRG. On the one hand, these diagnoses match the original source of the presented scores more directly than $ATC \rightarrow ICD$, on the other hand, they are much scarcer and only cover a small proportion of the population at hand. Furthermore, the contribution of the main and additional diagnoses and their weighted proportions to a person's individual multimorbidity score is unclear and has never been discussed before for the Austrian reimbursement system.

Summarizing, the evaluated multimorbidity measures based on ICD-10 are translations of their predecessors which have been originally developed for ICD-9. They appeared to be inappropriate for the study and application at hand mainly due to the small coverage of ICD-10 diagnoses for the selected population in GAP-DRG.

Alternative approaches: incorporating pharmacy data

Furthermore, methods incorporating pharmacy data in a more direct way are evaluated solely based on selected literature. While $ATC \rightarrow ICD$ provides an automatic and therefore relatively objective deduction of ICD-9 diagnoses from reimbursed prescriptions, these alternatives are based on translations defined by experts. Comorbidity scores presented above are developed utilizing diagnoses from hospitals and are mostly meant to estimate short-term mortality. In comparison, approaches based on pharmacy data tackle the detection of chronic diseases in larger populations. They are included due to their apparent similarity to $ATC \rightarrow ICD$ and to cover all possibilities to estimate a multimorbidity index with the data available.

Starting with a publication from Chini et al. [Chini et al., 2011], which has been thoroughly discussed in several studies utilizing data from GAP-DRG, associated literature is examined. While [Chini et al., 2011] is promising at first sight, there seems to be at least one typing error[19] in *table 1*, listing all applied coherence between ATC and groups of chronic conditions. Furthermore, the exact selection of included and excluded codes does not seem to match the Austrian reimbursement system.[20]

In other studies, such as [Cricelli et al., 2003, Böhm et al., 2013, Katschnig et al., 2012], only very selective sets of diagnoses are included, lacking the demand of generalizability and wide coverage of health-related conditions. This approach is also common in studies based on data from GAP-DRG, where input from (medical) specialists is feasible for a few conditions. In contrast, the morbidity score utilized in this study aims to summarize all kinds of diagnoses by providing a rough estimate instead of a precise valuation of the overall condition of a person.

Furthermore, national healthcare and reimbursement systems are often key components for the deduction of diagnoses from pharmacy data. For example, in Italy a specific medication is only allowed for reimbursement in case a respective diagnose is present, as described in [Chini et al., 2011, Maio et al., 2005]. As a result, the inference is much more reliable than in Austria where these rulings do not exist.[21] A similar situation can be observed in [Huber et al., 2013] for the Swiss healthcare system. As a result, applicability and comparability cannot be assumed by default.

Summarizing, diagnoses deducted from reimbursed medications are utilized in many published studies and are therefore understood as a common approach to compensate for missing information. Applying translation tables from other projects and countries to the setting at hand would require more in-depth review and involvement of experts,

[19] The ATC codes for *Psychiatric disorders* lists the code *N05AA* (Phenothiazines with aliphatic side-chain) twice. According to related sources ([Von Korff et al., 1992, Clark et al., 1995]) referenced in [Chini et al., 2011], *N06AA* (Non-selective monoamine reuptake inhibitors) is most likely meant.

[20] e.g., *A10BX* (Other blood glucose lowering drugs, excluding insulins) is listed for *Diabetes*, but does not appear in Austrian's pharmacy data at all.

[21] There are other rulings concerning reimbursement ranging from the requirement of prescriptions from medical specialists to the so-called *Chefarztplicht*, a system established in 2005 to control medication by health-economic criteria, listed in the *Erstattungskodex*.

which is not viable in the context of this project. Due to the need for a broad estimation of the burden of disease for a large part of the entire population, the diagnoses provided by the Austrian $ATC \rightarrow ICD$ project appear fitting. Additionally, they have already been evaluated [Filzmoser et al., 2009] and experience showed that this information is well established and accepted.

2.3 Statistical analysis

According to the study protocol, the measured and recorded (dichotomous) primary diagnosis of myocardial infarction (ICD-10 codes I21 and I22 with subcodes) in 2006 and 2007 will be examined. The occurrence of such an event is identified by linked inpatient episodes. Because of this rather limited approach, the limited longitudinal information (a total of two years of data are available), and the absolute rarity of the events considered, model performance and an estimate of sensitivity and specificity are important. In addition, covariates of individuals such as age, sex, socioeconomic status, and multimorbidity score are included.

Cohorts are compared using various univariate and multivariate statistical techniques. Covariates are used to describe differences between cohorts and as independent variables in regression models. Pairwise matching of individuals is used to prevent confounding due to matched variables. Although the variables available for matching may intermediate slightly between exposure and outcome, it is expected to reduce bias. Crude rates and statistical tests will be used to compare the resulting matched cohorts.

As described earlier, the intention of this study is to investigate whether evidence can be found that parents are more likely to have an early myocardial infarction. If evidence can subsequently be found to support the initial assumption, it could be used as an additional factor in clinical guidelines, as pointed out by the medical advisor to this study.

Therefore, several methods are applied. First, cross-tabulations of raw numbers and subgroups are presented to document the setting. Next, logistic regression models are chosen as one of the most common and understood methods for dichotomous dependent variables. To fit the overall setting and goal of the study, a binary decision tree algorithm *fast and frugal tree*, focused on decision making and communication in medicine is applied. Next, gradient-boosting machines, an ensemble of several rather simple so-called weak learners, are used as an example of a typical black-box model where lower error rates are expected. In combination with gradient boosting machines, various techniques for balancing the dependent variable are applied and discussed. Finally, propensity score matching is used to reduce bias.

Various caveats are applied to obtain valid model performance indicators. In addition to the various measures, cross-validation and balancing are used.

2.3.1 Descriptive statistics and tests

The first section of the statistical analysis includes univariate and multivariate descriptive statistics and tests for the selected cohorts, as suggested in the study protocol and by the study advisors.

Beginning with a rough outline of the final dataset, the number of observations, excluded persons, and relationships are summarized. The development of the final dataset from the entire population of Austria and the content of the database GAP-DRG, the co-insured population down to the selected individuals and cohort assignment are summarized as a flowchart, inspired by the *PRISMA Statement* [Moher et al., 2009], as depicted in figure 3.73 on page 124. The *PRISMA Statement* is a specialized reporting tool for systematic reviews, meta-analysis, and other types of research like the evaluation of intervention. The resulting flowchart gives a comprehensible overview over the formation of the final cohort, applied selection criteria, and the size of the populations involved at every step.

Univariate statistics and tests for the entire working dataset and various contrasting comparisons, split by cohort assignment, outcome, gender, and age groups, as well as a combination of several variables are presented and discussed in chapter 3.5.1. The resulting tables include test statistics in the rightmost column. The Wilcoxon signed-rank test [Wilcoxon, 1945] is applied for continuous and the Pearson's chi-squared test [Pearson, 1900] for categorical variables. All univariate tests are expected to show highly significant differences between the cohorts due to the large size of the observed population. Nevertheless, these test statistics are relevant to interpret the relative size of (univariate) differences and to gather first hints about the potential importance of single variables.

In addition, the study protocol proposes cross-tabulation by cohort, sex, and age groups to represent the prevalence and prevalence rates of myocardial infarction. The results are discussed in 3.5.2.

It should be noted that arbitrary age groups are listed in the protocol, including groups outside the defined range that are used as selection criteria. Specifically, the protocol mentions two age groups, 60-64 and 65-69, that are mostly or entirely outside the defined range of 30 to 60 years. Instead, more granular age groups of 5[22] years are used.

According to the study protocol, tables for each cohort are faceted by sex, while age groups are used as rows. In addition, the sum of both genders is included. Rates are presented as cases per thousand population (‰).

2.3.2 Logistic regression

Multivariate logistic regression models are calculated. They allow to evaluate the magnitude of the influence of each independent variable including its own and the overall significance of each model. Interaction between variables (e.g., age and gender), direct comparison of related models, and confirmation of results by cross-validation are included.

[22] 6 years are used for the highest stratum of 55 to 60 years

Two different models are estimated. In addition, one variation of both models includes an interaction between gender and the corresponding age variable, resulting in a total of four variants.

Model 1 includes the independent variables sex, cohort assignment, socioeconomic status, grouped Charlson score with three distinct characteristics, and age in 2007 divided by 10, called *age07_10th*, to ease the interpretation of resulting coefficients

Model 2 contains the same regressors except for age, which is replaced with 5-year age groups, starting at 30-34 through 55-60, as suggested in the study protocol

The resulting coefficients (i.e., log-odds) transformed into odds ratios and their 95% confidence intervals are presented as tables and forestplots. Odds ratios are chosen instead of log-odds to facilitate interpretation of effect sizes.

These forestplots allow a quick visual assessment of the results of a model. To increase readability, the axis is scaled accordingly and limited to a maximum of up to 2,5. Profile likelihood confidence intervals are determined with the R function *confint.glm* from the MASS package [Venables and Ripley, 2002b]. The very low axis intercepts, which represent the general improbability of being affected by a positive outcome, are not included.

Stepwise model selection with both forward and backward searches was performed to find the optimal variables, using AIC as the benchmark. In each case, no better version than the full model with the total number of variables can be identified.

In addition, k-fold cross-validation is applied to capture out-of-sample performance measures. To do this, the data is randomly partitioned into 100 exclusive partitions for testing and training. To obtain reproducible results, a fixed seed is defined and a different partition is created for each model. Then, 100 models are computed for each set of variables by omitting one partition at a time. Thus, each partition is omitted exactly once and can be used as a test set to calculate model predictions and performance indicators[23].

Several performance indicators are collected for each model. The dispersion of the out-of-sample prediction quality of the models can be observed by visualization of the individual ROC curves [Fawcett, 2006]. In their clustered version, the dispersion is presented as boxplots, including a median of the ROC and the corresponding AUC with information on the dispersion of these measures. In addition, the distribution of each estimated coefficient is presented, including their median and dispersion. A correlation matrix of the estimated coefficients with the AUC of the corresponding model provides additional impressions.

[23] to save memory and space, specialized objects from the R extension ROCR [Sing et al., 2005] are utilized

2.3.3 Decision Trees

A *Fast and Frugal Tree* (FFTree) is a set of bivariate rules for making decisions with rather few variables [Gigerenzer and Brighton, 2009]. FFTrees are transparent and easy to apply and interpret [Gigerenzer and Todd, 1999, Gigerenzer et al., 1999]. Especially in medical decision making, simple decision strategies and heuristics lead to faster and better decisions than an overwhelming amount of data and complex models [Marewski and Gigerenzer, 2012].

This methodology has been developed as a decision support tool when resources like time are limited and where the accuracy to do the right thing is required to be high [Martignon et al., 2008]. Because of its simple, adaptive, and deterministic heuristics, the authors propose FFTrees for high-risk environments such as accident and emergency departments, where physicians must decide whether or not a patient is likely to have a disease.

In [Martignon et al., 2008], the predictive performance of FFTree is measured by simulation and compared to common benchmarks of machine learning algorithms. They showed that especially the predictive accuracy is high for the focused classification tasks.

Hence, FFTrees fit the protocol and intention of this study well. As this algorithm is not widely used, the resulting models are directly compared to more common classifiers like logistic regression and *Classification And Regression Tree* (*CART*) [Loh, 2011].

Four FFTree models are calculated and presented in section 3.5.4.

reference The reference model is a very basic tree, only involving two variables, age and sex.

small For the small tree the variable of interest *cohort* is added.

full In the full tree also the Charlson groups and the SES (variable *soes_mean*) are included. The pruning heuristic is mostly deactivated because cohort assignment would be left out otherwise as most unimportant variable for decision making.

full default In comparison to the full model, the default parameters of the FFTree routines are left unchanged. As a result, pruning is applied and only the four most important decisions are included, lacking cohort assignment.

In each case, multiple FFTrees are computed using cross-validation and the best result is selected. As suggested by the authors and consistent with the intent of FFTrees, the results are presented as rich visualizations like a one-page handout. Nevertheless, many relevant components such as reference models, ROC curves that are cross-validated with other performance statistics, a summary of the test data, a structured presentation of the resulting decision tree itself, and illustrations of the discriminatory power of each decision are included in a compact and appealing format.

For the reference model, the two resulting trees are included in the plot to show their similarities and differences. Their relative performance to other trees in the same model and to a logistic regression classifier (LR, blue dot) and a CART (red dot) is shown in the ROC plot in the lower right corner of the corresponding figures.

In addition to the ROC curve described above, there are several other components in the FFTrees visualization. The topmost area summarizes the provided (test) data. Then, the central part, labeled with the number of the selected result itself, visualizes the decision tree. On the left side, the observations are classified as not affected, i.e., no mi, while the positive cases are on the right side. These decisions are also visualized using the test data provided. The bottom section contains various performance statistics. Starting with a contingency table for the test data, several model summaries and an ROC curve can be found in the lower right corner. It is noteworthy that the same colored icons[24] are used for visualizing results of the decision tree and in the summarizing contingency table.

Four performance measures can be found between the contingency table and the ROC plot. The first two are specificity (abbreviated *Spec*), i.e., the true negative rate, and sensitivity, labeled as *Hit Rate*, which is the true positive rate, also known as recall. Next, the estimated sensitivity index ([Vision, 1985]) d', labeled as D', represents the difference between sensitivity and specificity. It indicates the quality of detecting a true signal. Last, the area under the ROC curve is listed as the main performance measure not only for the selected tree but for the entire model.

All models are trained and tested on the same set of data. The training data is a 75% randomly selected sample from the entire dataset. All remaining observations are utilized as test data. The sample is stratified for the variable cohort. Therefore, 75% of each cohort *intervention* and *control* are selected for training.

2.3.4 Model performance measures

Several different performance and outcome measures are calculated for the FFTree models, including:

N.train number of observations in the training dataset

N.test number of persons used for testing

contingency table consisting of the following content:

TP **T**rue **P**ositive, correctly classified events

FP **F**alse **P**ositive, incorrectly classified as event

TN **T**rue **N**egative, correctly classified as no event

[24] circles represent true negatives and triangles true positives; green stands for correct classification and red for a wrong prediction

FN **F**alse **N**egative, incorrectly classified as no event: these observations have an unrecognized positive outcome

AUC **A**rea **U**nder the ROC **C**urve

sens recall, HR sensibility, recall, **H**it **R**ate: $TP/(TP+FP)$

spec FAR specificity, **F**alse **A**larm **R**ate: $TN/(FP+TN)$

PPV, precision **P**ositive **P**redictive **V**alue, precision, rate of TP from all selected outcomes: $TP/(TP+FP)$

NPV **N**egative **P**redictive **V**alue: $TN/(FN+TN)$

FPR **F**alse **P**ositive **R**ate, fall out: $FP/(FP+TN)$

FNR **F**alse **N**egative **R**ate: $1 - sensibility$

FDR **F**alse **D**iscovery **R**ate: $FP/(TP+FP)$

ACC **ACC**uracy: $(TP+TN)/(TP+FP+FN+TN)$

F1 F1 score, harmonic mean of precision and recall: $(2*TP)/(2*TP+FP+FN)$

Prevalence rate of cases in test dataset: $(TP+FN)/(TP+FP+FN+TN)$

DetectionRate rate of correctly selected positive cases in entire test dataset: $TP/(TP+FP+FN+TN)$

DetectionPrevalence rate of correctly recognized positive and negative cases in entire test dataset: $(TP+FP)/(TP+FP+FN+TN)$

BalancedAccuracy average of sensibility and specificity: $(sensibility+specificity)/2$

Kappa κ statistic, a measure of rater agreement between the actual outcome in the test dataset and the classifier's predictions [McHugh, 2012, Tang et al., 2015].

2.3.5 Gradient Boosting

Generalized Boosted Regression Modeling following Friedman's *Gradient Boosting Machines* (GBM, [Friedman, 2001]), are applied next. Additionally, several approaches to balance the outcome variable are introduced.

Results from the GBM cannot be interpreted as directly as the coefficients from logistic regression or binary decisions from the FFTrees. Nevertheless, boosting does not sustain from overfitting as many other methods and is a "general method for improving the accuracy of any given learning algorithm" [Schapire, 1999]. It is described as "techniques to obtain smaller prediction errors (in regression) and lower error rates (in classification) using multiple predictors" [Drucker, 1997]. Schapire states in [Schapire, 2003] that "logistic regression and boosting are in fact solving the same constrained optimization problem, except that in boosting, certain normalization constraints have been dropped".

It can therefore be expected that the GBM models perform at least as well as the logistic regression model. In case the rarity of the outcome event has a significant impact, balancing might even show an improved performance.

Gradient Boosting Machines as defined by [Freund and Schapire, 1997, Friedman et al., 2000, Friedman, 2001] are a gradient-descent based formulation of boosting methods according to [Natekin and Knoll, 2013]. Boosting implies that new and improved models are computed iteratively by adding basis functions to the previous ones. Each step is evaluated by an arbitrary loss function, resulting in a more accurate prediction. The area under the ROC curve (AUC) introduced earlier is used as the loss function in the presented application. ROC curves and the AUC are also used for the final plot and comparison of the resulting models. Boosting thus sequentially applies a classification algorithm to (reweighted) versions of the training data, which according to [Friedman et al., 2000] leads to dramatic performance improvements in most cases.

Since there are two groups to be compared, the binary outcome variable is modeled as a Bernoulli distribution, analogous to logistic regression. Alternatively, the exponential loss function AdaBoost and huberized hinge loss are supported by the R software package *gbm* for binary dependent variables. They are also evaluated on the present data, but do not show improved performance compared to models based on the Bernoulli distribution. Using the software package CARET [Kuhn, 2008], repeated 5-fold cross-validation is applied to estimate the GBM models and their validation error hyperparameter described in [Natekin and Knoll, 2013]. These procedures are run in parallel on 15 processors (without load balancing), resulting in a significantly lower total training time than in a single-core implementation.

Prior to this, the entire data set is randomly split into a 75% sample for training and the remaining 25% sample for testing. All predictors are trained and tested using the same data sets, with only the training set being further split for cross-validation. Thus, the test data allow for true out-of-sample validation.

2.3.6 Balancing

A balanced variable implies that its classification categories are almost equally distributed. In contrast, a dataset is imbalanced when a common class predominates and is understood to be the default setting, while only a small fraction of special cases are included, as pointed out in [Chawla et al., 2002]. This imbalance is (fortunately) the case for myocardial infarction, especially in younger individuals.

The problem with unbalanced outcome variables is that the cost of misclassifying a rare (but potentially interesting) sample is often much higher than the cost of doing the reverse. On the other hand, if one of the two classes occurs only at x%, the maximum error rate for the case where all labels are classified with the dominant label is also x%, which should be rather small in contrast to the misclassification of a small fraction of the more common category.

Four different approaches to balance the training data are applied, resulting in 5 models (including the original, unbalanced one) for each set of variables.

model 1 & 2 are two previously introduced model setups, where only variables describing age differ. All the following balanced models are also implemented in these two versions, including an appropriate indicator.

weighted additional cost is added in case a classification error is present in the minority class. The applied cost is directly proportional to the relative frequency of myocardial infarction in the training dataset.

down sampling abbreviated *down*: randomly removes cases from the dominating class.

up sampling abbreviated *up*: randomly replicates persons in the scarcer class.

SMOTE *Synthetic Minority Over-sampling TEchnique* [Chawla et al., 2002]: down sampling and at the same time, creating artificial cases in the minority class by interpolating between existing ones

Not only an optimally performing classifier but also the potential differences of these balancing approaches concerning the outcome measure AUC and the ROC curves are of interest. Especially *SMOTE* appears to be fitting well because it is stated to show better performance in ROC space, which is also utilized as a loss function for boosting [Chawla et al., 2002]. Although threshold-dependent metrics like sensitivity and specificity might gain most from balancing the training dataset and thereby moving the ROC curve to its optimum, also an improved AUC can be expected.

These additional models are only trained with GBM. The main objectives are the evaluation of the impact of balancing and different balancing algorithms in contrast and to assess the overall impact of balancing on out-of-sample model performance. Results are compared only using the AUC of each variant. Due to the insignificant differences, the balanced datasets are not evaluated with any other algorithm.

All resulting ROC curves are visualized in figure 3.88 on page 146. Although the optimized viridis color palette [Garnier, 2016] is applied, the single models cannot be distinguished. In detail, slight variations can be spotted but their overall result is practically identical.

2.3.7 Propensity score matching

In observational studies, unlike randomized trials, compared groups may differ by more than chance, which is referred to as bias. While bias in observed variables can be detected and accounted for, bias in unobserved background variables remains unknown in most cases. As a result, unaccounted bias can bias the data and models. Therefore, estimation of causal effects in observational studies is not possible. [Stuart, 2010]

Naturally, it is practically and ethically impossible to conduct a randomized study for the research question of this project. In particular, when variables are not primarily collected and predefined in the study design, but available data are used secondarily, it is not possible to remove unobserved bias in observational studies. Nevertheless, adjusting for observed confounders reduces bias and increases the quality of results and conclusions. In [Rubin, 2004], the authors state that it is required to balance the observed variables at least on average between the groups. They suggest propensity score matching in combination with blocking of the most relevant covariates to achieve this balance. As a result, unobserved covariates can be expected to be less biased on average and propensity score methods are therefore a fitting improvement of observational studies, despite they are still inferior to real randomization concerning causal inference.

Although matching on observed covariates will improve the balance of the dataset, unknown and unobserved confounders cannot be controlled at all. As a result, conclusions about causal inference can still not be achieved. Furthermore, the quality of observed confounders is critical as stated by [Arnold et al., 2010]. This is also the case for this project, where incomplete information on, e.g., family relations, comorbidities, and socioeconomic status (SES) is derived from administrative claims data.

According to the study protocol, the two cohorts have to be matched on individual SES, age, and gender. It has been agreed that *propensity score matching* is the method of choice to accomplish this prerequisite.

Therefore, the matched cohorts are balanced on the variables SES, age, and gender, leading to the assumption that the results are not influenced by these covariates anymore. Matching on the clustered comorbidity score is expected to reduce potential bias even further.

Several software implementations providing general matching routines are available in the R ecosystem. Two specific solutions, *MatchIt* [Ho et al., 2011] and *optmatch* [Hansen and Klopfer, 2006], are tested and applied in this project. Both support a variety of matching procedures including propensity scores, stratification (i.e., exact match, blocking), matching with and without replacements, a rich set of distance metrics and analytical functions. Flexible interfaces, rich documentation, and sophisticated integration into common R workflows are highlights of these packages.

The presented matching procedures consist of two major steps. First, the propensity scores are calculated for each individual using logistic regression. Cohort assignment is used as regressand while all confounders are used as regressors in the logistic regression model. Therefore, the propensity score can be interpreted as the conditional probability of a person being assigned to one cohort in comparison to its counterpart [Rosenbaum and Rubin, 1983, Austin, 2011]. Second, for each person's estimated propensity score, the most similar match is selected from the opposing cohort. Several matching functions, e.g., *exact matching*, *full matching* and *nearest neighbor matching*, can be used to determine the optimal pairs. Because the cohorts are not of the same size, the resulting population is drastically reduced to less than twice the size of the smaller group (in the case of one-to-one matching a maximum allowed distance). By allowing multiple selection of persons, i.e., replacement, the number of matched pairs can be increased, but single individuals are included several times.

Applying this procedure to small samples of around 20.000 persons of the study cohort succeeded in a few minutes without significant problems. Unfortunately, resource consumption and evaluation time increase drastically with growing datasets. Tuning the parameters of the distance function and testing different matching functions hardly improved the situation. Figure 2.1 shows the runtime of a propensity score matching procedure with logistic regression as distance function and nearest neighbor matching. Datasets with only a few thousand records are randomly simulated to circumvent the convergence warnings of the GLM model. Additionally, a linear regression model with a quadratic coefficient is fitted to the collected runtimes. Although some systematic differences between the fitted model and the plotted points can be observed, a slight quadratic coherence between the number of cases and runtime is obvious. Especially memory consumption increases quadratic with additional cases and exceeds the resources of the *GAP-DRG* application server quickly.

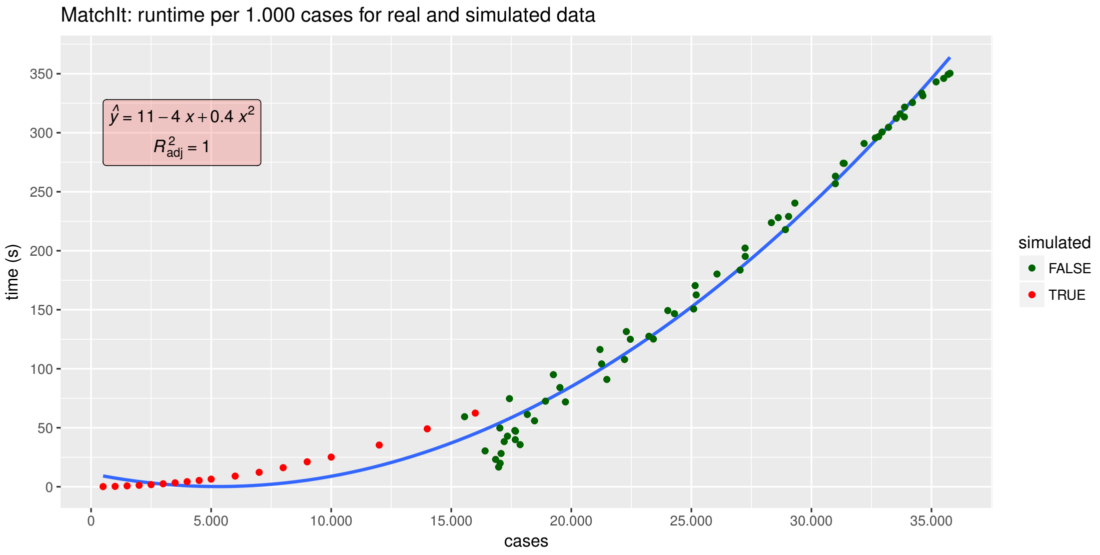


Figure 2.1: Propensity score matching: performance of R's MatchIt package

Digging into the inner structure of the matching function revealed that the previously described two steps are strictly separated. As a result, a matrix of all potential matches and their propensity scores is created and subsequently applied to the matching method. Stratification, maximal relevant distance, and other tuning parameters are also passed to the matching functions, although optimizations could be applied beforehand. On the one hand, this approach provides high flexibility where new matching and distance functions can be easily integrated and combined. On the other hand, the entire matrix of distances is calculated despite it would not be necessary in case exact matching, i.e., stratification, is applied to discrete variables.

As a result, it is not possible to apply the matching functions of the favored packages directly to the entire dataset. A simple algorithm is implemented by the author, which applies the matching procedure to each stratum defined by age and gender separately. As a result, the functionality of the *MatchIt* package can be directly utilized with the available resources. The only disadvantage is the resulting data structure which does not allow to examine and review the results directly because 62[25] sets are generated. Therefore, functions extracting and combining results have to be developed as well.

Four matched datasets are calculated utilizing these algorithms. They are stratified by age in years and gender. Two matched populations are created using replacement while each person is only included once in the two smaller ones. The logistic regression model used as the scoring function describes cohort assignment to the continuous variable SES as required by the study protocol. Furthermore, SES and the three comorbidity classes are utilized for scoring in the two result sets. Nearest neighbor matching of individual propensity scores in each stratum is applied. This algorithm resulted in very similar results as most other default matching criterions (despite exact matching) while minimizing resource consumption and calculation time.

In the results section, the sizes of the matched datasets are summarized by the cohort. Finally, the cohorts are compared as described in chapter 2.3.1 on page 23, revealing, on the one hand, how well the matching procedure balanced the datasets on matched and unmatched variables and, on the other hand, showing the differences of the outcome criteria between the groups.

[25] two genders times 31 age groups ranging from 30 to 60 years

CHAPTER 3

Results

3.1 Co-insurance: exploratory data analysis

In this section, the applied process and results from the data exploration and data quality assessment of co-insurance data are discussed in detail.

Altogether[1], 4.388.605 co-insured pairs with 2.386.052 unique insured persons (54.37% of all pairs) and 3.328.267 unique persons depending on another one (75.84% of all pairs), summing up to 5.240.670 unique persons can be identified.

Personal information is collected in two manifestations, one describing the insured and the other one the dependent person. This information might be missing (*NA* meaning *not available*). Further clinical data is not required for cohort selection and is therefore not collected or analyzed at this stage.

Starting with a univariate data profile of the extracted information, the most important variables are explored, checked for quality issues, and discussed in detail.

The following variables are involved:

birthyear, birthyear_insured year of birth

age07, age07_insured (approximate) age in 2007

sex, sex_insured gender / sex

deathyear, deathyear_insured year of death (exact date is also available)

[1] One person can be compulsorily insured and co-insured at the same time with differing or even the same person in alternating roles. Therefore, the total number of apprehended persons is smaller than the sum of unique insured and co-insured partners.

pop_forschung, pop_forschung_insured whether the person belongs to a standardized population in *GAP-DRG*[2]

age_difference absolute difference of years of birth between the insured and dependent person

hints_metadata how often this pair occurs in the insurances' metadata

hints_prescriptions how many prescriptions have been filled in this constellation: The number of packages per prescription or other details such as, e.g., the number of distinct dates are not considered.

hints_ambulatory how many hints from the ambulatory outpatient sector, i.e., registered doctors and specialists without outpatient departments of hospitals, are collected.

hints_inpatient number hospital discharges of any length from a public (LKF) hospital

3.1.1 Univariate data profile

A univariate summary of all variables listed above, depending on their type were conducted.

It is important to keep in mind that a single observation in the data set represents the relationship between two individual persons. Therefore, all counts refer to these relationships and not to individuals. Personal information (e.g., age07, gender) describes the dependent person unless otherwise specified in the variable name.

coinsured pairs
15 Variables 4.388.605 Observations

birthyear

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4267564	121041	131	0.999	1985	21.18	1941	1953	1976	1992	1999	2004	2006

```
lowest : 1879 1880 1881 1882 1883, highest: 2005 2006 2007 2008 2009
```

age07

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4267564	121041	131	0.999	22.08	21.18	1	3	8	15	31	54	66

```
lowest :  -2  -1   0   1   2, highest: 124 125 126 127 128
```

[2] The *Forschungspopulation* (research population) in GAP-DRG is a defined subset of the total population in the database. It is filtered depending on the quality constraints and the fact that there needs to be at least a single reimbursement recorded for each member. In subsequent steps, the analyzed population will be limited to this population.

sex

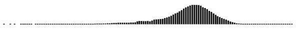
n	missing	distinct
4033293	355312	2

```
Value              F       M
Frequency    2355774 1677519
Proportion     0.584   0.416
```

pop_forschung

```
Value          FALSE    TRUE
Frequency     477113 3911492
Proportion     0.109   0.891
```

birthyear_insured



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4301568	87037	131	0.999	1962	13.62	1937	1945	1957	1965	1971	1976	1979

```
lowest : 1876 1879 1881 1884 1885, highest: 2008 2009 2010 2011 2012
```

age07_insured

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4301568	87037	131	0.999	44.58	13.62	28	31	36	42	50	62	70

```
lowest :  -5   -4   -3   -2   -1, highest: 122 123 126 128 131
```

sex_insured

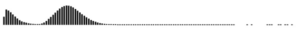
n	missing	distinct
4220016	168589	2

```
Value              F       M
Frequency    1713742 2506274
Proportion     0.406   0.594
```

pop_forschung_insured

```
Value          FALSE    TRUE
Frequency     320908 4067697
Proportion     0.073   0.927
```

age_difference



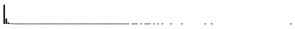
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4196967	191638	114	0.999	23.29	13.65	1	3	17	27	32	36	39

```
lowest :   0   1   2   3   4, highest: 123 124 126 127 129
```

hints_metadata

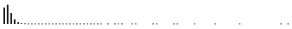
```
       n  missing  distinct   Info   Mean    Gmd   .05   .10   .25   .50   .75   .90   .95
 3432650   955955        20  0.898  2.156  1.415     1     1     1     2     3     4     5

Value            1       2       3       4       5      16      17      18      20      21
Frequency  1447339  967056  541934  242955  115937      43      18       5       1       1
Proportion   0.422   0.282   0.158   0.071   0.034   0.000   0.000   0.000   0.000   0.000
```

hints_prescriptions



```
      n  missing  distinct   Info   Mean    Gmd   .05   .10   .25   .50   .75   .90   .95
 853235  3535370       518  0.977     11  15.84     1     1     2     3     8    23    50

lowest :    1    2    3    4    5, highest:  777  829  937  966 1338
```

hints_ambulatory

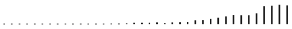


```
       n  missing  distinct   Info   Mean    Gmd   .05   .10   .25   .50   .75   .90   .95
 2493118  1895487       127  0.989  6.629  6.116     1     1     2     5     9    14    18

lowest :   1   2   3   4   5, highest: 398 402 410 411 412
```

hints_inpatient

```
      n  missing  distinct  Info   Mean    Gmd   .05   .10   .25   .50   .75   .90   .95
 190379  4198226        84  0.64   1.69  1.166     1     1     1     1     2     3     4

lowest :   1   2   3   4   5, highest: 101 104 112 116 134
```

deathyear



```
     n  missing  distinct   Info  Mean    Gmd   .05   .10   .25   .50   .75   .90   .95
 61415  4327190        38  0.992  2003  5.675  1992  1996  2001  2005  2007  2009  2009

lowest : 1972 1973 1974 1975 1976, highest: 2005 2006 2007 2008 2009
```

deathyear_insured

```
      n  missing  distinct   Info  Mean    Gmd   .05   .10   .25   .50   .75   .90   .95
 100678  4287927        38  0.982  2005  4.369  1995  1998  2004  2007  2008  2009  2009

lowest : 1972 1973 1974 1975 1976, highest: 2005 2006 2007 2008 2009
```

Several important facts are shown in this first data profile.

The distributions of the numerical variables initially appear to be usual for the data originating from GAP-DRG, although some outliers can be observed. Individual implausible values (which should be unusual in practice), such as age differences between insured persons and persons in need of long-term care of more than 100 years, are already easy to identify.

The same statement can be made for birth and death years. Occurrences such as year of birth before 1900 or co-insured individuals who died many years ago are clearly impossible but common problems with administrative benefit data. Fortunately, most outliers seem to be rather rare.

Interpreting these simple univariate data profiles from top to bottom, the following main observations can be extracted:

birthyear, age07 (of the dependent person):

- considering a recruitment period in the years 2006 and 2007, both the lowest and highest values are implausible
- these outliers are present but not dominating at first sight
- most persons can be classified as children and young adults (median at 15 years)
- there are spikes in the histogram which are known to be data artifacts located mostly at 1st of January
- the year of birth is missing for 121.041 (2,76%) relationships
  - because age is a major characteristic for selection and observation, individuals including their association with insured persons without a known age cannot be included in the study in the subsequent analysis
- as the age is calculated in 2007 and some records might be younger, negative values (i.e., persons born after 2007) are plausible

sex (of the dependent person):

- there are significantly more females than males recorded as dependent in a relationship
  - this observation is most likely different for children and grown-ups as it seems to be more common for female adults to be co-insured with a (male) spouse
  - apart from that, females are known to be single parents more often
- sex of the co-insured partner is missing for 355.312 (8,1%) couples

pop_forschung (of the dependent person):

- about 10,87% of all couples associated with a dependent person are not included in the research population
- the actual proportion of affected persons might be different as,

  - on the one hand, the data quality of persons not included in the research population tends to be worse and they might therefore occur in more - presumably wrongly assigned - relationships
  - on the other hand, there might not be any recorded reimbursement of a person not included in the research population which decreases the number of possible sources

birthyear_insured, age07_insured

- concerning outliers and implausible values, the same conclusion as for dependent persons can be drawn
- as expected, the median age of insured persons is slightly above 40
- the year of birth is missing for 121.041 (1,98%) relationships
  - because age is a major property for selection and observation, individuals including their association with dependent persons without a known age cannot be included in the study in subsequent analysis

sex_insured

- there are significantly more males than females recorded as insured in a relationship
  - this observation is most likely different for children and grown-ups as it seems to be more common for female adults to be co-insured with a (male) spouse
- sex of the compulsorily insured partner is missing for 168.589 (3,84%) couples

pop_forschung_insured

- about 7,31% of all couples associated with an insured person are not included in the research population
- again, the actual proportion of affected persons might be different
- overall, this value reflects a better data quality concerning personal information of insured persons in comparison with dependent persons

age_difference

- ignoring some very high values, two main, visually distinguishable groups can be identified
- the first group on the left-hand side with a smaller spread ought to refer to adult couples
- the second group refers to the relationship between adults and children
- means and medians are distorted

hints_metadata

- the most complete source of hints for a relationship
- the low number of total hints from metadata per couple results from the data source's construction

hints others

- appears to be less complete in comparison to hints from metadata
- large values due to the nature of the corresponding sources of information
- a large number of hints in several cases is especially of interest and might be utilized as a quality indicator for a link between two individuals

year of death for insured and dependent:

- implausible values recorded before the database ought to begin (2006)

In summary, it is relevant for the subsequent differentiation of cohorts that two focal points can be identified in the distribution of age differences between insured and dependent persons. They are expected to distinguish between (adult) partners and the relationship between parents and minor children.

There are significantly more females listed as dependent and more males as insured persons. Although this fact can be regarded as plausible, a deeper look into the age and gender distribution of pairs must be conducted. It is important to mention that there are (only) two values for the gender variables and unknown sex is encoded as missing.

Most of the individuals recorded are also members of the research population of the data source (*research-population*). This cohort does not play a major role in the genealogical part of the project, but will be important for later analysis. Nevertheless, it is interesting to note that the insured in this cohort tend to be more common than their partners. One possible reason could be hidden in the data source and its history. Since the claims data originate in the billing processes of Austrian social insurance institutions, the information on insured persons might be better or more complete compared to their co-insured dependents.

For most of the personal variables, a significant number of unavailable values (NA) can be observed. This missing information must be treated separately and with great care, especially since the following distinction of these couples into children, couples with children, and couples without children depends on the age of the persons involved.

Tables 3.1 on the next page, 3.2 on the following page, and 3.3 on page 41 give another slimmed down overview of the dataset distinguished by the data type (numerical and categorical). This additional summary is supposed to present a much denser overview than the first one above. In addition to the extracted information on the location

Table 3.1: Univariate summary of numeric variables

	n	mean	sd	iqr	nunique	nzeros	miss	miss%
birthyear	4.267.564	1.984,92	20,10	23	132	0	121.041	2,76
age07	4.267.564	22,08	20,10	23	132	110.124	121.041	2,76
birthyear_insured	4.301.568	1.962,42	12,62	14	132	0	87.037	1,98
age07_insured	4.301.568	44,58	12,62	14	132	31	87.037	1,98
age_difference	4.196.967	23,29	12,42	15	115	90.422	191.638	4,37
hints_metadata	3.432.650	2,16	1,46	2	21	0	955.955	21,78
hints_prescriptions	853.235	11,00	27,04	6	519	0	3.535.370	80,56
hints_ambulatory	2.493.118	6,63	6,05	7	128	0	1.895.487	43,19
hints_inpatient	190.379	1,69	2,46	1	85	0	4.198.226	95,66
deathyear	61.415	2.003,44	5,56	6	39	0	4.327.190	98,60
deathyear_insured	100.678	2.005,06	4,38	4	39	0	4.287.927	97,71

Table 3.2: Univariate summary of numeric variables: distribution of values

	min	1%	5%	25%	50%	75%	95%	99%	max
birthyear	1.879	1.926	1.941	1.976	1.992	1.999	2.006	2.008	2.009
age07	-2	-1	1	8	15	31	66	81	128
birthyear_insured	1.876	1.924	1.937	1.957	1.965	1.971	1.979	1.984	2.012
age07_insured	-5	23	28	36	42	50	70	83	131
age_difference	0	0	1	17	27	32	39	46	129
hints_metadata	1	1	1	1	2	3	5	8	21
hints_prescriptions	1	1	1	2	3	8	50	141	1.338
hints_ambulatory	1	1	1	2	5	9	18	27	412
hints_inpatient	1	1	1	1	1	2	4	10	134
deathyear	1.972	1.984	1.992	2.001	2.005	2.007	2.009	2.009	2.009
deathyear_insured	1.972	1.990	1.995	2.004	2.007	2.008	2.009	2.009	2.009

and distribution of the data, missing values and their potential impact are pointed out separately.

The following interpretations can be derived from these tables.

First, the data quality of age and gender seems to be better for the insured than for dependents. This fact could be influenced by the fact that insurance companies have more well-maintained information on their direct clients and dependent persons with poorer data quality are (falsely) associated with more insured persons, thus distorting this overview.

Table 3.3: Univariate summary of categorical variables

	n	miss	miss%	unique	freq
sex	4.033.293	355.312	8,10	3	F: 2.355.774 M: 1.677.519
pop_forschung	4.388.605	0	0,00	2	TRUE: 3.911.492 FALSE: 477.113
sex_insured	4.220.016	168.589	3,84	3	M: 2.506.274 F: 1.713.742
pop_forschung_insured	4.388.605	0	0,00	2	TRUE: 4.067.697 FALSE: 320.908

Second, the completeness and values of the various sources of co-insurance cues are essential to this analysis. While insurance metadata accounts for most (up to about 80%) of all relations, very few actual imputations are usually available from this source (the most frequent value, i.e., mode is 1, median is 2). About 20% of all pairs are inferred from other sources based on reimbursement procedures. In particular, outpatient contacts add a lot of information, while the frequency of leads from prescription data appears to be highly skewed. It is likely that this information interacts with other variables and, in particular, influences the extracted role of individuals (child, parent, spouse without children).

Basic descriptive statistics are presented in the following subsections. The selection of relevant properties is based on general knowledge about the GAP-DRG database, initial insights from the univariate profiling, and the scope of this study.

3.1.2 Sex / Gender

Relationships concerning the gender of insured and dependent individuals are displayed in figure 3.1 on the following page. Although this plot represents the magnitude of relationships and not of unique persons, it can already be stated that females are more commonly co-insured with males than vice versa.

It is anticipated that some people have more than one co-insurance documented. Even higher numbers of relationships per person are plausible for a smaller subset of individuals.

Figure 3.2 on the next page illustrates the number of occurrences (abscissa) of each depending person. The number of individuals is represented by the log10-scaled ordinate. There is one bar for each gender (of the dependent person). Additionally, a boxplot for all pairs is shown on top of the graph. Because there are single individuals with a very high number of relationships, the plot is limited to a maximum of 20 occurrences.

Interpreting the (log-scaled) bars, several coherences become clear. There seems to be large differences between dependent persons with (red, blue) and without (green)

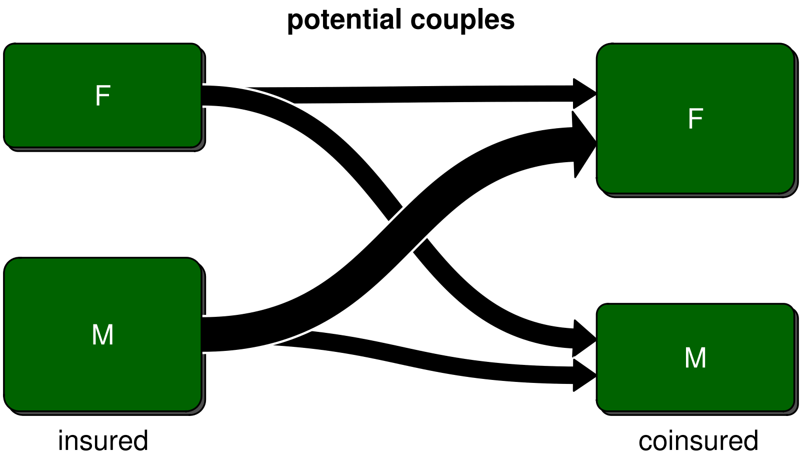


Figure 3.1: Flowchart of co-insurance depending on gender

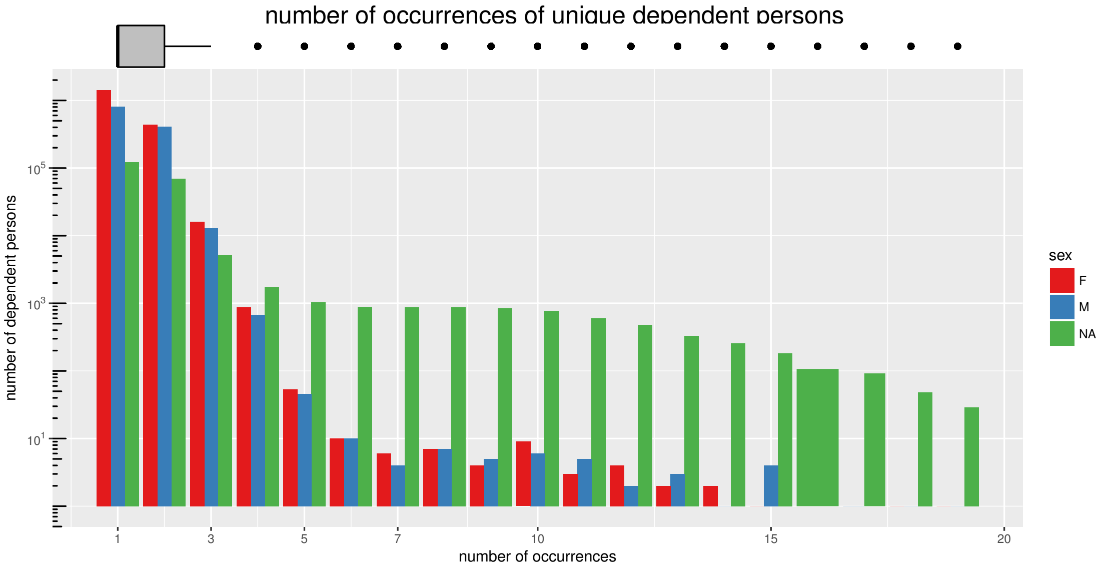


Figure 3.2: Number of relationships of unique, co-insured persons split by sex

a known gender. While the number of occurrences of the first group is decreasing exponentially from over one million individuals per sex for one relationship to about 50 for 5 relationships, people without a known gender tend to have more connections.

The boxplot indicates that most people have very few, mostly one or two, occurrences altogether. This fact is a relevant argument for the presented method of cohort selection.

It is important to mention that the log-scaled y-axis tends to hide some differences. The total number of potential relationships without a recorded sex for the co-insured person is 355.312 (8,1% of all relationships). Altogether, there are 207.496 unique co-insured persons lacking recorded gender.

Figure 3.3 zooms in on dependent persons with between two (beginning at 3) to 20 co-insurances without distorting the scale of the y-axis. As a result, the actual magnitude can be examined more directly.

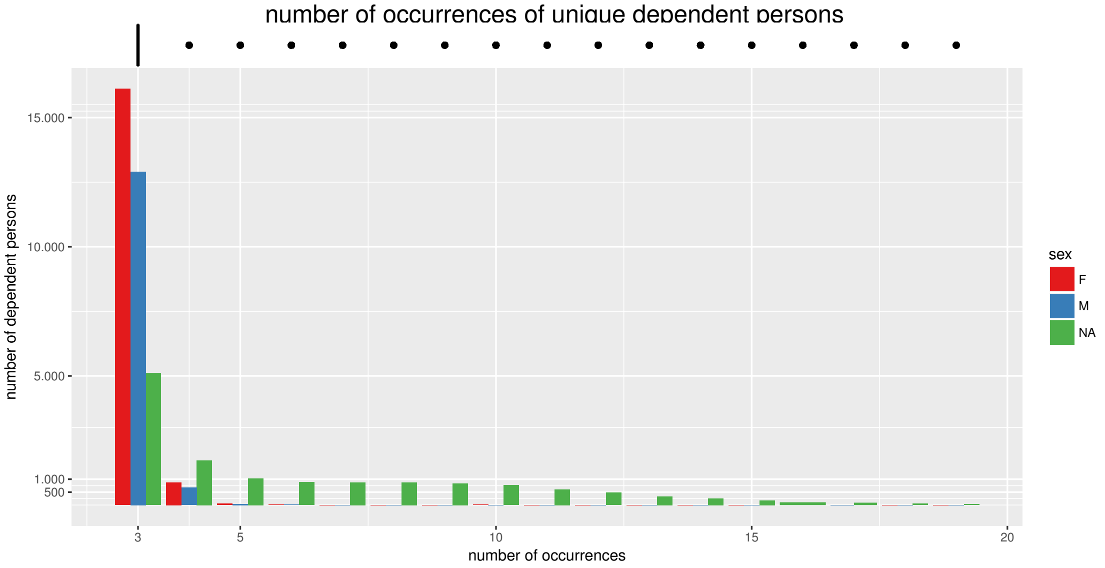


Figure 3.3: Number of relationships of unique, co-insured persons split by sex: absolute numbers

Figure 3.4 on the next page illustrates the same information for the second part of each pair, the insured persons.

Although the number of persons with an additional co-insured partner is decreasing exponentially, multiple occurrences seem to be more likely. There are still around 10.000 cases of insured people with five co-insurances, decreasing to 10 co-insurances for about 1.000 individuals. Moreover, at this point (10 occurrences), the number of persons without a known gender (green bar) begins to dominate.

The number of males in figure 3.4 on the following page is always higher than the number of females, which is the exact opposite in comparison to the barchart in figure 3.2 on the preceding page. Initially, it seems that males are more likely to be the insured part of a pair, ignoring the fact that co-insured children are still included and possibly skewing the results.

The exact meaning and reasons of the larger number of insured persons with more than a few co-insurances are not clear at this point and might be related to a very similar finding in section 3.1.5 on page 66 concerning the standardized research population. It can be speculated that the total number of co-insurance is also correlated with the profession of

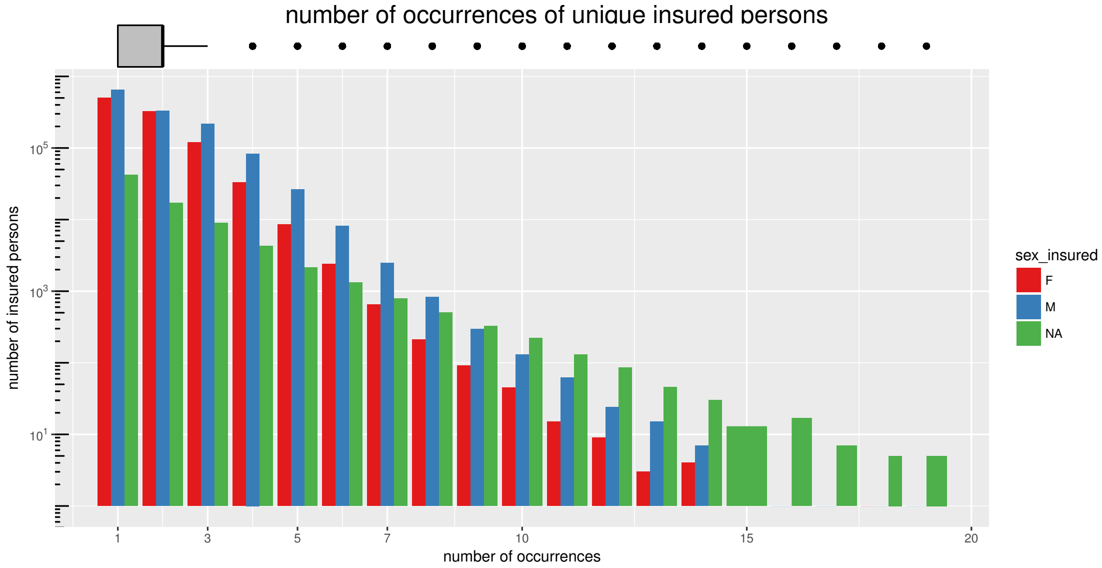


Figure 3.4: Number of relationships of unique, insured persons split by sex

the insured person and the corresponding insurance institutions. Nevertheless, insurance institutions are not included in this analysis. Summarizing, most individuals of the collected population has a rather small number of co-insurances and therefore fits the expectations.

Altogether, there are 207.496 co-insured and 77.829 insured unique persons where the variables *sex* and *sex_insured* are missing. In table 3.4 on the next page, missing values of any one or both parts of each relationship are cross-tabulated. Line by line, a combination of variables is encoded, beginning with complete cases where both genders are available. Each red cell containing the value 0 represents a missing variable. In the first two columns, the absolute and relative number of cases are listed. The section at the bottom of the table provides the sum of missing entries for each univariate variable.

Table 3.5 on the facing page extracts the most relevant information, including the total number of relationships affected. In companion to table 3.3 on page 41, it can be concluded that the missing values of the variables *sex* and *sex_insured* are only partly overlapping.

3.1.3 Age

Age and difference of age, respectively, of insured as well as co-insured persons are the most important variables for cohort selection. Therefore, the completeness of this personal information is crucial as a pair must be omitted entirely if the year of birth of at least one partner is unknown.

Altogether, there are 41.415 co-insured and 28.333 insured unique persons where the variables *age07* and *age07_insured* are missing. Table 3.6 on the facing page summarized

Table 3.4: Combinations of missing values for sex

∑	∑ %	sex	sex_insured	missing
3.917.042	89,25 %	1	1	0
302.974	6,90 %	0	1	1
116.251	2,65 %	1	0	1
52.338	1,19 %	0	0	2
∑ NA		355.312 8,1 %	168.589 3,84 %	523.901

Table 3.5: Missing values: sex, sex of insured

	complete	1 missing	both missing	sum missing
pairs affected	3.917.042	419.225	52.338	471.563
relative	89,2 %	9,55 %	1,19 %	10,8 %

missing values of any one or both parts of each relationship. In contrast to table 3.1 on page 40 it can be concluded that only about 10% of all missing values of these variables are overlapping. Nevertheless, pairs with at least one unknown age must be excluded during cohort identification.

Table 3.6: Missing values: age, age of insured

	complete	1 missing	both missing	sum missing
pairs affected	4.196.967	175.198	16.440	191.638
relative	95,6 %	3,99 %	0,37 %	4,37 %

The combination of missing year of birth and missing gender is more complex. Altogether, there are 16 combinations for all 4 variables, i.e., age in 2007 and gender of the co-insured and insured. Table 3.7 on the following page summarizes all occurring combinations with a color-coded pattern where a red cell or the content 0 signifies that a variable is missing.

While gender is completely known for 89,3% and both years of birth are available for 95,6% of all relationships, all personal information is completely present for 88,9% pairs. Gender is missing more often for both, the insured and dependent person, than age. Furthermore, the information of the dependent person is missing in more cases. Sex of the co-insured partner seems to be the most limiting factor.

Next, in figures 3.5 on page 47 and 3.6 on page 48 a tableplot is utilized to visualize the parallel univariate distribution(s) of personal information including age, sex, and death. The dataset is arranged by the age of the co-insured person in 2007 and split into

Table 3.7: Combination of missing values for sex and age

relationships	%	age07	sex	age07_insured	sex_insured	missing
3.902.969	88,93 %	1	1	1	1	0
7.268	0,17 %	0	1	1	1	1
208.509	4,75 %	1	0	1	1	1
6.763	0,15 %	1	1	0	1	1
56.483	1,29 %	1	1	1	0	1
93.668	2,13 %	0	0	1	1	2
42	0,00 %	0	1	0	1	2
334	0,01 %	1	0	0	1	2
72	0,00 %	0	1	1	0	2
29.006	0,66 %	1	0	1	0	2
59.423	1,35 %	1	1	0	0	2
463	0,01 %	0	0	0	1	3
3.593	0,08 %	0	0	1	0	3
273	0,01 %	0	1	0	0	3
4.077	0,09 %	1	0	0	0	3
15.662	0,36 %	0	0	0	0	4
$\sum$ NA		121.041 2,76 %	355.312 8,1 %	87.037 1,98 %	168.589 3,84 %	731.979

1.000 equally sized bins. Each bin is displayed as a single line with the mean and spread emphasized for numeric variables. Categories are colored according to the apparent proportions. Missing values are highlighted in green for categorical variables and as a gray bar for numeric ones.

Several conclusions can be drawn from the tableplot in figure 3.5 on the next page. Naturally, the variable *age_difference* is missing in all cases where the age of the co-insured person (variable *age07*) is not known, while the age of insured persons (variable *age07_insured*) seems to be located at its mean for the same bins. While an in-depth analysis of death is presented in the following chapter, it can already be concluded that mostly older people are recorded to be deceased.

The distributions of gender of both partners in dependence of age imply mirror-inverted ratios. While the sex of co-insured persons in the second column shows an increasing proportion of females, the opposite is true for compulsorily insured partners in column five. Furthermore, the number of persons without a documented gender mostly affects persons without a known age and younger people[3].

[3] The slight increase of dependent persons without a known gender at the lower end of the second column might be a result of high age and death or misscoding due to the year 2000 which is discovered

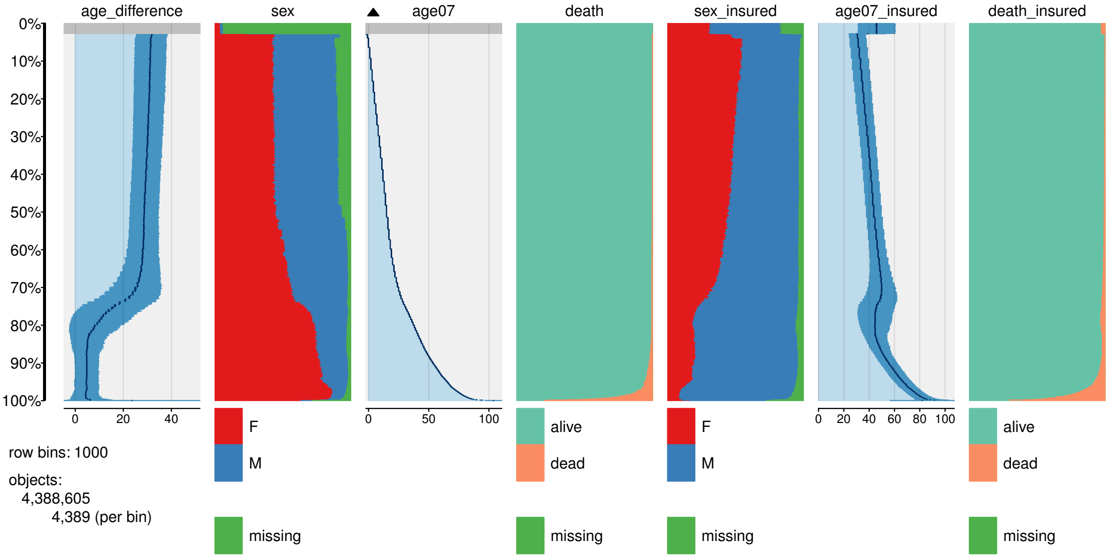


Figure 3.5: Tableplot of age, gender and death of insured and dependent, sorted by age of co-insured

Most interestingly, also the gender of the insured part of a relationship is missing more often in cases where the age of the dependent person is unknown. This correlation is confirmed by table 3.7 on the facing page where both variables are missing in about 0,45% of all relationships.

The difference in age in the first column shows a kink with a corresponding non-linear segment in the second and third last column. These structures might be related to the difference of co-insured adults and the relationship between adults and children, although no firm conclusions are possible using this visualization.

While the tableplot in figure 3.5 is sorted by the age of the insured person, arranging all bins according to the difference in age indicates new structures hidden in the dataset.

In figure 3.6 on the following page new structures can be observed, described from top to bottom. First, couples where at least one year of birth and consequently the difference in age is not known yet again show a large proportion of missing values concerning gender. Second, there is a group of adult couples with nearly equal age where predominantly females are co-insured with males. Most persons recorded to be dead are located in this group, but missing gender does not seem to be overrepresented. Third, a large section between 30% and nearly the bottom of the graph is located between about 20 and 40 years of difference in age. Most relationships between parents and children are expected to be part of this group due to the appropriate ages and difference in age as well as the balanced gender of co-insured individuals in the second column. The increase of insured

and documented later.

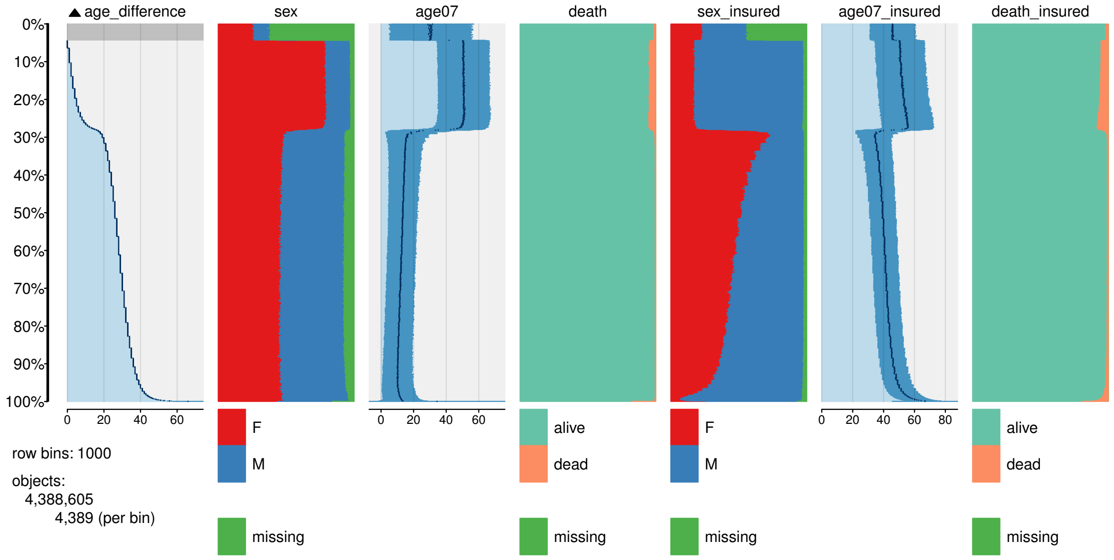


Figure 3.6: Tableplot of age, gender and death of insured and dependent, sorted by difference in age

males, analogous to the growing difference in age, is notably. Especially the peak of insured females at the top of the fifth column is surprising. It can be speculated that single or unmarried younger mothers are responsible for this structure.

Figure 3.7 on the next page shows the number of relationships per age of the co-insured persons, split by sex[4]. The graph is designed after a typical population pyramid and is cut at the age of 100 years to omit the documented outliers.

Concluding from figure 3.7 on the facing page, there seems to be a vast structural difference between children and young (until the age of about 25) adults and adults older than 25. While the number of relationships of the younger co-insured group is rather similar between the two genders, there is a large difference for older persons. It is possible that the second maximum for women (on the right-hand side), located around the age of 38 is caused by (unemployed) mothers who are supposedly co-insured with their spouses.

Figure 3.8 on the next page shows the same information for unique co-insured persons instead of the total number of relationships. The main differences between the total number of relationships in figure 3.7 on the facing page and the number of unique persons cleansed by persons in multiple relationships can be identified for persons below the age of about 20. While the crude numbers do not change much for older co-insured persons, it gets cut by nearly a third for children, most significantly for the youngest ones. It

[4] As sex of the co-insured persons is used as parameter in the graph, persons without a known gender are silently dropped. This marginal error is accepted deliberately to keep the analysis more comprehensible.

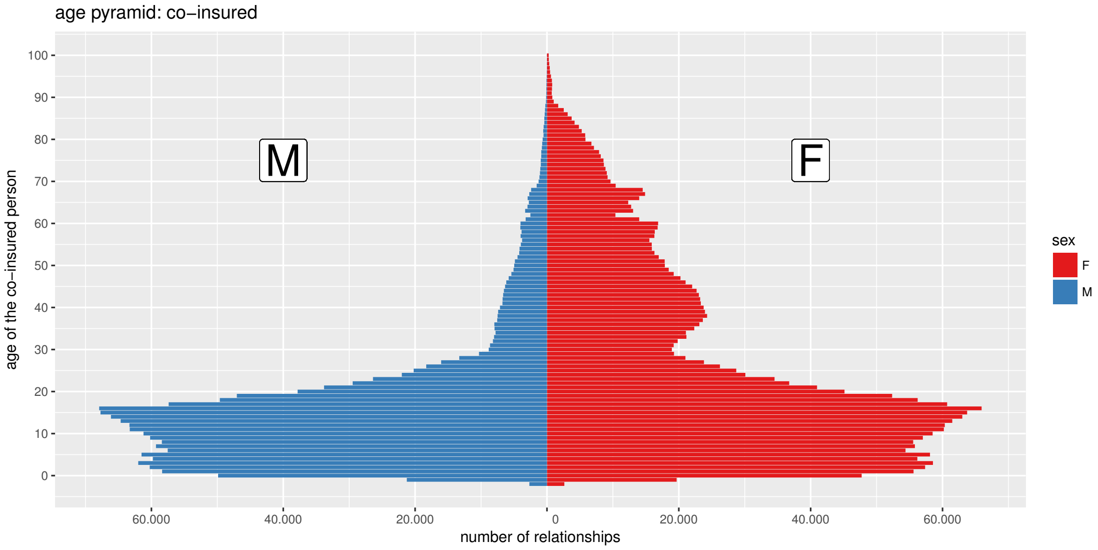


Figure 3.7: Age pyramid of relationships: co-insured

can be speculated that children tend to be co-insured with both their parents in about half the cases. Although there are still severe outliers and missing values not included in these graphs, this increases the trustworthiness of the data quality as the general trend complies with the expectations.

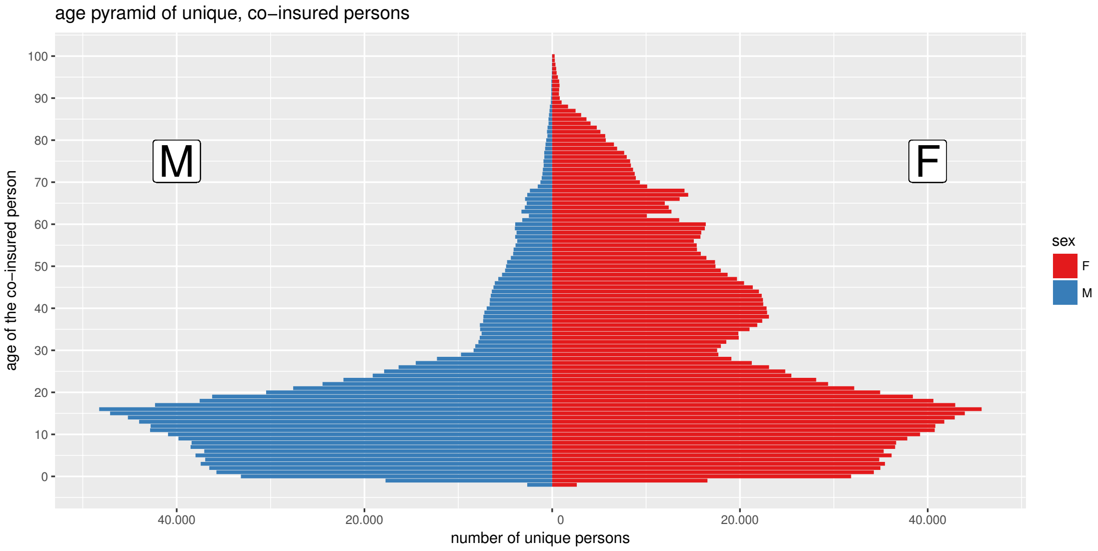


Figure 3.8: Age pyramid of unique persons: co-insured

The same analysis is also shown for the insured side of each pair. In figure 3.9 the total number of relationships and in figure 3.10 on the facing page the number of unique persons is displayed as a population pyramid.

Both graphs have one clear maximum for each gender and (despite some outliers) gain pace at the age of 20. The maximum for women is located around the age of 37 and for men around 42 years. While the age distribution of co-insured females is nearly symmetric and fades out at about the age of 60, the distribution for males has a long tail. This might be due to the fact that females are more commonly co-insured with males, but children seem to be associated with both parents.

Additionally, the difference between the sum of relationships and the number of unique persons involved varies between both genders. For females, the shape of the distribution does not change notably, and the number of cases gets roughly halved. For males, the main part of the distribution gets about halved as well, but the long tail does not change much. This might be a result of males being in a relationship are co-insured with their spouse and children during their middle ages, and later only with their partners. Therefore, the number of unique males in comparison to the number of relationships roughly halves between the age of 22 and 65, but does not decrease by the same factor for higher ages.

Ignoring the outliers, the overall picture fits the expected distribution considering stereotypical constellations in our society.

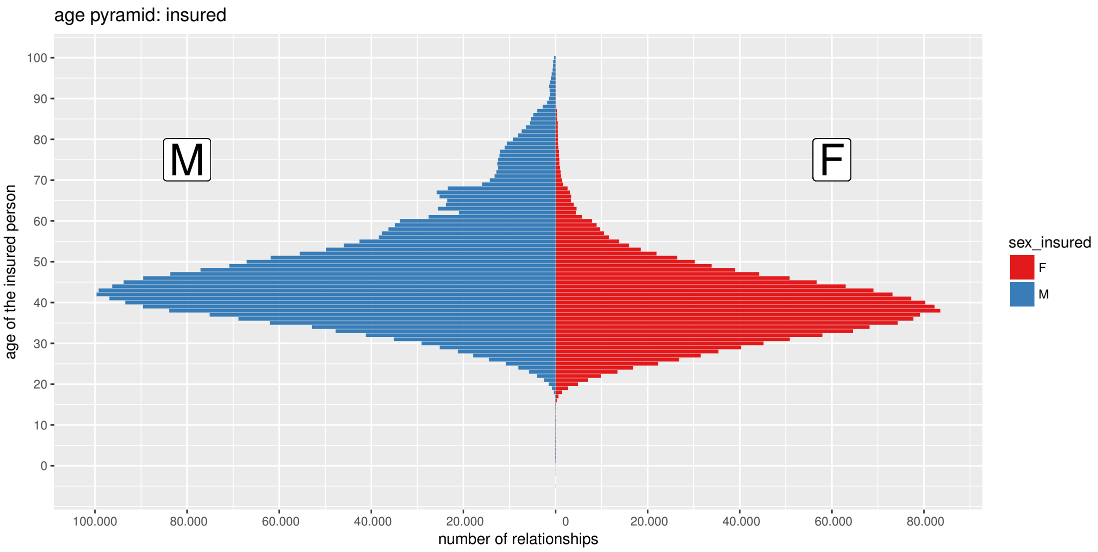


Figure 3.9: Age pyramid of relationships: insured

Furthermore, the difference in age between the insured and co-insured person is of relevance. Figure 3.11 on the next page shows the absolute difference in age for each

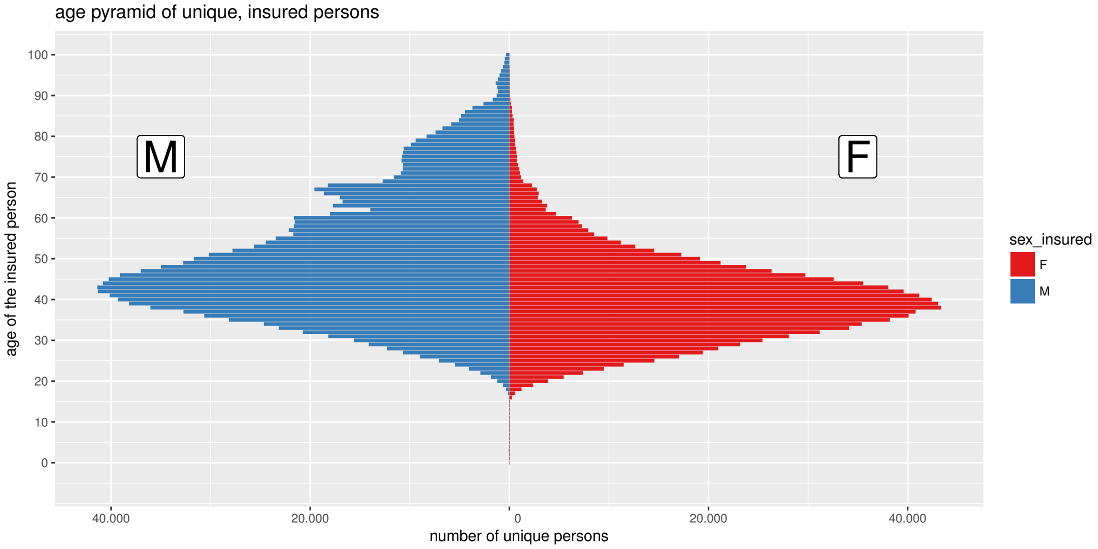


Figure 3.10: Age pyramid of unique persons: insured

relationship up to 75 years split by gender of the insured person and figure 3.12 for co-insured ones. The age difference between an adult and a child is expected to equal the age of the parent at the birth of the child.

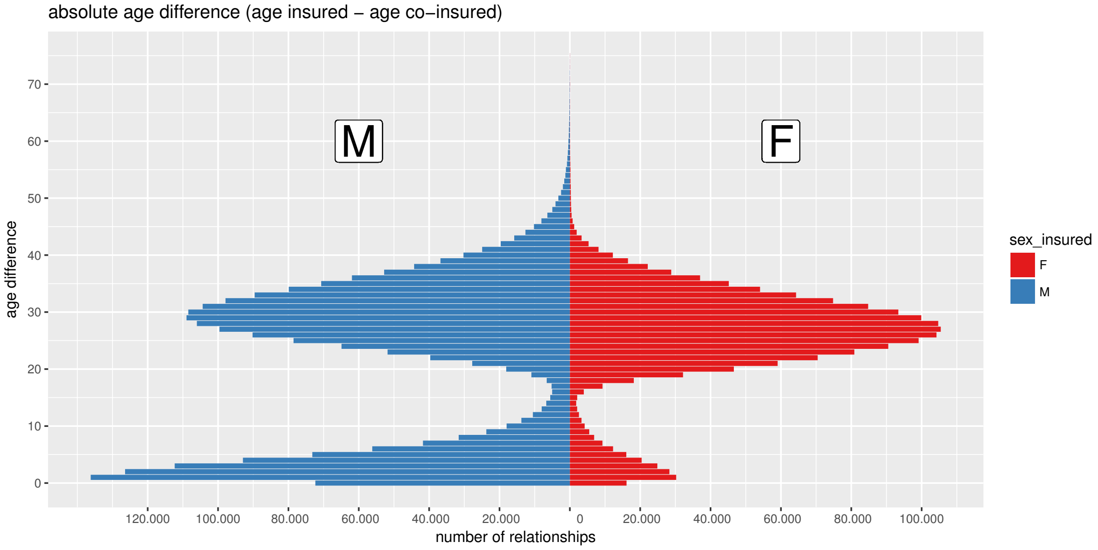


Figure 3.11: Age difference split by gender of insured

Two main groups can be spotted for each gender (of the insured person) in figure 3.11.

The upper one, with an average of about 30 years for males and 28 years for females, most likely depicts a relationship between children and parents. Males tend to be slightly older when being co-insured with children, while females are hardly older than 45 years than their children. Although the spread and maximum values seem to be higher for males, the difference between both genders is very small.

The second, lower group is located roughly between 0 and 14 years (where the bars for a difference of 14 and 15 years are the smallest). It represents co-insured adults. As already concluded before, females are less often the insured part in a relationship of adults.

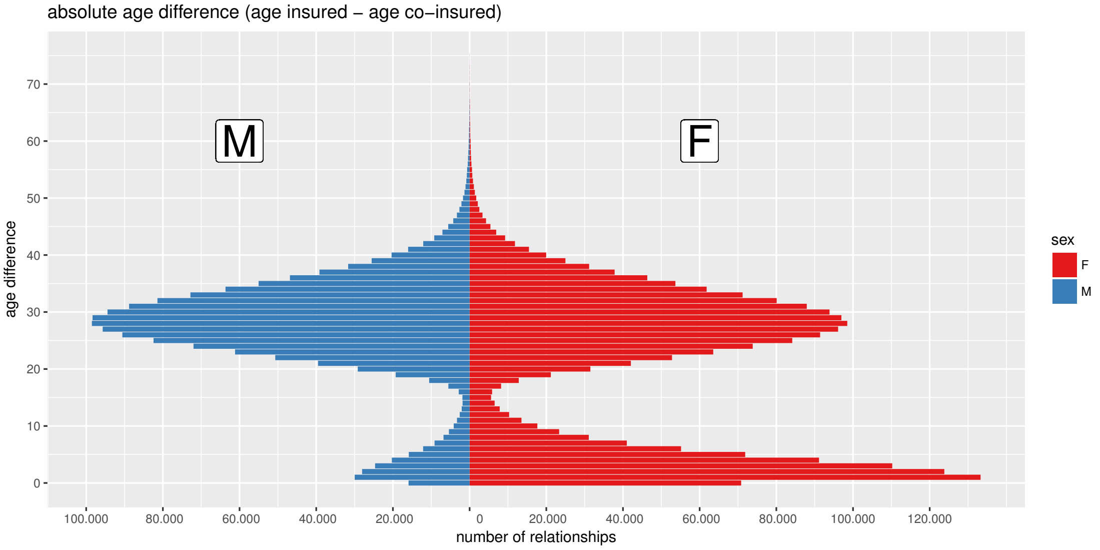


Figure 3.12: Age difference split by gender of co-insured

Moreover, figure 3.12 shows two main groups for each gender of the co-insured person. The upper group is expected[5] to mostly represent children (in contrast to their parents in figure 3.11 on the previous page) while there are co-insured couples of the same age in the lower group.

The upper structures are very symmetrical and can be located between about 15 and 60 years difference in age with a maximum at 28 years. According to Statistik Austria's "statistics of natural population movement"[6] Austrian mothers got their first child on average at the age of 28 in the year 2007, which would correspond with the data from GAP-DRG.

The lower group is located between 0 and 14 years difference with a maximum of one year. It includes co-insured couples of a similar age. Again, the distribution of males and

[5] It is also possible that co-insured spouses have an age difference of more than 20 years although this is not the most likely case in our society.

[6] Durchschnittliches Gebär- bzw. Fertilitätsalter der Mutter nach Lebendgeburtenfolge seit 1984, last accessed 2016-01-30

females are symmetrical in shape, but there are more relationships where women are the dependent partner.

Figure 3.13 and figure 3.14 on the next page show the number of co-insured and insured persons born per year, faceted by sex. Additionally, the same information is presented for persons with unknown gender and a smoothed[7] curve is drawn to emphasize the general trend. In figure 3.14 on the following page the scale of the y-axis is adjusted according to the maximum values occurring.

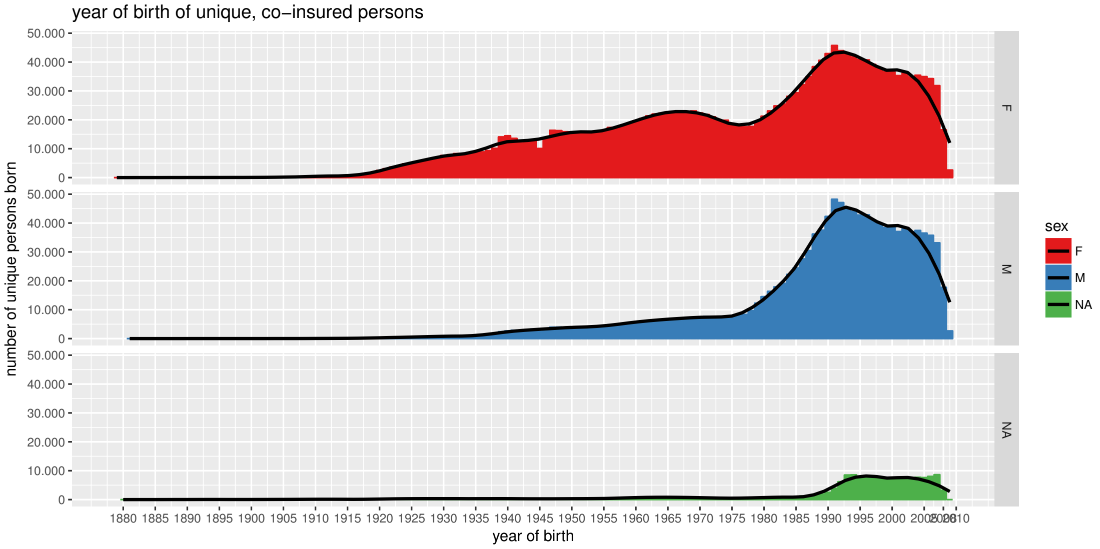


Figure 3.13: Year of birth of unique, co-insured persons split by gender

Because age is always calculated for the year 2007, the distribution of age equals the distribution of persons per year of birth. The shape of the curve for persons without a known gender seems to be slightly more like the curve for males than females. Altogether, these special cases are rare but follow the general trend and are therefore not suspected to be caused by outliers concerning the year of birth.

A gap can be spotted in the year 1945 for females in figure 3.13 and males in figure 3.14 on the following page. Furthermore, the numbers of people born in the years 1940 and 1947 seem to be higher than the local trend would suggest. These observations can be explained by historical incidents and fit the expectations well.

Thus far, age has been analyzed independently from the structure defined by the relationships[8]. Nevertheless, specific differences in age are defined in the study protocol to distinguish couples, parents and children from unwanted information and data errors.

[7] The smoothed line is calculated using *LOESS*, "Locally Weighted Scatterplot Smoothing" with a rather small span of 0,15

[8] The conjunction of missing values implied by relationships is summarized in table 3.6 on page 45.

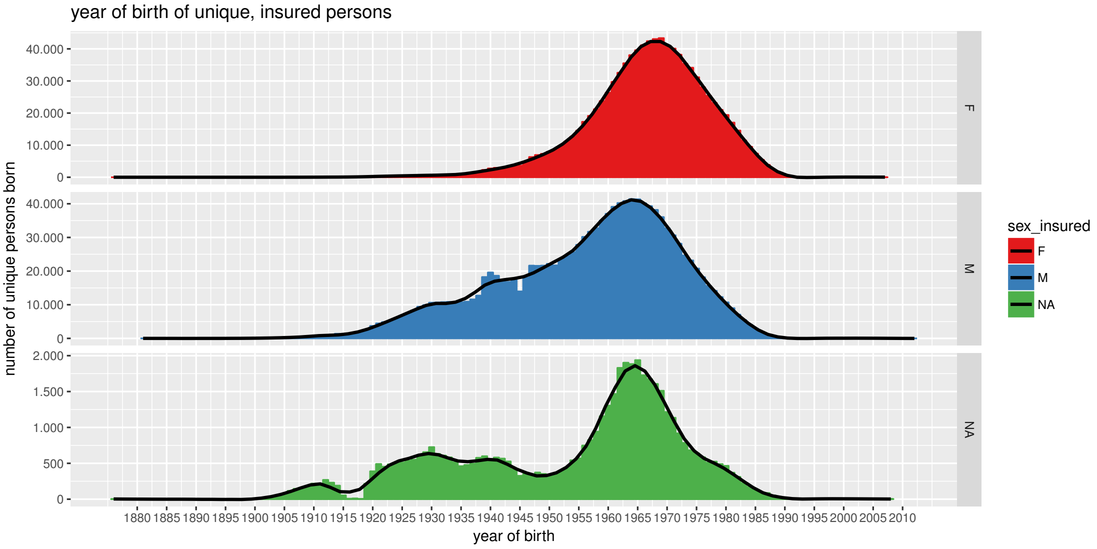


Figure 3.14: Year of birth of unique, insured persons split by gender

Therefore, the specific interaction of the age of insured and co-insured persons is relevant. As for the analysis of differences in age, only cases where the year of birth is known for both partners can be included.

As described in chapter 2 on page 9, a two-dimensional density is calculated and displayed as 3-dimensional shape and contour plot[9] in figure 3.15 on the next page. The 3-dimensional visualization is rotated and tilted to give an optimal view of the two salient shapes. Therefore, the axes of both plots are not aligned equally.

While the density estimation in figure 3.15 on the facing page is appropriate for visualization, the calculated values themselves are hard to interpret. Therefore, hexagonal binning is henceforth utilized to visualize various *age-age* matrices.

Figure 3.16 on the next page visualizes the entire dataset. Roughly the same structure as in figure 3.15 on the facing page can be spotted. In addition to that, more information on the surrounding noise is available. In both graphics, two main clusters can be identified.

The larger, higher, and rounder one on the bottom of the graph involves children and young adults co-insured with an older adult. The base of this cluster is located between the age of 0 and roughly the end of the 20's for the dependent and between 20 and mid-50's or 60 years for the insured person. It shows a clear linear relationship between the co-insured and insured person, although there is a spread of about 20 years.

The second cluster appears not as significant, holds fewer cases, and is structured

[9] The baseline or background of the contour plot is manually colored in *white* to emphasize the apparent structure.

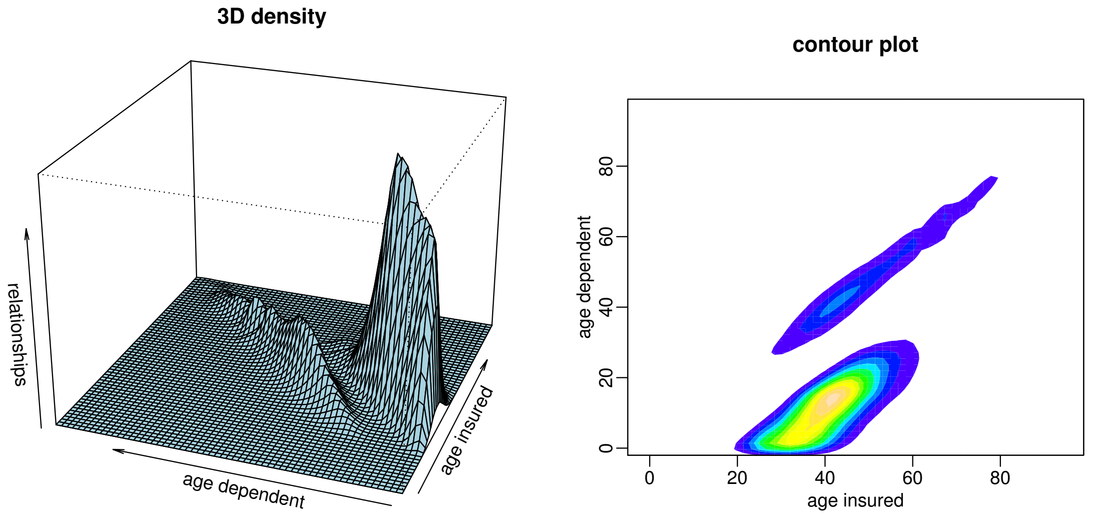


Figure 3.15: Age of insured and dependent

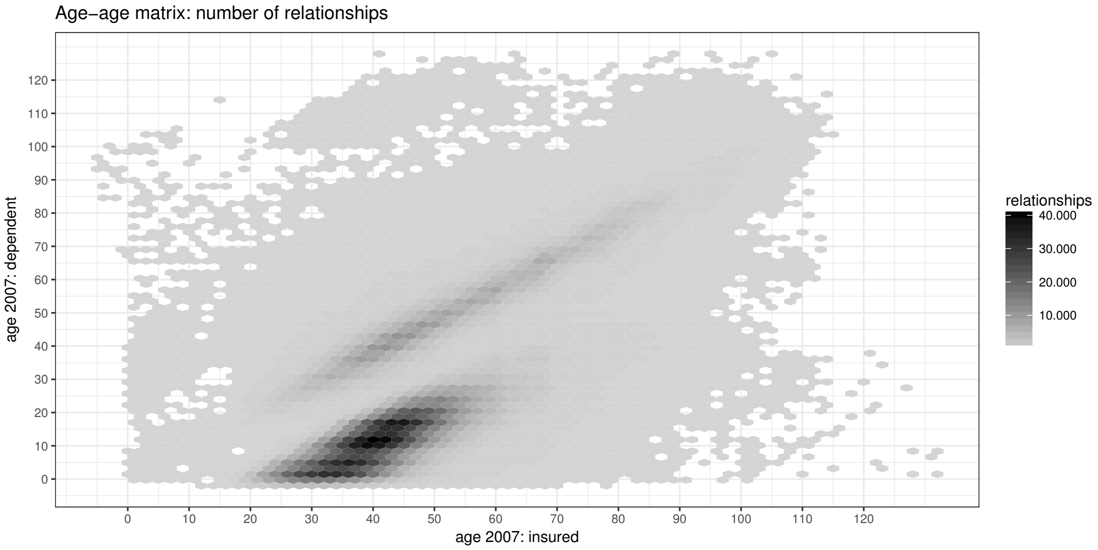


Figure 3.16: Number of relationships per age of insured and dependent person

differently. For both the insured and dependent part of a relationship, it is located roughly between the age of 30 and 80 with a maximum in the 40's, while the compulsorily insured person seems to be older. A clearer linear relationship for the first cluster with a smaller spread can be observed. This cluster most likely holds co-insured (adult) partners.

Furthermore, there seems to be background noise nearly all over the entire graph. Although some of these cases might actually occur, some seem to be very unrealistic and can be classified as errors. Unfortunately, it is not possible at this stage to select both clusters for further analysis because this would affect the classification of parents and spouses without children massively. Additionally, this data only holds direct co-insurance and does not consider mediated relationships (e.g., a father being associated with his child only because both are co-insured with their mother/wife).

In figure 3.17 the same information as in figure 3.16 on the preceding page is faceted by the sex of both the insured and dependent person. To consider as much data as possible, unknown sex is included in separate graphs and age is shown with the full range.

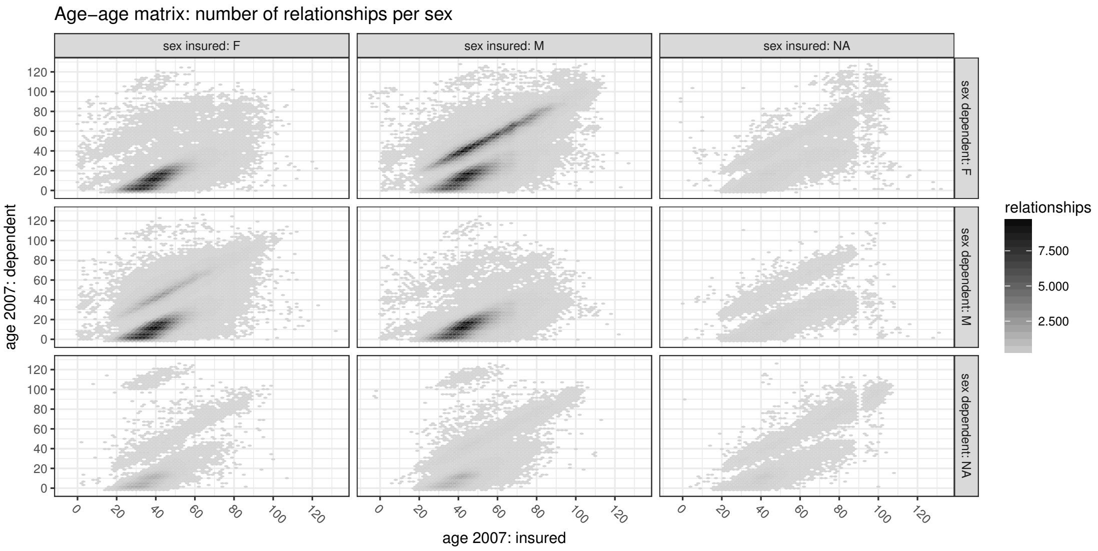


Figure 3.17: Number of relationships per age and sex of insured and dependent person

At first sight, figure 3.17 is very similar to previous images. Nevertheless, there are several important differences to spot. Most of co-insured adults are of different sexes and can therefore be interpreted as heterosexual couples. In most of these cases, a female is co-insured with a male partner. The opposite direction is not only rarer but also has a smaller spread. This difference cannot be observed with the same intensity for children being co-insured with an adult. Furthermore, the sex of dependent children is more often unknown than for other groups, what might be a result of their larger number of observations. While the clusters for persons with unknown gender have shapes like the general trend, extreme outliers seem to be overrepresented, e.g., co-insured persons above the age of 100.

A log-scale is utilized in figure 3.18 on the facing page. Due to a software error[10] in

[10] Details are documented in issue number 1608of the project's official bug tracker.

ggplot2 [Wickham, 2009b], 2d binning must be applied instead of a hexagonal grid.

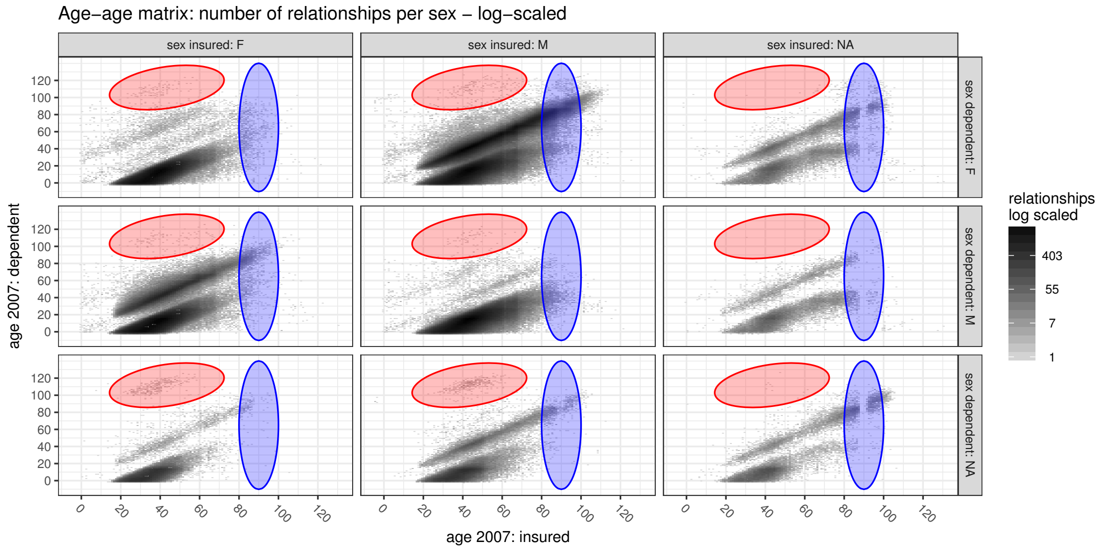


Figure 3.18: Number of relationships per age and sex of insured and dependent person - log-scaled, 2D bins

Previously identified clusters appear emphasized in figure 3.18, although the marginal areas of the apparent structures are larger and are harder to discriminate. There is a rather strange blank *line* where fewer individuals than in the surrounding areas are occurring, encircled by a blue ellipse. This structure appears to be most outstanding for insured persons without a documented gender at the age of around 90. This is most likely an artifact of an unknown data cleansing process. Furthermore, there is another vague but distinguishable cluster centered around the insured person's age of 40 and the dependent person's age of 110, marked by a red ellipse. It is also most prominent where the co-insured person's gender is missing. This structure might be a result of a wrongful interpretation of a two-digit year of birth (i.e., the year 2000 error). In case the exact year of birth of a co-insured child is stored only in a short, 2-digit form or is retrieved from the social insurance number, the wrong century might be preceded.

In figure 3.19 on the following page the data is restricted to complete information considering age and gender and is limited to a maximum age of 100 years.

It can be observed that co-insured adult couples have partners of rather similar age, where males tend to be the insured part and are slightly older. There are more relationships documented between about the age of 30 and 60 with a maximum around 40 years, which hints to an overrepresentation of parents in comparison to couples without children. These variances fit the expectations well. Nevertheless, some of these structures might be influenced by the source of information. Furthermore, there are hardly any co-insured same-sex couples occurring more often than the overall background noise.

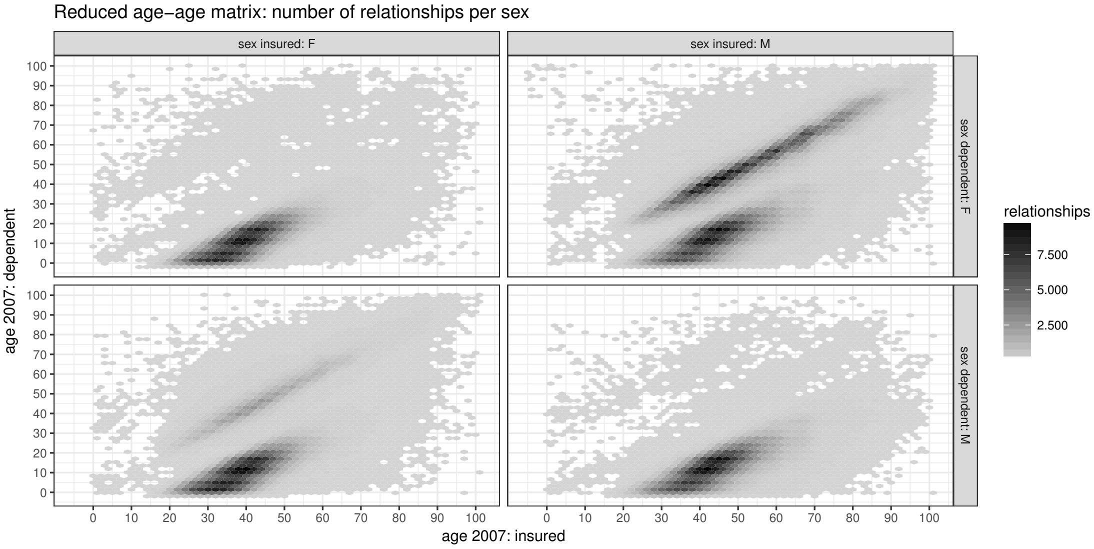


Figure 3.19: Number of relationships per age and sex of insured and dependent person cleaned data

Summarizing, the overall distribution of age and difference in age seems very promising for the task at hand. About 5% of all relationships are not usable because of missing values. Additionally, there are outliers and implausible values present where, e.g., one or both insured persons are underage, persons are over 100 years old or the difference in years is above 70. As some of these cases might be real outliers and others just data errors, the grand picture fits the expectations.

The analysis of the age for both the insured and dependent person reveals interesting structures and confirms the data quality, although much background fluctuation is present. Overall, it complies with general knowledge or stereotypes about the Austrian society.

3.1.4 Death

Death, either as a state or event, are not part of the cohort selection. It is also not explicitly defined how to handle deceased people.

There are several possibilities to include, exclude, or partially include this corner case (e.g., should they be excluded altogether including their children and spouses, should they be ignored in the following analysis, should they participate according to their time at risk). Especially regarding the focused diagnoses which might cause immediate death, total exclusion could also remove important cases. On the other hand, observing persons who have already died potentially distorts the results too.

Crude numbers of relationships concerning death until the year 2007 (i.e., before 2008) are summarized in the following contingency tables. First, table 3.8 gives the total number of relationships where a death on either side occurred. Row percentages can be found in table 3.9, column percentages in table 3.10 and the proportion of each element in relation to the total number of relationships in table 3.11 on the next page.

Table 3.8: Cross table of death for relationships between dependent (row) and insured (col)

	FALSE (insured)	TRUE (insured)	Sum
FALSE (co-insured)	4.289.490	52.756	4.342.246
TRUE (co-insured)	33.263	13.096	46.359
Sum	4.322.753	65.852	4.388.605

Table 3.9: Cross table death: row percentages

	FALSE (insured)	TRUE (insured)	Sum
FALSE (co-insured)	98,79	1,21	100,00
TRUE (co-insured)	71,75	28,25	100,00
Sum	98,50	1,50	100,00

Table 3.10: Cross table death: column percentages

	FALSE (insured)	TRUE (insured)	Sum
FALSE (co-insured)	99,23	80,11	98,94
TRUE (co-insured)	0,77	19,89	1,06
Sum	100,00	100,00	100,00

Less than 2,5% of all relationships are affected by the death of at least one partner. About 1% of co-insured and 1.5% of insured cases are involved. The difference might be caused by the ratio of the total number of insured (2.386.052) and co-insured (3.328.267) persons. The number of wrongfully missing values can hardly be determined in GAP-DRG because there is no direct indicator for the state of living.

Table 3.12 on the following page summarizes these findings.

Altogether, 45.297 (1,36% of all) unique co-insured persons and 54.258 (2,27% of all) unique insured persons have died before the year 2008. Both proportions are larger than the relative number of affected relationships.

Table 3.11: Cross table death: total percentages

	FALSE (insured)	TRUE (insured)	Sum
FALSE (co-insured)	97,74	1,20	98,94
TRUE (co-insured)	0,76	0,30	1,06
Sum	98,50	1,50	100,00

Table 3.12: Relationships affected by death

	alive	1 dead	both dead	sum death
pairs affected	4.289.490	86.019	13.096	99.115
relative	97,7 %	1,96 %	0,3 %	2,26 %

It can be assumed that persons who have already been recorded as dead are not in an active relationship or consume any goods and services from the healthcare system. Nevertheless, this is not the case for administrative claims data in the presence of data quality issues.

Furthermore, the fact that a person has already died affects the completeness of other variables. In table 3.13 and table 3.14 the distributions of sex of unique persons who have died before 2008 are summarized for co-insured and insured persons.

Table 3.13: Sex of co-insured persons who have died before 2008

	F	M	NA	Sum
persons	30.852	10.165	4.280	45.297
relative	68,1 %	22,4 %	9,45 %	100 %

Table 3.14: Sex of insured persons who have died before 2008

	F	M	NA	Sum
persons	3.966	44.555	5.737	54.258
relative	7,31 %	82,1 %	10,6 %	100 %

It can be concluded that the gender is unknown for about 11% of all persons who have died before 2008. Considering the crude missing values presented in table 3.3 on page 41, especially the difference for insured persons is notable. This may also be related to the exact date of death which is presented later.

Interestingly, variables about age are not missing for deceased persons at all. In detail, the year of birth is missing for exactly 0 co-insured and 0 insured unique persons who have died before 2008.

The age distribution in 2007 of co-insured persons (i.e., not the age when they have died but the age considered in the analysis) split by gender is plotted in figure 3.20. As for the other age pyramids, persons without a known sex or year of birth are omitted and the maximum age is truncated at 125 to omit outliers.

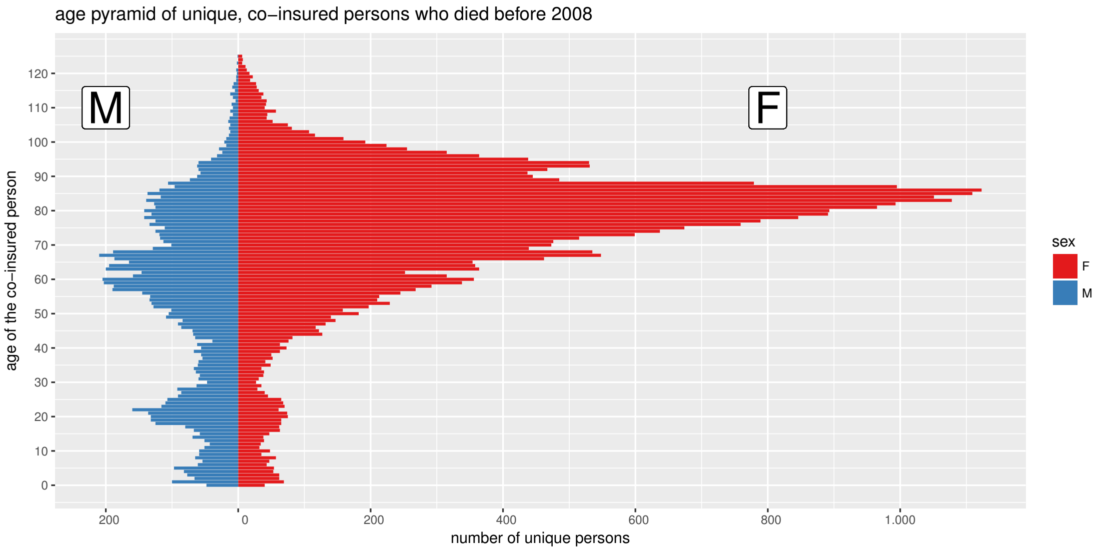


Figure 3.20: Age pyramid of unique, co-insured persons who died before 2008

Figure 3.20 gives a mixed impression for both genders. There is approximately the same number of deceased children and adults between the age of 20 and 50. As these absolute values are on the one hand rather small and on the other hand more meaningful in relation to the total number of persons per group as shown in figure 3.8 on page 49, another plot with relative values is presented in figure 3.21 on the next page.

Figure 3.21 on the following page gives a more precise and mostly expected picture of the age distribution of dead, co-insured persons. Although there is a very small proportion of young children and adults who have died, most cases can be found above the age of about 60. Additionally, there is a recorded date of death for most cases of a small number of people with uncommon or even unrealistic ages over 100 years. It can be assumed that the remaining cases are a direct result of the year-2000 error documented in figure 3.18 on page 57.

In figure 3.22 on the following page the relative number of insured persons who have died before the year 2008 is plotted per age and gender. As before, persons without a known gender and age above 125 years are omitted.

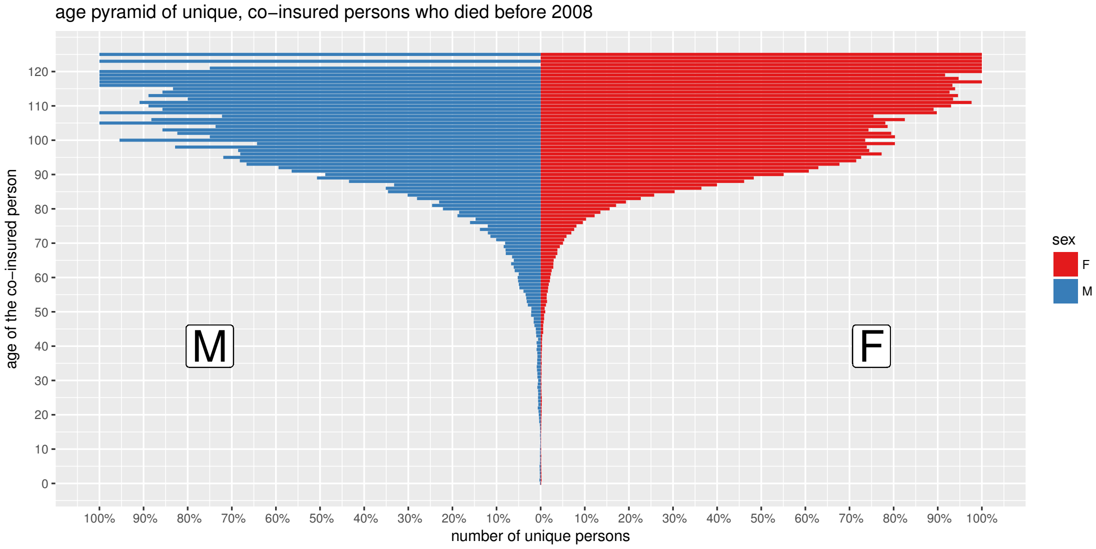


Figure 3.21: Age pyramid of unique, co-insured persons who died before 2008 relative to the total number of persons in each group

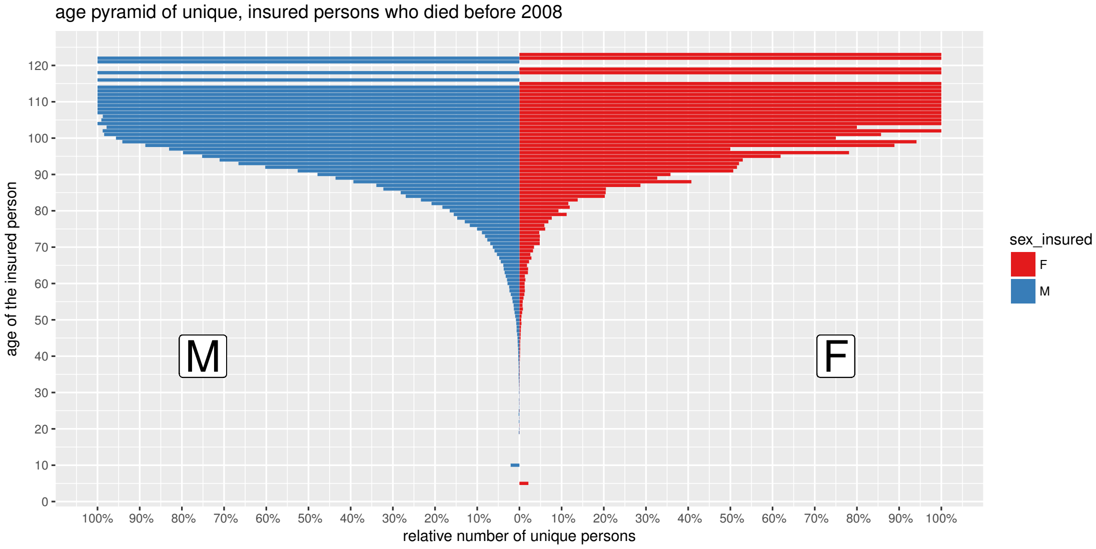


Figure 3.22: Age pyramid of unique, insured persons who died before 2008 relative to the total number of persons in each group

The result is mostly similar shaped as in figure 3.21. Young adults and children seem to be even rarer, which complies with prior findings (e.g., in figure 3.10 on page 51). Most interestingly, there are salient groups at the age of 4 and 10 for males and 5 and 8 for females, respectively, where a higher ratio of insured persons has died. Considering the very small number of children being compulsorily insured (and having a co-insured partner), the unclear implications and practical possibilities in the Austrian social insurance system for such cases, these are most likely errors or artifacts of the data generating process.

Next, the distribution of the date of death is examined. The number of co-insured individuals who have died per distinct day faceted by sex is presented in figure 3.23.

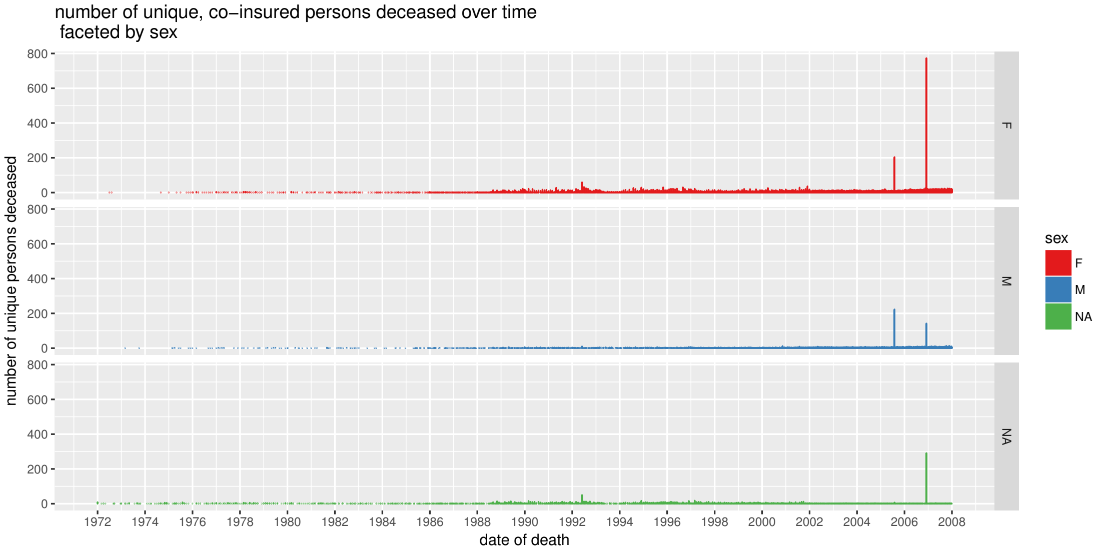


Figure 3.23: Number of unique, co-insured persons deceased over time faceted by sex

Extreme outliers can be spotted per gender in figure 3.23, distorting the overall picture. They are most likely a result of some cleanup procedures in the original data source. These outliers are extracted in table 3.15.

Table 3.15: Number of deceased co-insured persons per day: outlying values

death date	F	M	NA
2006-12-04	773	142	291
2005-07-31	204	223	7
1992-05-31	59	10	50

Removing these outliers, figure 3.24 on the following page shows a more comprehensible picture. The y-axis is scaled accordingly for each gender to emphasize the available information. The black line represents smoothing[11] to show the overall trend hidden by the present outliers.

[11] *LOESS*, "Locally Weighted Scatterplot Smoothing" with a rather small span of 0,15

It can be concluded that there are co-insured persons in the dataset who are recorded to be dead for several years. The number of deceased persons per date is increasing, although there are several outliers, especially for females. Altogether, there are 31.571 co-insured persons who are reported to have died before 2006 and 13.726 co-insured persons who have died in 2006 or 2007.

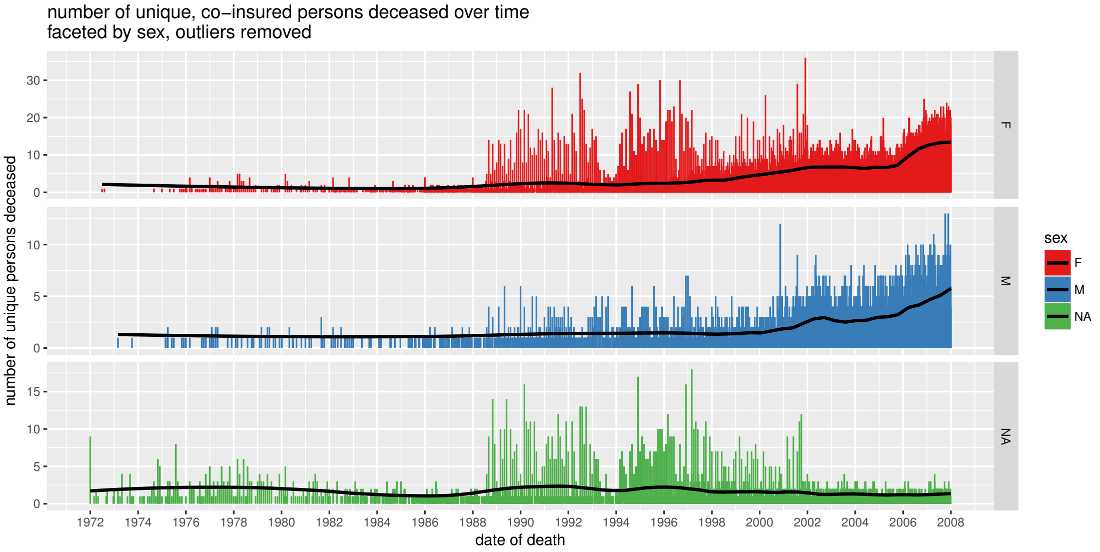


Figure 3.24: Number of unique, co-insured persons deceased over time faceted by sex without outliers

The two most extreme outliers for compulsorily insured persons are extracted in table 3.16. According to the proportion of genders of insured persons, the most extreme values are recorded for males only.

Table 3.16: Number of deceased insured persons per day: outlying values

death date	F	M	NA
2006-12-04	37	169	89
2005-07-31	14	102	11

Finally, all unique, compulsorily insured persons who have deceased before 2008 are plotted in figure 3.25 on the facing page, but without previously documented outliers. Most interestingly, larger fluctuations in the late 1990's and a volatile increase at the beginning of 2006 (where the data collection starts) can be observed.

Altogether, there are 30.225 co-insured persons who are reported to have died before 2006 and 24.033 co-insured persons who have died in 2006 or 2007.

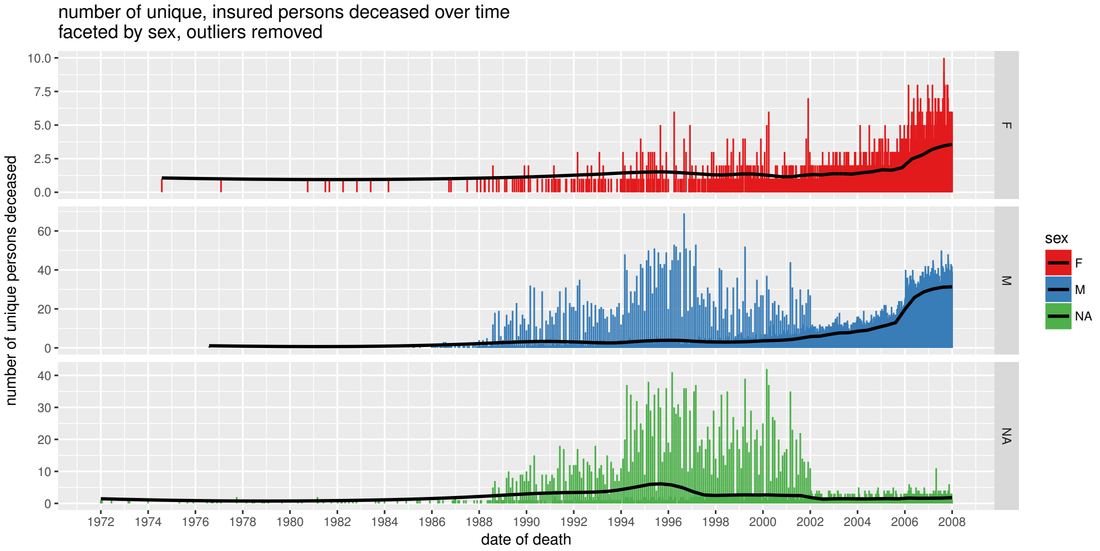


Figure 3.25: Number of unique, insured persons deceased over time faceted by sex without outliers

Summarizing, the overall number of deceased persons is rather small in comparison with the extracted cohort. Overall, the age and gender distribution of dead persons is plausible and corresponds to the entire cohort. Outliers concerning the variable age are mostly reported as dead, which is a promising hint to reliable data. There also seems to be a slight correlation between missing values for sex and the fact that a person is recorded as deceased.

Most interestingly, there are several thousand deceased persons who are still in the dataset, even if their date of death is before the year 2006, the beginning of the data collection. Furthermore, the day of death of these persons does not seem to be evenly distributed but shows several outliers mainly in the 1990's. Extreme outliers at 4th December 2006 and 31st July 2005 might be the result of data cleansing processes.

The source and reason of these structures and the presence of deceased persons are unclear. Nevertheless, it can be speculated that especially the large variations in the 1990s', where rather large peaks can be observed, are a result of the data recording process. The so-called *Zentrale Melderegister* (ZMR), a central, digitalized register of all Austrian citizens/inhabitants, has been established in 2001 and finally activated in 2002.[12] Social security institutions used this cleaned information since then and might

[12] More details can be found in the press release from the Ministry of Internal affairs http://www.ots.at/presseaussendung/OTS_{}20011205_{}OTS0157/(last visited 2020-05-08) and in a publication series of the *Austrian community association* (*Österreichischer Gemeindebund*), issue 2-2001, March 2001, Vienna: `http://gemeindebund.at/images/uploads/downloads/2014/Publikationen/RFGs/2001/RFG_2-2001_-_Zentrales_Melderegister.pdf`.

have also implemented data quality procedures. Notably, even persons who have died many years ago are reported to be co-insured or have consumed services of the healthcare system.

As the study protocol suggests, all hints for co-insurance will be utilized in the subsequent steps of the cohort selection. In the final analysis, persons who are already deceased at the beginning of 2006 will be removed. No additional precautionary measures will be implemented for the small number of persons who have deceased in the years 2006 and 2007 (5.371 or 0,16% of co-insured and 10.547 or 0,44% of insured)[13].

3.1.5 Standardized research population

A standardized *research population* has been introduced in GAP-DRG to enhance overall data quality and reliability in research projects. Moreover, the profiling gave the impression that data quality (e.g., completeness of personal information) and the number of co-insurances per person might correlate with a person belonging to this selected cohort.

Figure 3.26 on the next page shows the number of dependent persons (log-10 scaled ordinate) with a specific number of co-insurances (i.e., the occurrence of a person's ID as a dependent part of a relationship). There is one bar for each group of persons being part of the research population or not. The chart is truncated at 50 co-insurances per person to omit outliers. Additionally, if there is one group (e.g., *TRUE*, meaning someone belongs to the research population) missing completely, the other bar gets plotted with the full (doubled) width.

It can be deduced that dependent persons who are also part of the research population have most likely less than 10, and in most cases less than 5 partners where hints for a relationship exist. On the other hand, dependent persons who are not part of the research population compose the dominating group of people with 5 or more co-insurances. This sheds a good light on the data quality as it can be expected that in most typical relationships a person is co-insured with only a few insured ones. This is the case for nearly the entire subset of dependent persons associated with the research population.

Next, figure 3.27 on the facing page displays the same information for the insured population.

Most interestingly, the association is mirrored in comparison with the same plot of dependent persons in figure 3.26 on the next page.[14]

The explanation for this pattern is most likely hidden in the assembly of the insured and dependent persons. As observed in the univariate analysis, significantly more persons being co-insured are not part of the research population than the insured ones (roughly 11% in comparison to about 7%). The same group tends to have more co-insured partners

[13] These are the absolute and relative numbers of unique persons having any relationship, who are between 30 and 60 years old in the year 2007 and have died in 2006 or 2007.

[14] As this is not the initially expected result, the entire source of the data had to be checked thoroughly. This did not reveal any severe problems with the data source or database queries.

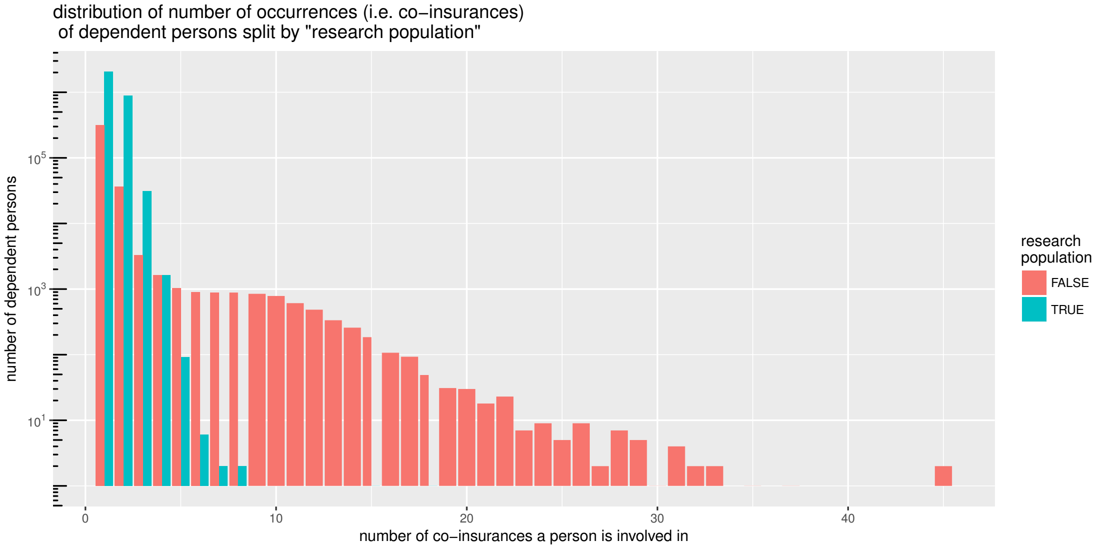


Figure 3.26: Distribution of number of occurrences (i.e., co-insurances) of dependent persons split by "research population"

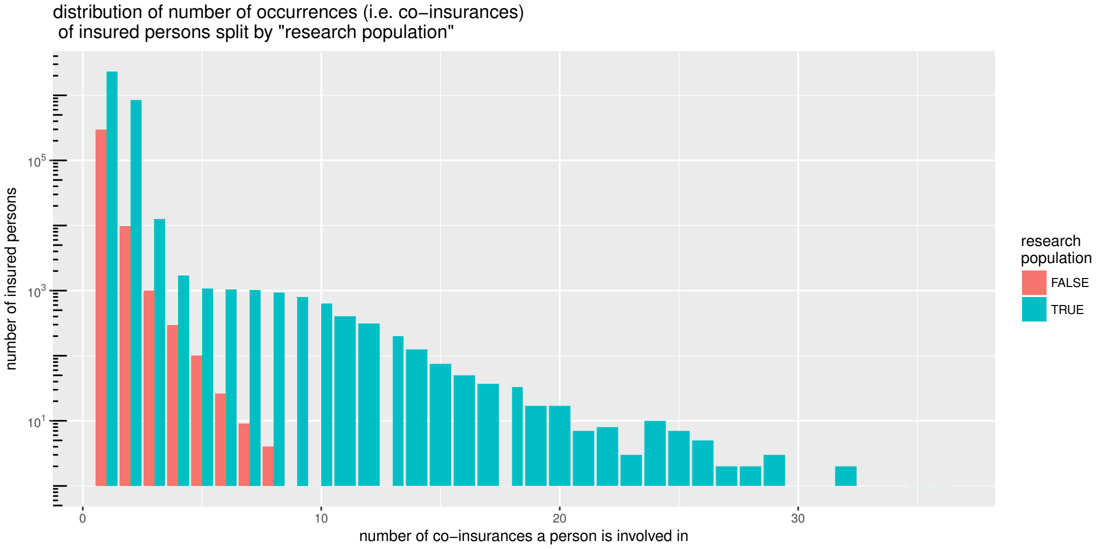


Figure 3.27: Distribution of number of occurrences (i.e., co-insurances) of insured persons split by "research population"

documented. Furthermore, there are 3.328.267 dependent persons and only 2.386.052 insured ones (of a total 5.240.670 persons and 4.388.605 pairs). Furthermore, claims data from the Austrian health insurance system does only record co-insurances in case a service is reimbursed by the dependent partner of a couple. Therefore, it is more likely that a co-insured person gets (also wrongly) associated with an insured person than vice versa.

As a result, there must be a larger group of insured persons associated with co-insured ones which are not originating from the research population. As there is a larger share of insured persons belonging to the research population, they are also more likely to be associated multiple times with dependent persons.

This liaison between insured and co-insured couples in relation to the corresponding research population is described in the following contingency tables. First, in table 3.17 the total number of relationships is summarized. Row percentages can be found in table 3.18, column percentages in table 3.19 on the facing page and the proportion of each element in relation to the total number of relationships in table 3.20 on the next page.

Table 3.17: Cross table of membership of the research population for relationships between dependent (row) and insured (col)

	FALSE (insured)	TRUE (insured)	Sum
FALSE (co-insured)	104.455	372.658	477.113
TRUE (co-insured)	216.453	3.695.039	3.911.492
Sum	320.908	4.067.697	4.388.605

Table 3.18: Cross table research population: row percentages

	FALSE (insured)	TRUE (insured)	Sum
FALSE (co-insured)	21,89	78,11	100,00
TRUE (co-insured)	5,53	94,47	100,00
Sum	7,31	92,69	100,00

Summarizing, multiple occurrences per person in conjunction with the standardized research population of GAP-DRG gives a complex picture. Introducing additional variables such as sex, age, and date of death complicates the analysis even more. No further structures of interest could be found with basic exploratory data analysis, even though there seems to be a trend that the existence of missing values correlates with the number of occurrences. Furthermore, there is a rather small total number of persons, where multiple co-insurance can be found, but all other relevant information is complete, and the person belongs to the research population.

Table 3.19: Cross table research population: column percentages

	FALSE (insured)	TRUE (insured)	Sum
FALSE (co-insured)	32,55	9,16	10,87
TRUE (co-insured)	67,45	90,84	89,13
Sum	100,00	100,00	100,00

Table 3.20: Cross table research population: total percentages

	FALSE (insured)	TRUE (insured)	Sum
FALSE (co-insured)	2,38	8,49	10,87
TRUE (co-insured)	4,93	84,20	89,13
Sum	7,31	92,69	100,00

As a result, all relationships are utilized for cohort extraction as defined by the study protocol, independently of one or both participants belonging to the research population. The final cohort has to be limited to GAP-DRG's research population according to the best practice recommendations.

3.1.6 Hints for co-insurance

Finalizing the exploratory data analysis and quality assessment of the data on co-insurance, different types of hints to a relationship are described. All four sources

- metadata
- prescriptions
- inpatient episodes
- ambulatory outpatient contacts

are expected to represent a different point of view. Especially, the comparison of co-insurances defined in the personal metadata of the social security institutions with relationships documented in the claims data is of interest.

In figure 3.28 on the following page, the Pearson correlation between variables related to hints and age as well as the difference in age of the involved persons is plotted for the entire dataset.

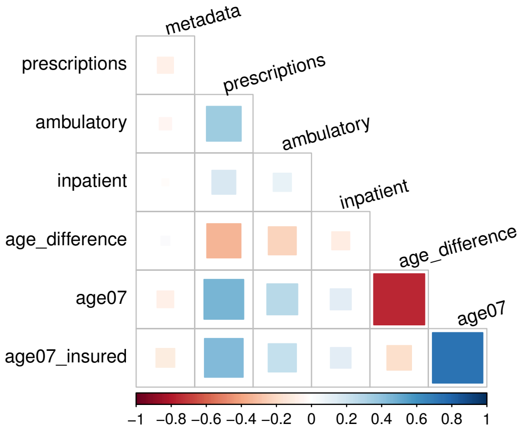


Figure 3.28: Pearson correlation of hints for co-insurance and age

The plot shows the number of prescriptions correlating with age and the number of ambulatory contacts. The latter one has also a slightly positive correlation with age, while hints extracted from the metadata have none or a vague negative correlation with all other variables. It can be concluded that older persons have more recorded reimbursements. Naturally, the age of insured and dependent persons is highly correlated.

The same graph for Spearman's ρ rank correlation[15] (figure 3.29) is calculated as there are many known outliers and possibly non-linear but monotonic relationships in the dataset.

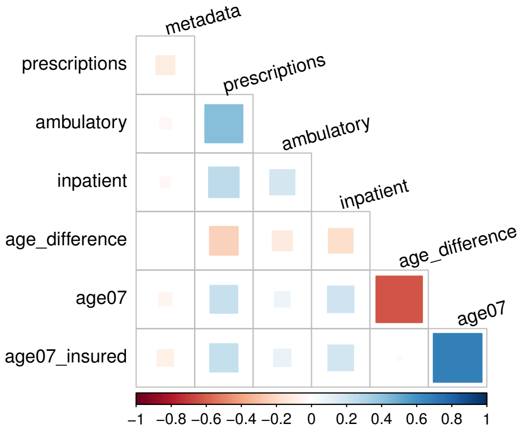


Figure 3.29: Spearman's rank correlation of hints for co-insurance and age

Again, a strong positive correlation between the age of insured and dependent persons as well as between hints from ambulatory contacts and prescription data can be observed,

[15] Kendall's τ required much longer to be calculated for the given dataset and has therefore to be omitted.

while the correlation between prescriptions and age is mitigated significantly. The absence of correlation for the variable metadata can be explained by its low variability[16].

Building upon table 3.1 on page 40 where the univariate information of each numeric variable are summarized, table 3.21 shows the patterns of missing data for all 4 sources of hints to a relationship.

Table 3.21: Combination of missing values for all 4 sources of hints

relationships	%	metadata	ambulatory	prescriptions	inpatient	missing
54.171	1,23 %	1	1	1	1	0
13.381	0,30 %	0	1	1	1	1
67.779	1,54 %	1	1	0	1	1
1.229	0,03 %	1	0	1	1	1
490.359	11,17 %	1	1	1	0	1
22.558	0,51 %	0	1	0	1	2
360	0,01 %	0	0	1	1	2
10.313	0,23 %	1	0	0	1	2
126.849	2,89 %	0	1	1	0	2
1.059.170	24,13 %	1	1	0	0	2
53.518	1,22 %	1	0	1	0	2
20.588	0,47 %	0	0	0	1	3
658.851	15,01 %	0	1	0	0	3
113.368	2,58 %	0	0	1	0	3
1.696.111	38,65 %	1	0	0	0	3
$\sum$ NA		955.955 21,8 %	1.895.487 43,2 %	3.535.370 80,6 %	4.198.226 95,7 %	10.585.038

Altogether, about 57% of all relationships are supported by a single source of information only, dominated by metadata and ambulatory outpatient care contacts. Without both minor sources, about 3% of all relationships would get lost, while reducing to information originating only from social insurance institutions' metadata, nearly 22% would not be seized. As a result, inpatient data and conceivably reimbursed prescriptions could be dropped without major losses. However, relying solely on metadata could result in a larger bias.

The coherence of personal information and the number of hints per source are depicted, utilizing a tableplot in figure 3.30 on the following page. Two main sections can be found, the upper 22% where no metadata is available and the rest.

[16] There is exactly one hint for about a third of all relationships. Only up to 5 hints are recorded for 75% of all relationships, while nearly 22% of all pairs do not have any hint from metadata at all.

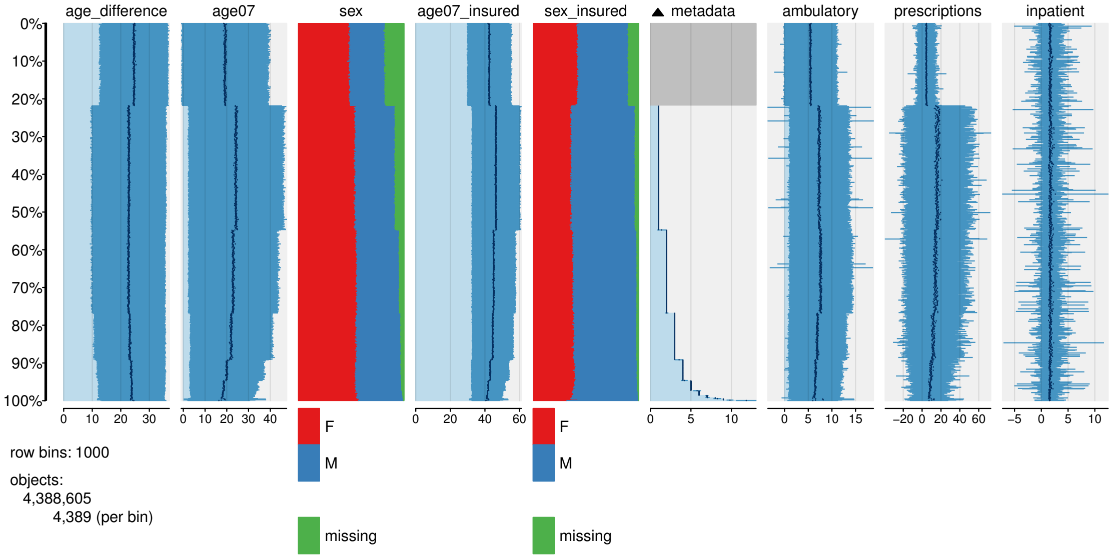


Figure 3.30: Tableplot of personal information and number of hints per source, sorted by hints from metadata

First, most likely influenced by the data generating process, missing values for both the gender of insured and dependent persons are correlated with missing information and the number of hints from metadata. Moreover, the number of hints from ambulatory outpatient care and prescriptions is lower for relationships where no metadata is available. Because the difference in age for couples without information from metadata is higher on average, relationships between children and parents might be overrepresented. Additionally, the higher proportion of females in the same section can be interpreted as another incidence.

Second, both the ages of the co-insured and insured partners seem to be slightly inverse proportional to the number of hints from the metadata, which complies with the marginally negative correlation coefficient. This trend implies that young adults and children tend to have more entries in the metadata tables, potentially because of multiple or changing insurances. On the other hand, it might also be an artifact of varying granularity of the records from different insurance institutions. Nevertheless, the gender of both partners tends to be more complete the more hints from metadata are available.

Figure 3.31 on the next page shows personal information and the source of hints for relationships where no evidence from metadata is available[17]. Sorted by the number of ambulatory outpatient contacts, the tableplot summarizes the nearly 1.000.000 relationships not backed by direct encoding in the insurance institutions' metadata. Due to the

[17] Nevertheless, the variable is displayed as empty in the last column of the plot to emphasize the applied subset.

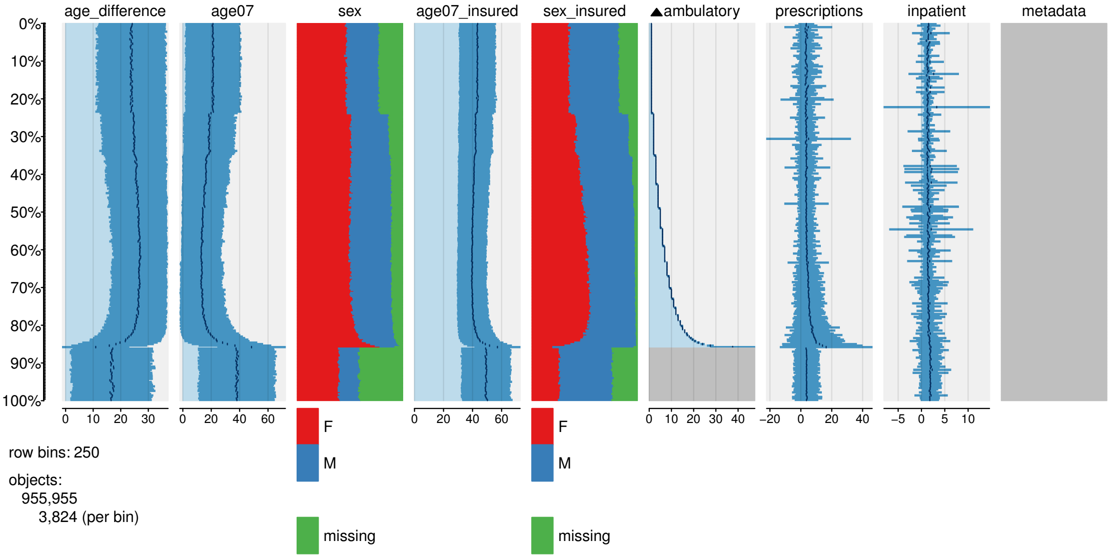


Figure 3.31: Tableplot of personal information and number of hints per source, sorted by hints from ambulatory outpatient care, where no hints of metadata are available

smaller number of datasets included, only 250 bins are calculated to keep the resolution similar to figure 3.30 on the facing page.

It can be observed that the number of missing values for gender is correlated with the number of hints for ambulatory care. Consequently, a large proportion of the corresponding variables are missing in case there are no hints from either outpatient contacts or metadata. This implies that findings induced from other sources than the insurances' metadata can be expected to be mostly valid. As described before, there seems to be a rather large number of children present in this subset, while patients with the most ambulatory contacts tend to be older.

Next, the number of hints per source, age group[18], and gender of the dependent (in figure 3.32 on the next page) as well as insured (in figure 3.33 on page 75) persons is summarized. It is important to note that every single boxplot comprises a different number of relationships, which cannot be determined directly from these figures. The label *med.*, an abbreviation of *medication*, is applied instead of *prescriptions* to enhance readability. Due to the skewed distribution of hints and rather extreme outliers, the y-axis is log10-scaled.

Figure 3.32 on the next page shows large differences between the sources in relation to age and gender. Generally, the largest number of hints per relationship is documented

[18] The age of insured and co-insured persons is dichotomized at age 27. Several cut-off points (e.g., 18, 20, 27) have been evaluated. In general, no interesting differences could be observed. As any of these values can be substantiated by previous findings and general knowledge about the Austrian social insurance system, the selected threshold is selected arbitrarily.

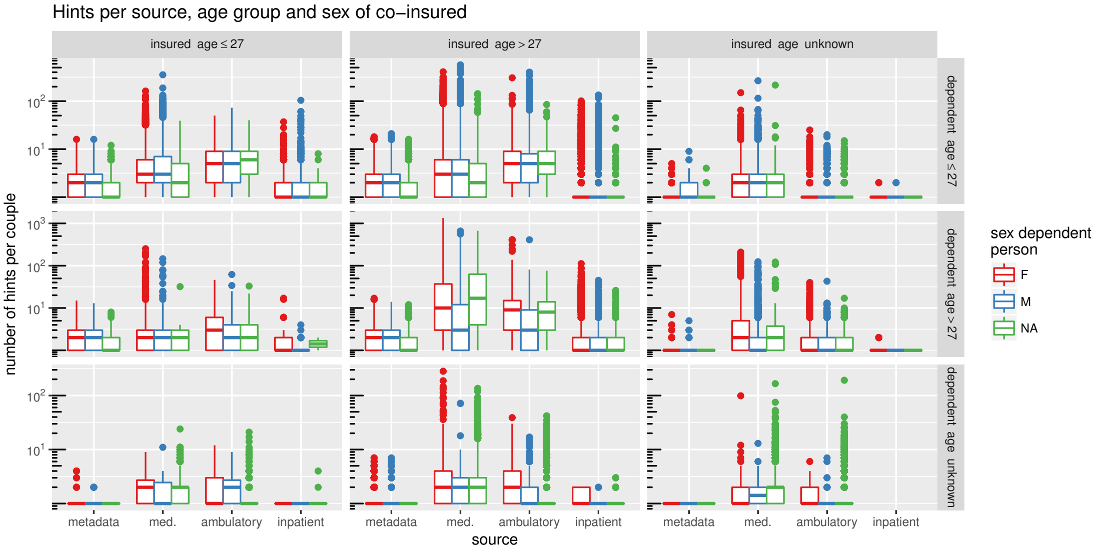


Figure 3.32: Boxplot for hints per source, age group and sex of the dependent person

for prescriptions (i.e., medication), tightly followed by ambulatory outpatient contacts. For *metadata*, a median of exactly 2 can be found for all relevant cases (i.e., neither age nor sex missing). Most interestingly, there seem to be a few cases with a large number of recorded prescriptions where the age of the co-insured person is unknown.

While the medians and spread of co-insured males and females are rather similar in most cases, there is a large difference for prescription and ambulatory outpatient data in the center plot where both ages are above 27. In general, more adult females are co-insured with males than vice versa. This fact could influence the distribution of hints per relationship in case co-insured males are not depending on their partner as often (i.e., not for the whole 2 years of available data). Because it is not possible to estimate the duration of a co-insurance, this assumption cannot be investigated any further.

Figure 3.33 on the facing page shows mostly the same results as the previous one. Every triplet of sex (male, female, not available) represents the same number of cases as in figure 3.32. Solely the proportion of the genders changes relative to each other. Two main observations can be found. First, if the age of the insured or dependent person is not known, it seems to be more likely that also the gender is not documented. Second, the differences between males and females in the center graph for the sources medication and ambulatory care are mirrored because heterosexual relationships between adults are more common in this dataset.

Introduced in figure 3.16 on page 55, the age-age matrix allows deeper insight into the source of relationships in dependence of sex and age. The sum of all hints per age and gender is presented in figure 3.34 on the next page for the cleaned dataset where persons

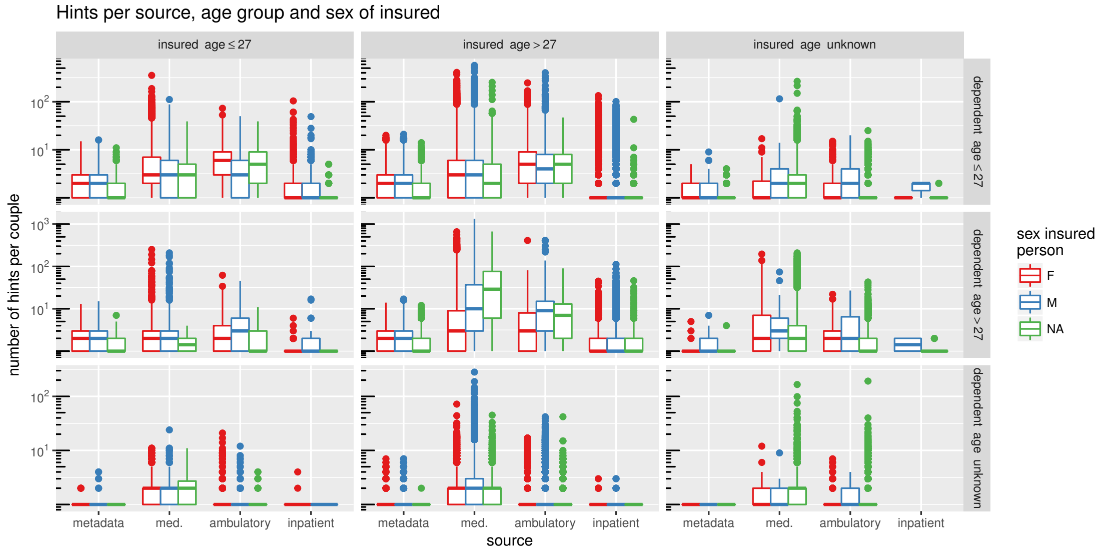


Figure 3.33: Boxplot for hints per source, age group and sex of the insured person

with unknown gender and extreme ages above 100 years are removed.

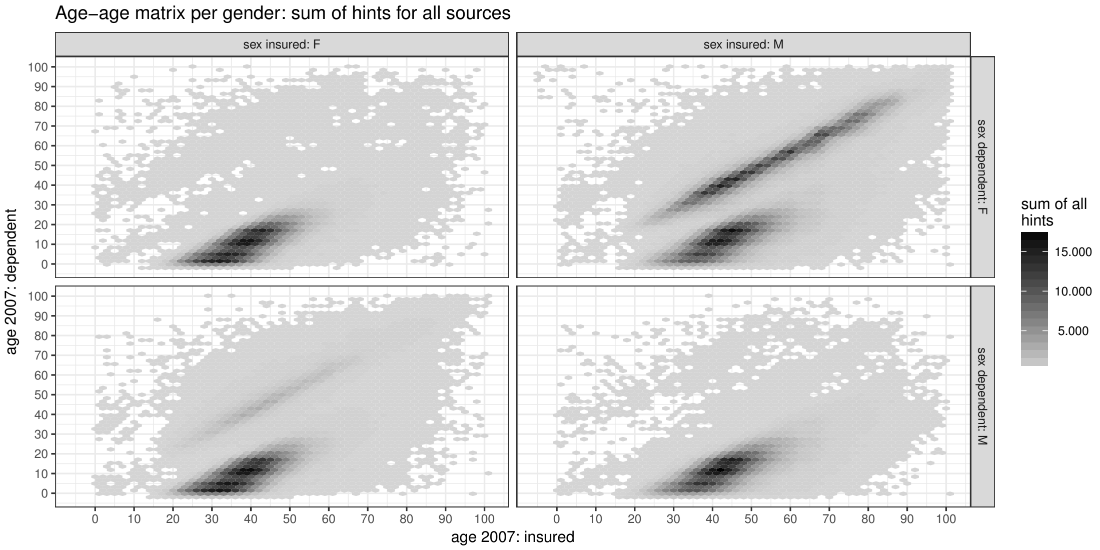


Figure 3.34: Age-age matrix for the sum of hints from all sources

The main distinction to figure 3.19 on page 58, which shows the number of relationships instead of the sum of hints, is the larger difference between the identifiable clusters and the surrounding noise. It can be concluded that there tend to be more hints for relationships associated with one of these clusters than for more uncommon or even unrealistic ones. As a result, the number of hints might be a quality indicator of an identified relation.

Figure 3.35 visualizes the median number of different sources in each bin, faceted by gender. In case the median is located exactly between two different values, their average is utilized and therefore also included in the color scale with the same color as the next higher entry. Not only the rather skewed sum of all hints but also the number of different sources hinting at a relationship can be utilized as a measure of certainty. Between 1 and 4 sources per relationship are possible. For any combination of age and gender of the insured and co-insured person, all 4 values might occur. Per combination, and as a result, also for any hexagonal bin, a distribution of the sum of sources exists with a different number of cases.

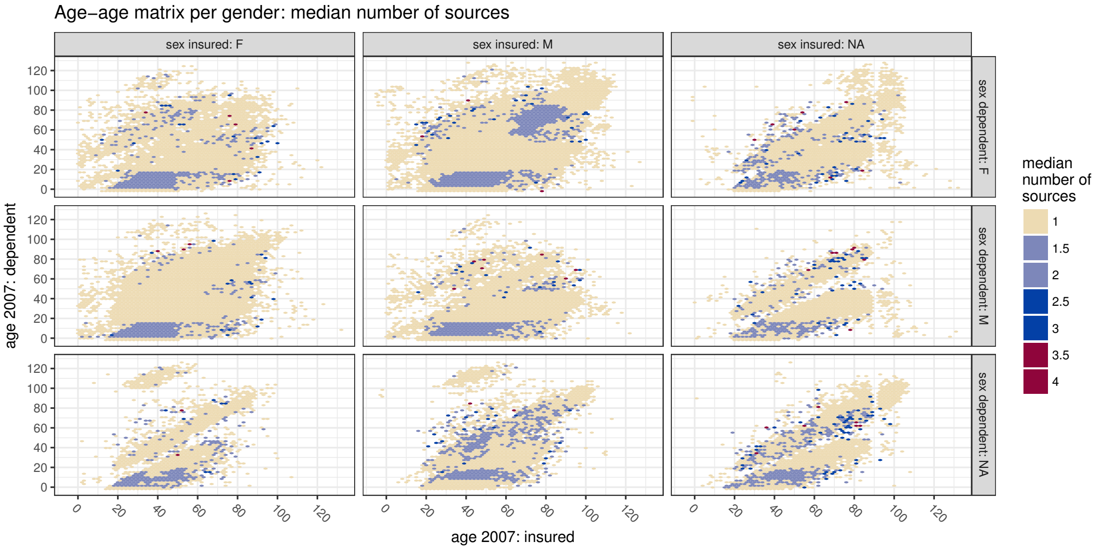


Figure 3.35: Age-age matrix: median number of sources

An unexpected result is displayed in figure 3.35. While the background noise outside the previously identified clusters mostly shows a median number of sources of 1, the clusters are not standing out as before. A larger median of the number of sources inside these clusters can only be observed for children and teenagers without younger adults as well as for older couples above the age of 60 for both partners, where a female is co-insured with a male. Most interestingly, a value above 1 seems to be widespread for pairs where either one or both genders are not known. These plots also show higher values sporadically. On the one hand, this might be a result of bins with rather few pairs, on the other hand, there might be unknown effects in place.

As the median of sources per bin equals one or two in most cases, the distribution of the number of sources seems to be rather skewed. It is therefore not necessary to investigate the lower quartile (25%). The upper quartile (75%) is shown in figure 3.36 for the reduced dataset where age and gender are not missing[19]. A similar color scale is implemented as in the previous plot. Values are rounded to full numbers.

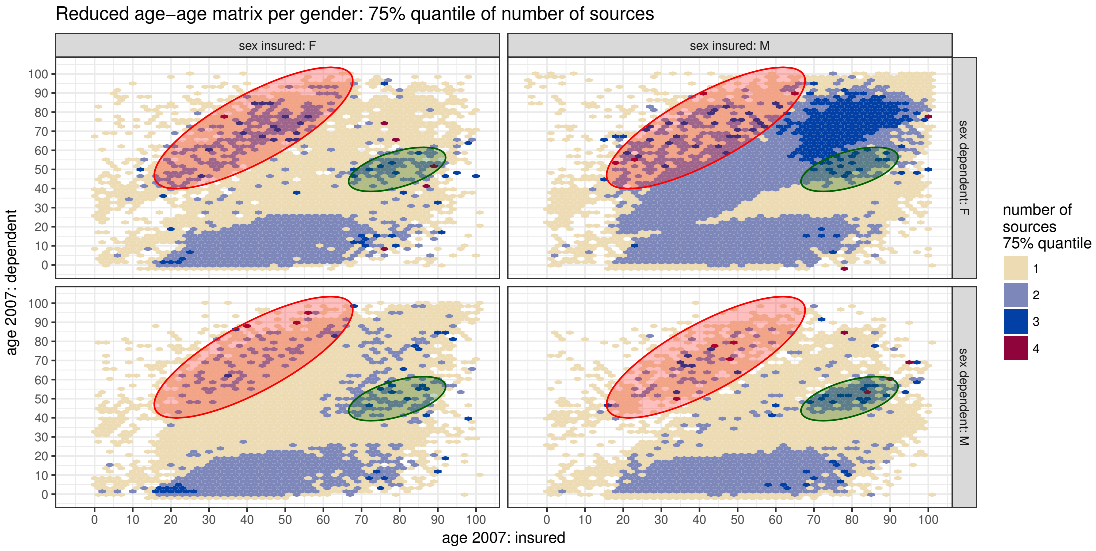


Figure 3.36: Age-age matrix: 3rd quartile of number of sources

In comparison with the preceding figure showing the median number of sources, all previously identified clusters can be found in figure 3.36. Again, there are more hints to a relationship for older couples (3rd quartile equals about 3) and newborns co-insured with young mothers. Most interestingly, the area of co-insured adults of similar age is only showing up for females co-insured with males. Furthermore, the cluster representing dependent children stretches till at least the age of 70 for the insured parent.

Two interesting structures outside of the plausible clusters can be identified in different sections of the plot. First, especially co-insured females with a difference in age of about 20 to 40 years are highlighted with a red ellipse. This cluster can already be spotted in figure 3.18 on page 57, displaying the number of relationships on a log-scale. Furthermore, it also appears for females co-insured with males where it is not clearly distinguishable from the main cluster around the main diagonal. Second, a smaller cluster, most notably for co-insured males with an age difference of around 30 years, is highlighted in green. It seems like this structure is part of the upper end of the cluster determining children in, e.g., figure 3.19 on page 58. Again, it appears in a less outstanding form for all combinations of genders and is hard to differentiate from the main cluster of females co-insured with males. Summarizing, both clusters appear primarily in couples with several different sources hinting to their relationship. They are mostly hidden in other analysis and their origin cannot be explained. If these structures would be replaced by

[19] The cleaned dataset is introduced with figure 3.19 on page 58.

random background noise, the cluster determining co-insured adults would occur nearly homoscedastic.

Both clusters can also be found in a similar plot presented in figure 3.37 showing the spread[20] with the number of sources for a relationship in an even more prominent form. A cluster holding dependent adult males co-insured with females is showing up more prominently than before. Altogether, this measure discriminates between plausible relationships and background noise more clearly than the absolute number of relationships did.

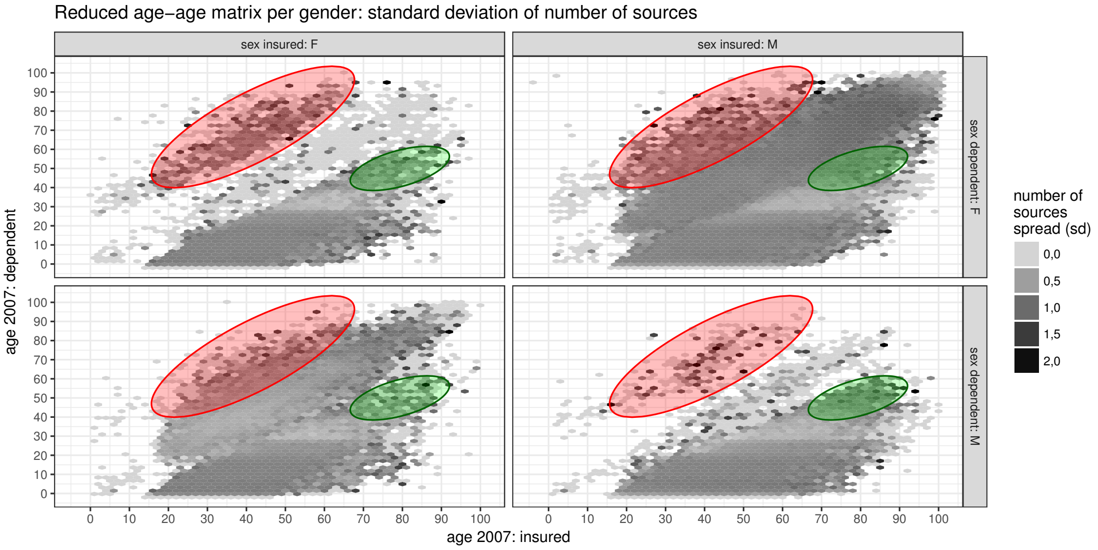


Figure 3.37: Age-age matrix: standard deviation of number of sources

Finally, the influence of a single source is analyzed in more detail. As pointed out before, mainly metadata and ambulatory outpatient care add significant amounts of information. Therefore, these two resources are focused on.

Figure 3.38 on the facing page shows the age-age matrix faceted by gender for all relationships which are not supported by metadata. Mostly the same clusters and noise can be observed as in figure 3.17 on page 56. Couples with one or both genders missing might be underrepresented while the distortion most likely caused by mishandling the year 2000 (highlighted with a red ellipse) is slightly more prominent. These impressions are most likely also affected by the lower number of datasets involved.

Nevertheless, figure 3.38 on the facing page is relevant to emphasize the contrast to figure 3.39 on the next page, where only relationships supported by metadata are included.

[20] The standard deviation is chosen as measure for spread. A robust measure is omitted as only values between 1 and 4 are possible, no extreme outliers are therefore present, and single bins have a rather low number of observations.

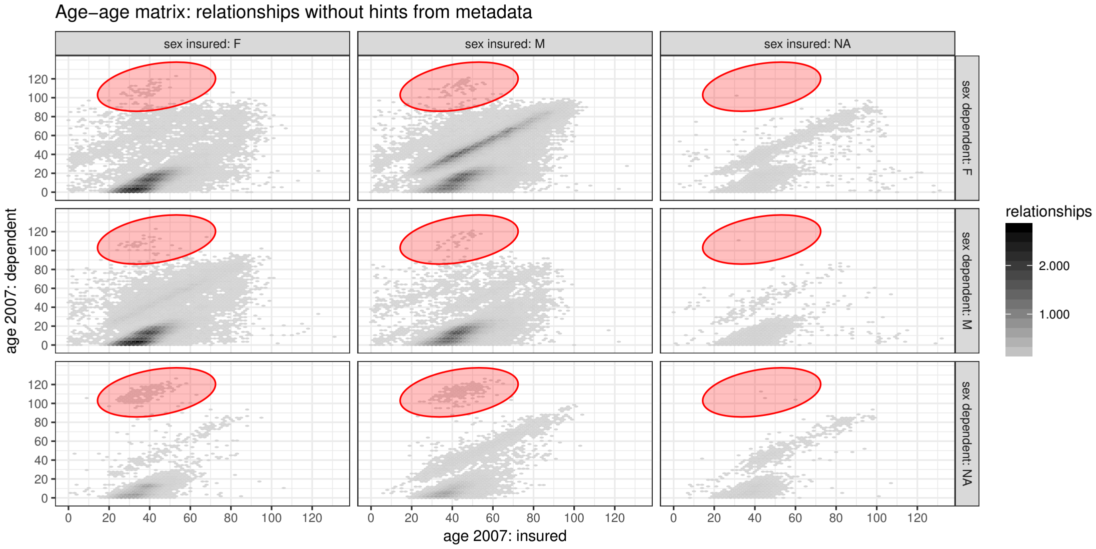


Figure 3.38: Age-age matrix without hints from metadata

Overall, there seems to be less background noise and hardly any compulsorily insured persons under the age of roughly 18.

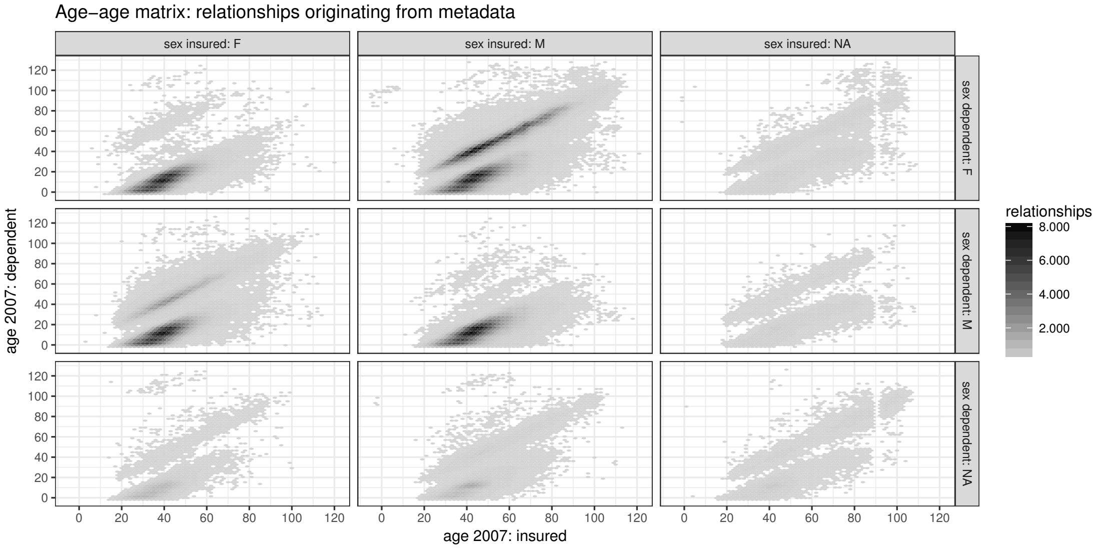


Figure 3.39: Age-age matrix: exclusive contribution of metadata

Similar results appear in figure 3.40 on the following page, where the contribution of ambulatory outpatient data is plotted for all cases where no metadata is available.

Influence from prescriptions and inpatient data is not considered here due to their relatively small proportion. The lack of insured persons under the age of 18 and the higher fraction of young mothers is most important. Furthermore, there are hardly any same-sex relationships recorded for adults and the erroneous encoding for children born after the year 2000, highlighted by the red ellipse, disappears nearly entirely. Extreme ages of 100 years and above hardly exist. Nevertheless, a small, still unexplained cluster highlighted in green, holding dependent persons roughly 20 years older than their insured partner still exists for couples with known gender.

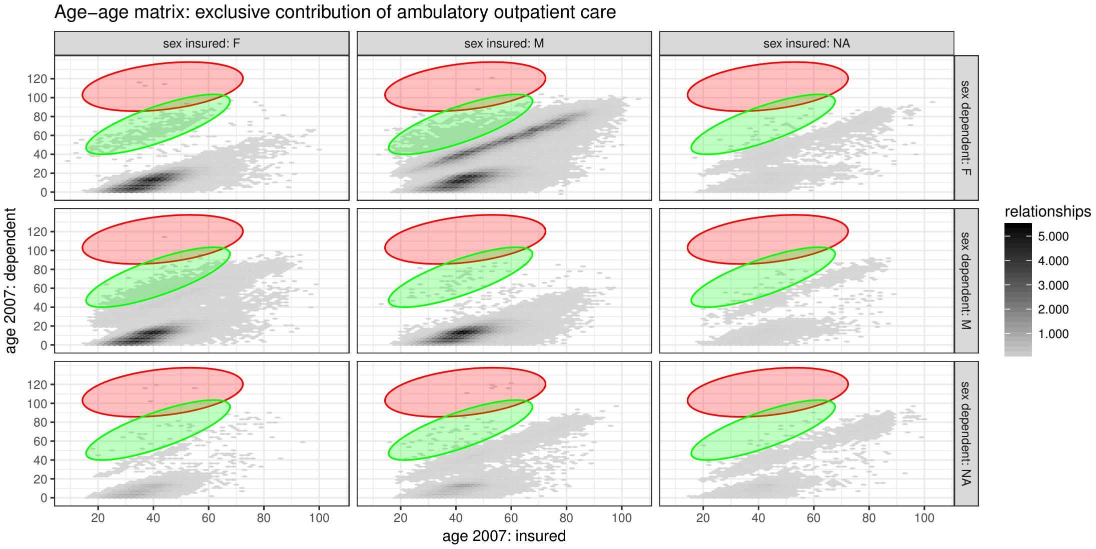


Figure 3.40: Age-age matrix: exclusive contribution of ambulatory outpatient data

Relationships supported by inpatient hospital data, presented in figure 3.41 on the next page, seem to be rather clean, mostly lacking extreme outliers. Although there are comparatively few relationships backed by this source, two groups stand out. Both the relationship between mothers and (very) young children and older couples above the age of about 60 seem to be overrepresented. The first group might be caused by child delivery conducted at a hospital and the second group due to the natural progression of epidemiology and diseases.

Finally, the exclusive contribution of reimbursed prescriptions is plotted in figure 3.42 on page 82. On the one hand, there are hardly any pairs where one or both genders are unknown. On the other hand, the background noise of relationships not associated with the main clusters is apparent. More detailed analysis not included here suggests that much noise in the entire dataset originates from this source, especially in relation to the total amount of additional information provided. An overrepresentation of insured persons under the age of 20 years is most notably. It can be speculated that this potentially wrong information is a result of the prescribing process established in Austria in 2006

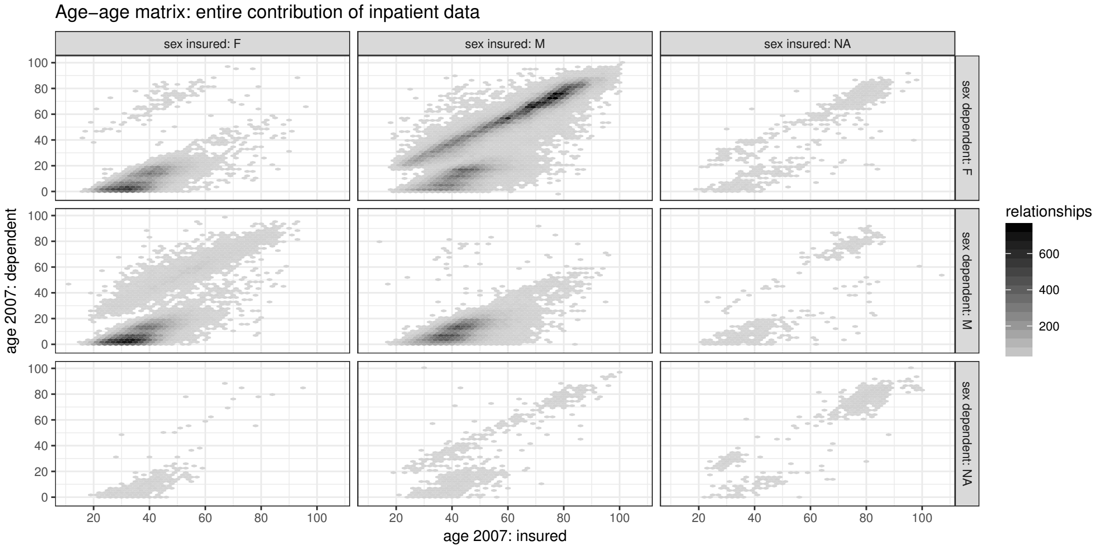


Figure 3.41: Age-age matrix: entire contribution of inpatient hospital data

and 2007.

Although these facts might suggest that this source must be avoided completely, there is another important structure to be found. The exclusive contribution of filled prescriptions shows a previously unobserved maximum for children around the age of 20, co-insured with adult females and in a less distinctive form co-insured with males. Assuming that the adult, insured part of these relationships are parents, the mothers would be between 18 and 35 older than their children.

Summarizing, all four sources show differing patterns, completeness, data quality, and interdependence. Although two of them, i.e., metadata and ambulatory outpatient data, contribute the most, and information from prescriptions adds noise to the dataset, every origin contributes new and most likely relevant information.

While losing adult couples is not expected to influence the study's outcome significantly, every missing child and therefore potentially misclassified parents as a childless couple lowers the discrimination of the cohorts. As a result, the unobservable power of the study would most likely suffer from excluding noisy information from prescriptions. In case clusters are extracted automatically without specified rules, not only the total number of relationships but also the spread of the number of different sources is worthwhile to integrate as an additional measure.

3.1.7 Conclusions

The most important variables related to extracted relationships from claims data have been described and explored thoroughly. From univariate summaries to simple multivariate

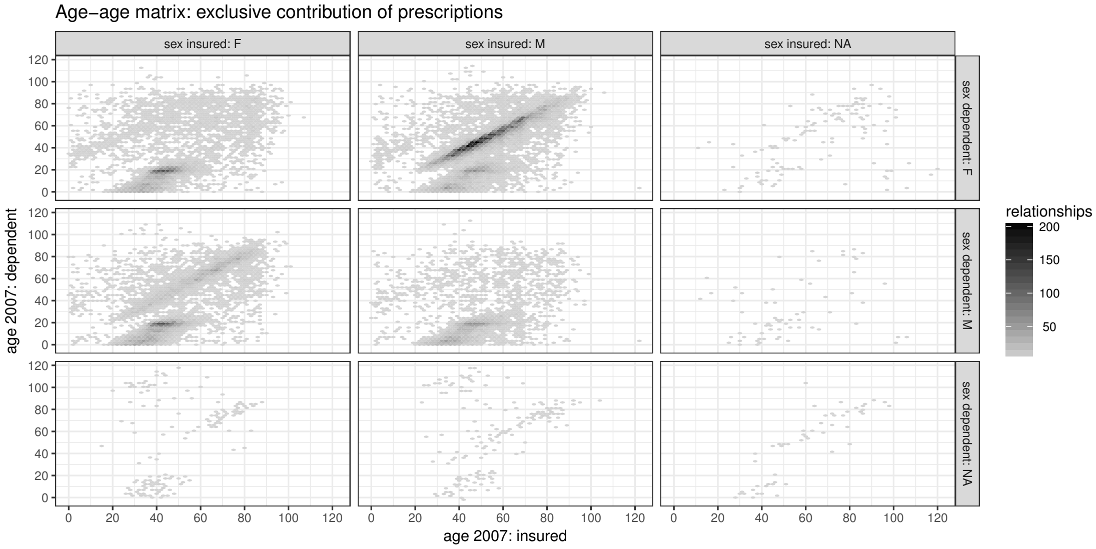


Figure 3.42: Age-age matrix: exclusive contribution of prescriptions

interactions and especially the structure of missing data, several aspects have been explored, visualized, and interpreted.

Naturally, many more multivariate exploratory analyses are feasible but not considered as relevant for the study protocol at hand. Influence of the concrete data generating process and data's originating social insurance institution, regional and temporal parameters are the most common additional variables which are expected to add bias. For example, age, common dependence of couples, number of children, and data quality are known to vary depending on the insurance institution and regional classification. Furthermore, for ambulatory outpatient contacts, only data mirroring former health insurance vouchers are utilized, ignoring information about actual visits to a doctor's office.

Summarizing, a mixed conclusion can be drawn. On the one hand, various explicable and even previously unknown structures have been identified. Missing data causing single relationships to be omitted from the final cohort selection process has been described, including its interdependence with various variables. Especially the year-2000 error and the influence of the *Zentrale Melderegister* have not been documented before. On the other hand, generally expected clusters have been found, suggesting reliable data and promising quality. Moreover, the necessity to include at least ambulatory outpatient care data as an additional source to insurances' metadata has been demonstrated.

During cohort selection, some of the most unrealistic couples are left out by design. Additionally, final cohorts must be evaluated for plausibility and dubious information should be left out entirely. Therefore, all relationships of a person involved in a couple where information (e.g., the age of the co-insured partner) is missing will be removed to ensure valid discrimination of couples without children and parents. As this procedure could introduce a selection bias by not randomly removing couples, the process will be evaluated thoroughly. Moreover, persons who are deceased before 2006 or are not part of the standardized research population of GAP-DRG are excluded from the final cohorts.

Summarizing, a solid foundation for the following cohort selection has been created. As a result, inclusion and exclusion of relationships and potential participants to or from the cohorts are induced and justified by findings in the available dataset. Furthermore, the collection of all available relationships provides a new means for other studies and data quality assessment.

3.2 Genealogical information and cohort extraction

In this section, the extraction procedure of the cohorts from the data on co-insurances is described in detail.

3.2.1 Selection criteria

The selection criteria defined in the study protocol for a relationship between children and an adult (supposedly a parent) and co-insured adults (purportedly spouses) with or without children are highlighted in figure 3.19 on page 58, including areas of mixed cases.

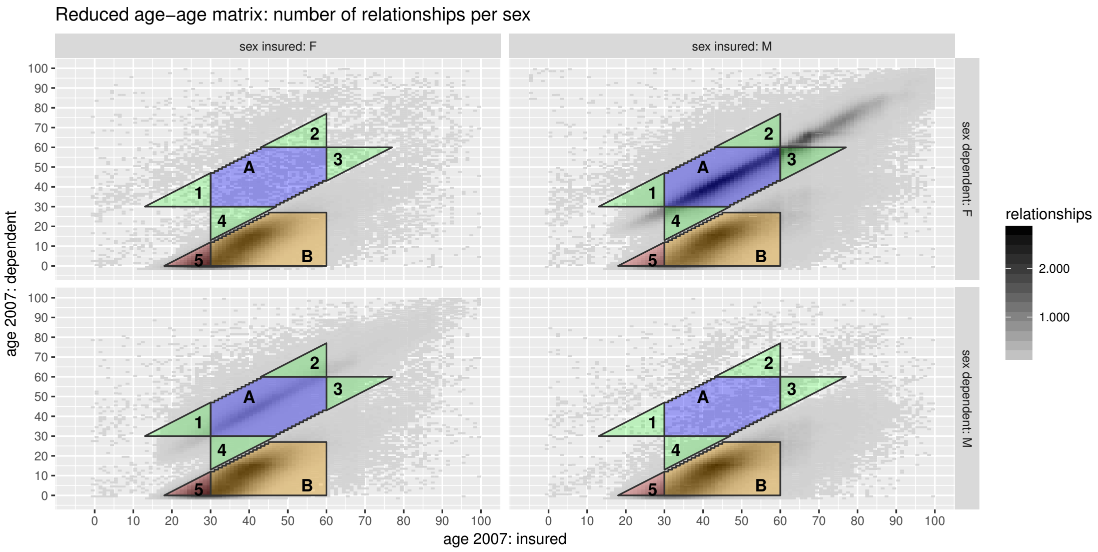


Figure 3.43: Reduced age-age matrix: cohort selection criteria

The following sections are highlighted in figure 3.43:

A All adult couples with (direct) co-insurance are included in these blue areas. It can be concluded that the difference in age of up to 17 years fits the apparent cluster and could even be reduced slightly. Both partners of each couple in this area are potentially included in the final cohorts.

B The relationship between a child co-insured with an adult who is himself/herself potentially part of the final population. In case the adult part of these relationships is not associated with another adult directly or indirectly (e.g., by a common co-insurance with a mutual child) fitting the defined criteria, he or she cannot be selected for further analysis.

1 - 4 Four green areas extend region *A* in different directions. One partner of each couple located there can be potentially included in the final cohorts (i.e., aged between 30 and 60), while the second part is too young or old, but the difference in age is fitting. It is crucial to respect these relationships for discrimination of the cohorts although these areas are located mostly outside the main cluster. Especially younger parents where one partner is born after 1977 (i.e., younger than 30 in the year 2007) and older couples where one partner is above the age of 60 could be misclassified or missed entirely in case these areas are neglected.

5 Young adults below the age of 30 who are associated with children are included in this area. Neither of both parts of these relationships can be included in the study cohorts. Nevertheless, adults located in this area are supposedly parents and could classify another adult in section 1 or 4 as a parent indirectly.

Before the cohorts are selected, the following datasets are extracted from the collection of relationships:

Relationships per person as either compulsorily insured or dependent. As presented in figure 3.26 on page 67, figure 3.27 on page 67 as well as section 3.1.2 on page 41, and concluded in section 3.1.7 on page 81, a small group of persons with many co-insurances exists. These few but extreme outliers can be interpreted as reporting errors. Furthermore, practical reasons and common knowledge require a limitation of the network size. During the following extraction of networks of relationships, single nodes (i.e., persons), which have a large number of connections, drastically increase the size and required calculations without adding reliable information. Therefore, this extract is utilized to limit the following datasets to persons with a maximum number of relationships.

Relationships for recursion is a specially prepared subset of all reported co-insurances, optimized to recursively traverse networks of relationships. Connections introduced by persons with more than 20 relationships are removed. There are two versions of this dataset. For the first one, no additional restriction is applied. It is utilized for illustration of the co-insurance networks presented in chapter 3.2.2 on the next page. The second one is essentially identical to the first one, with the difference that all relationships where a person is lacking information on gender are left out. Differences and their impact on the outcome are documented in chapter 3.4 on page 110.

Potential participants is a list of all persons who are potentially included in one of the final cohorts. Both compulsorily insured and co-insured persons are selected with the following restrictions applied. It is important to note that most restrictions overlap. Therefore, removed persons are excluded due to several criteria in many cases.

Age in the year 2007, between 30 and 60

Sex has to be known

Research population: persons must be part of the standardized research population

Death: persons who have died before the year 2006 are excluded

Relationships: persons with more than 20 or none[21] relationships are excluded

Based on these datasets, the exploration of co-insurance networks and cohort extraction is performed.

3.2.2 Co-insurance networks

Visualizations of some manually chosen networks are presented and interpreted in figure 3.44 to figure 3.46 on page 88. As described above, nodes represent persons and edges are defined by co-insurance. Nodes are colored and labeled by the person's gender (i.e., blue for *M*ales, red for *F*emales, green for *N*ot *A*vailable) and the person's age is appended as a note. The direction of each edge is indicated by a gray arrow from the co-insured to the compulsorily insured person. Naturally, this arrow must point in at least one direction, but can also go in both directions.[22] Edges are annotated with the absolute difference in age. Their width represents the number of sources, documented in 3.1.6 on page 69. All nodes, labels, and annotations are automatically arranged and thus may not be laid out optimally, although different objects ought to repel each other. Bounded by a dashed border, each network is labeled with an arbitrary title and a capital letter. To save space, some networks are grouped by content in individual figures.

Each presented network holds at least one person who is potentially selected to the final cohorts. It is also possible and probable that several persons in the same network are

[21] Due to additional filters applied to networks of relationships described in the following chapters, persons might lack any accepted co-insurance. It is therefore not possible to identify any partner (of children), which removes them from the collection of possible included study participants.

[22] In case the arrow points in both directions, both partners are co-insured with each other in alternating roles. There are two overlapping but visually not distinguishable arrows printed, possibly resulting in a slight misrepresentation of the weights (i.e., the arrow with the larger width is dominating) and an overlap of annotations with equal content.

obliged to be included. In the following descriptions, persons will be referred by their gender and age as unique identifiers because the exact layout of each network might change due to the dynamic creation of this document.

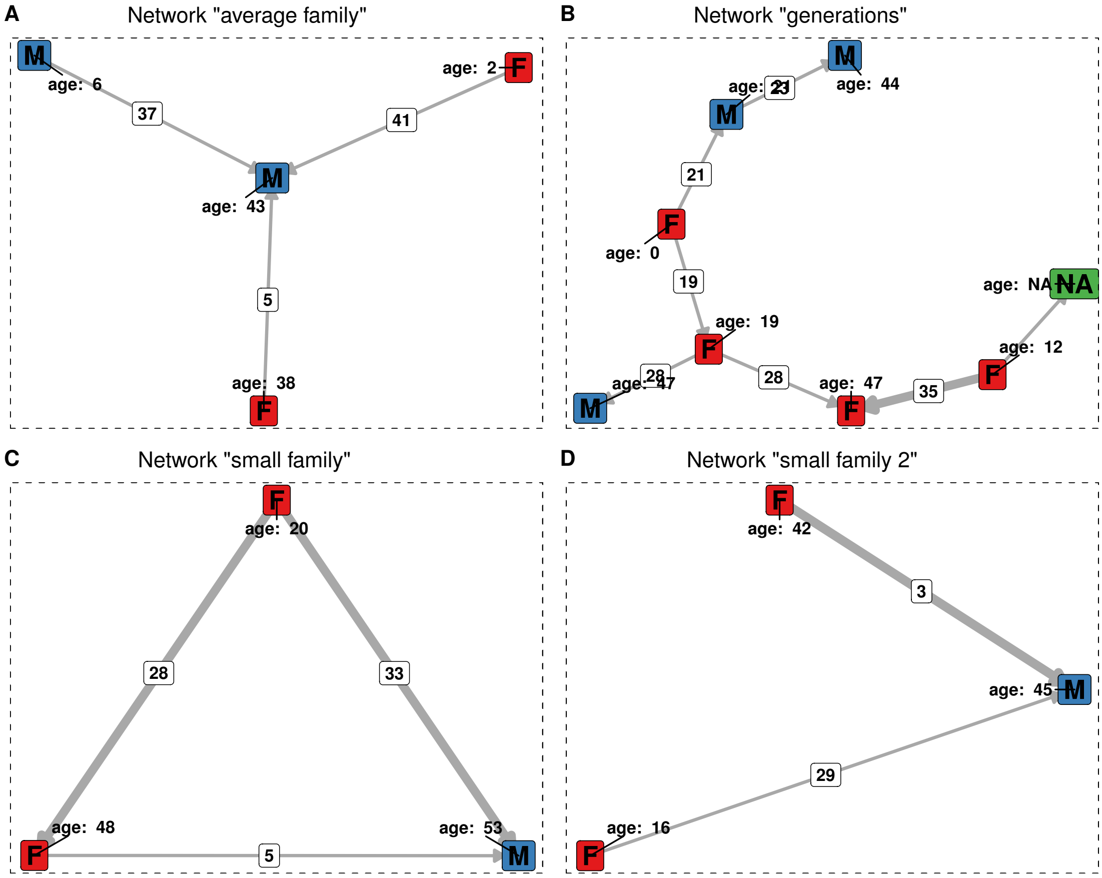


Figure 3.44: Network examples "families"

Figure 3.44 shows the networks of 4 different families. They are supposed to show widespread cases including potential pitfalls in the extraction process.

A: "average family" The first network is meant to represent a rather typical family, where two children and a female are all co-insured with the same male. No data is missing and all differences in age as well as absolute age are fitting the expectations. While the adult male is identified as a parent directly, the (adult) female is classified indirectly due to co-insurance with her partner.

B: "generations" This network shows a more complex case. At least two *complete* families and one male co-insured with his father can be identified. First, a family consisting of 47-year-old parents (male and female) with two female children at the age of 12 and 19 can be identified. The 12-year-old child is additionally co-insured with a person of unknown age and gender, which might be a data error. Differences in age between all four family members fit the selection criteria and are plausible. Second, the older daughter has a child (newborn female), which is also associated with a 21-years-old assumed father. Both young adults could be a valid family but are excluded due to their age. Third, the young father is co-insured with a 44-year-old adult male, presumably his father. As the last one is not associated with an adult partner with a maximum distance of 2, he has also to be excluded. Summarizing, this network holds two parents who are included in the final cohorts. Although this network is larger than the other ones in the same figure and shows erroneous data, it seems to be valid and plausible. It can be concluded that ignoring such more complex networks due to their size or invalid data would (non-randomly) remove eligible data. Additionally, this case supports the restriction of the maximum distance for identification of children and partners to 2.

C & D: "small families" Finally, two examples for small families with one child each are presented as a contrast to example *B*. Network *C* shows a cycle in case the directions are ignored.

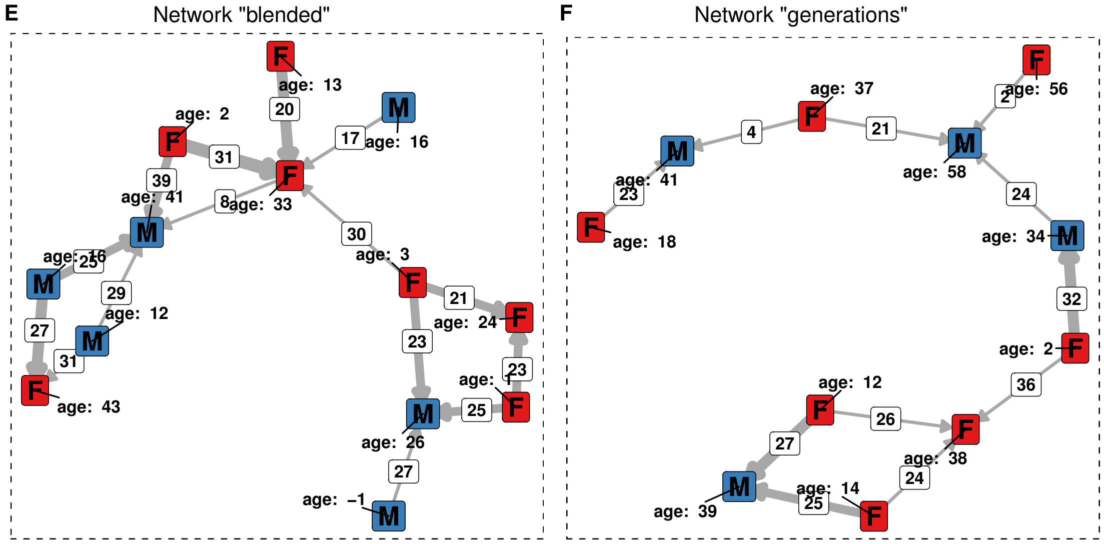


Figure 3.45: Network examples "blended"

Figure 3.45 shows two more complex cases of patchwork families which are plausible and can be considered as common. These illustrations emphasize the variety of family networks in the dataset, the potential of the presented method to identify compound structures, and the risks of applying too restrictive clauses during cohort extraction. Both networks allow speculation about family histories.

E: "blended" Up to three valid couples with children can be identified. Beginning with the youngest one, a couple aged 24 and 26 in the year 2007 has three children born in 2004, 2006, and 2008 (age 3, 1, -1). The associations between children and adults are backed by several sources while the adults are not co-insured directly. Their oldest daughter is also depending on another 30-year-old woman. The second family consists of a 41-year-old male and a 43-year-old female whose relationship is determined by two common children. Additionally, the male is also associated with a 33-year-old female directly by co-insurance and indirectly by a common child at the age of 2. This female might have two older children with a difference in age of 17 and 20. It can be speculated that she founded a blended family with a 41-year-old male. Altogether, there are three persons qualifying as parents and are included in the final cohort.

F: "generations" Two generations can be identified in this network. Starting with a couple in their late 50s, two children aged 34 and 37 can be found. Both have spouses and children on their own. Similar to the network *E*, a 38-year-old female seems to have two children associated with one male and another, much younger child with a second partner. Altogether, 7 persons are included in the study. While the younger generation is (most likely correctly) identified as parents, the older one is regarded as a childless couple because their children are both older than 27. It is important to note that this example does not seem to be atypical and as a result, misclassification, especially of older couples, is likely to be common.

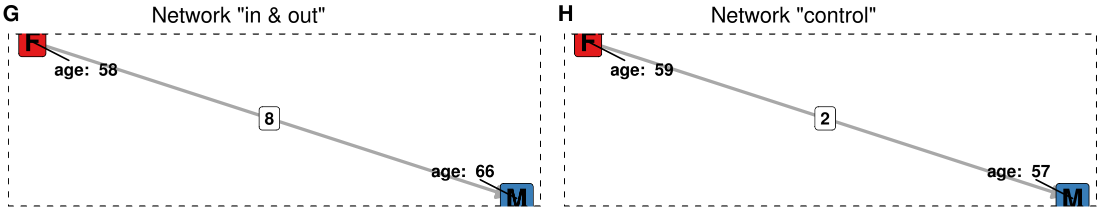


Figure 3.46: Network examples "small"

Figure 3.46 illustrates couples without further relationships. Both can be clearly classified as spouses without children according to the study protocol. Network *G* shows a special case where the male part is more than 60-year-old and is therefore not included in the final cohort, although he contributes by identifying his 58-year-old wife as being in a relationship. This example is located in section 3 of figure 3.43 on page 83 and describes a valid and plausible case.

Figure 3.47 on the next page illustrates two networks with unclear interpretation, possibly leading to deficient classification. In both cases, it seems to be hard to define whether a specific person must be regarded as parent or childless. It can be speculated that these examples show a more complex reality which cannot be strictly discriminated into two

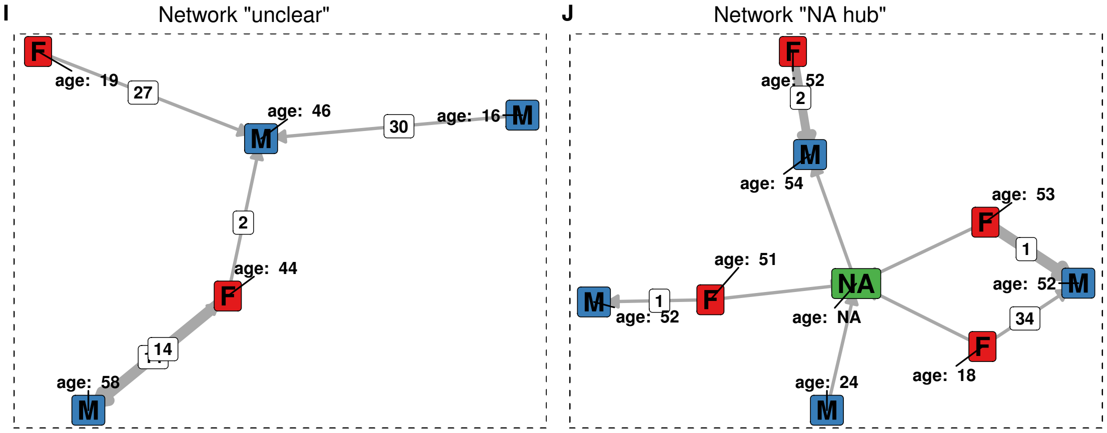


Figure 3.47: Network examples "unclear"

distinct groups. Additional variables and information (e.g., number of hints, number of sources, duration of co-insurance) might give a clearer picture, but also bare the risk of more complex (arbitrary) selection criteria, introduction of additional bias and withdrawal of manual appraisal of resulting networks due to their multivariate nature. Because there is no arrangement in the study protocol for this situation and no clear suggestions can be deduced from the data itself, these novel findings are solely documented.

I: "unclear" The network consists of three adults aged above 40 years and two younger persons at the age of 16 and 19. Both younger ones are associated with one male only. Considering the selection criteria, all three adults are classified as being in a relationship, but only the 46-year-old male and the 44-year-old female are recognized as parents.[23] It is remarkable that the relation between the adult female and the 58-year-old male is based on mutual co-insurance, documented by several sources, while her co-insurance with the other adult male is only retrieved from a single source. Otherwise, the difference in age of couples is more commonly less than 10 years and the age of the two children would be indicative of her being their mother. Summarizing, despite the unclear situation, the defined selection criteria might conceive the situation correctly.

J: "NA-hub" Most likely, cases of misclassification are documented here. Three couples, two without children and one with an 18 year old daughter, and a single 24-year-old male can be observed. Due to their common relation to a person of unknown age and sex, everyone directly in contact with these participants is identified as a parent because of the short distance to the two youngest persons. Summarizing, 4 adults in this network are classified correctly and the two most likely end up in the

[23] The distance between the oldest male and the children is larger than 2. Therefore, their association is not respected.

wrong cohort. The person of unknown property is called an *NA hub* because he or she lacks personal details and connects otherwise disjunctive family associations without adding further information.

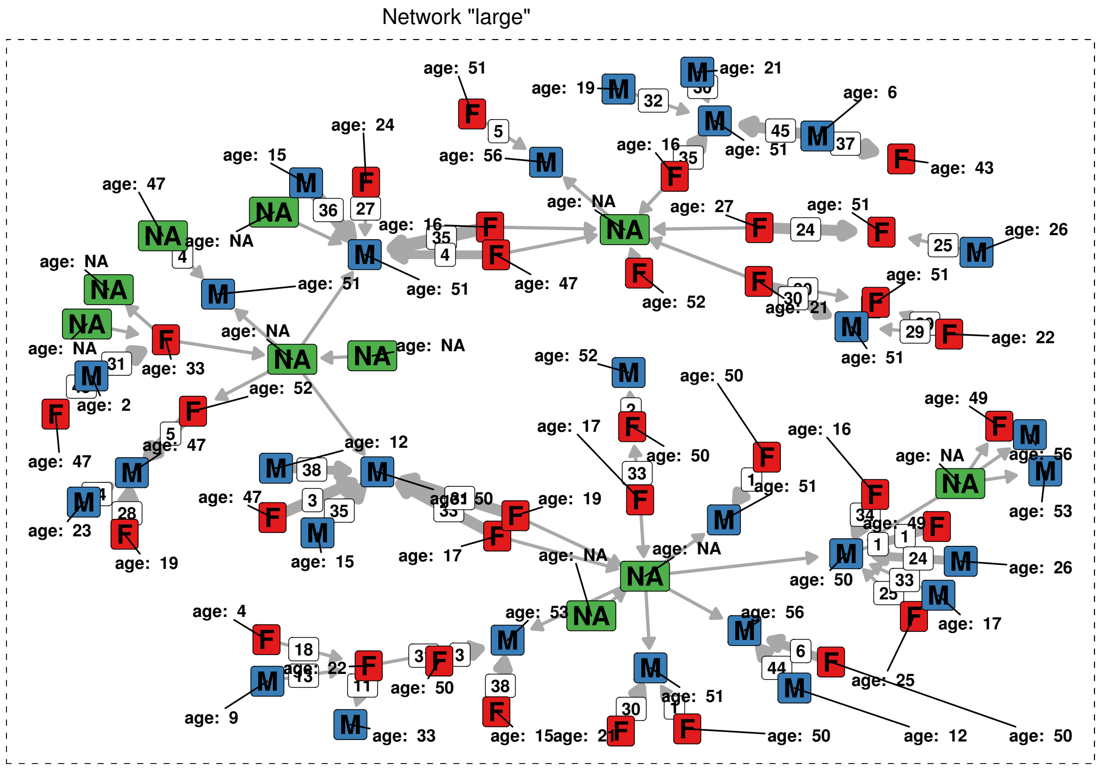


Figure 3.48: Network examples "large"

Figure 3.48 illustrates a large network consisting of several spouses, *NA hubs* and otherwise single persons. It can be clearly observed that nodes which are lacking personal information tend to connect to a larger number of other vertices. There seem to be several valid family associations and some potential misclassifications.

These *NA hubs* are left out in figure 3.49 on the next page. Notably, the entire network splits up in several smaller groups and single persons who are not associated with anyone anymore and are therefore not printed. As a result, some cases might be identified more correctly while others would be removed entirely during cohort selection, possibly leading to misclassification and nonrandom removal of persons. While the data quality assessment of co-insurances in chapter 3.1 on page 33 revealed missing personal property, there has not been enough evidence to omit these cases and their associations entirely. Therefore, it has been concluded to include as many cases as possible to respect the maximum amount of available information. Analyzing not only single persons and co-insurances but the entire network of relationships, new indications that these persons behave not only vastly differently than the majority of cases but that they do not add useful information but even distort the overall picture by e.g., acting as an *NA hub* stack up. Although only single samples can be checked manually and there is no possibility for verification with additional data, there is enough evidence to remove these associations for cohort selection.

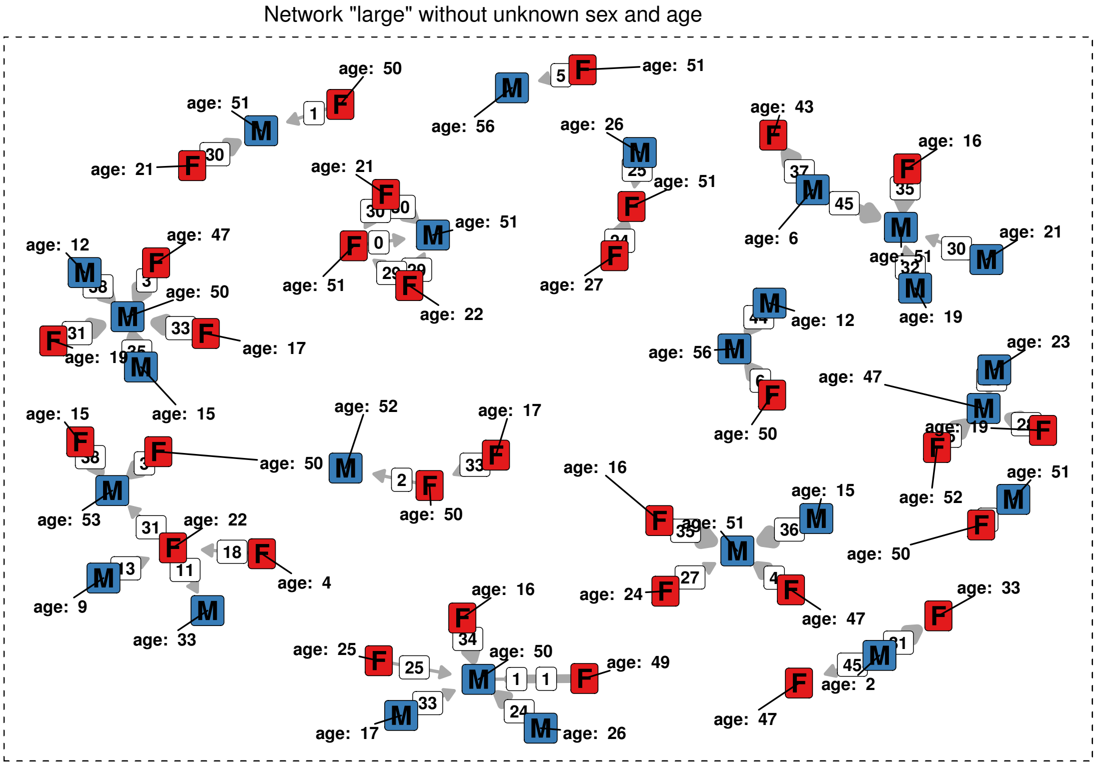


Figure 3.49: Network examples "large": without NA hubs

3.3 Additional personal information

3.3.1 Socio Economic Status (SES)

Figures 3.50 to 3.53 give a summary of both SES variables *soes* and *soes_mean* as well as their differences, split by 5-year age group, gender, and cohorts.

Figure 3.50 on the next page shows the distribution of the original (corrected) average of the historical and current SES. The plot consists of two main sections. In the upper part, the total number of people is directly represented by a stacked histogram. The lower part of the figure contains the same information represented as density curves. Since the area under each curve is equal to 1, the relative proportions can be interpreted more

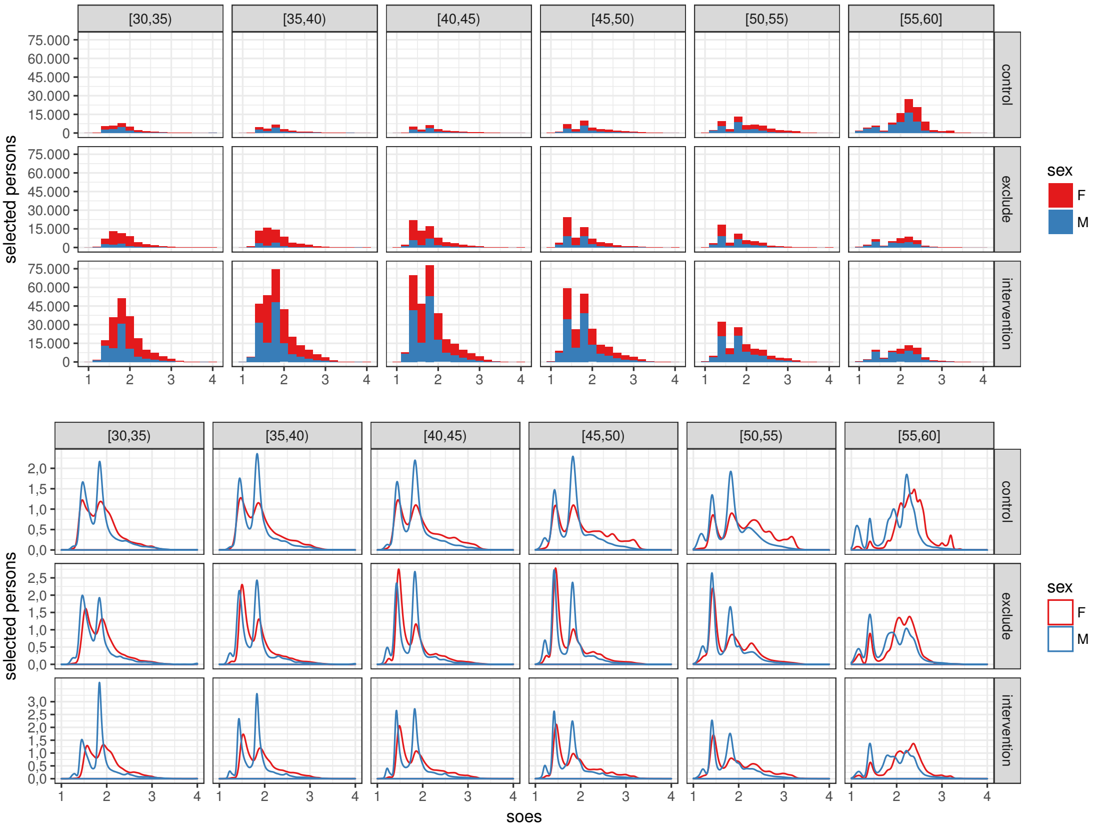


Figure 3.50: Distribution of SES (variable *soes*): split by age group and cohort

directly, while the absolute number of people involved remains hidden. Both areas are divided by age group (columns), assigned cohort (rows), and gender (color).

An increase (deterioration) in socioeconomic status with increasing age can be observed for all cohorts, with group control showing the strongest effect. The majority of individuals appear to be in the lowest third (i.e., between 1 and 2) of the SES spectrum, corresponding to comparatively good SES. This trend shifts only for the highest age group. It could be caused by several effects, either by an assumed general increase in SES with age, by more or different (historical) data included in the original calculations for older individuals, or by a correlation of the fact that older couples are married to each other (and not divorced, for example) with their social and economic well-being. Moreover, gender appears to be correlated with SES in several ways. Males are better off[24] in nearly all cases. On the other side, females seem to be more likely to have a SES above two than males.

[24] i.e., in the density plot, the blue line is above the red line on the left side of the graph and below the red line for most values above two

Other effects such as, e.g., the proportion of gender in the excluded group are discussed in chapter 3.4 on page 110.

In summary, there appear to be several correlations with respect to age, gender, and the cohort used. Additionally, a selection bias regarding older couples can be documented. Although the underlying reasons cannot be determined and discussed within the scope of this project, it seems important to include a variable representing SES in the following analysis.

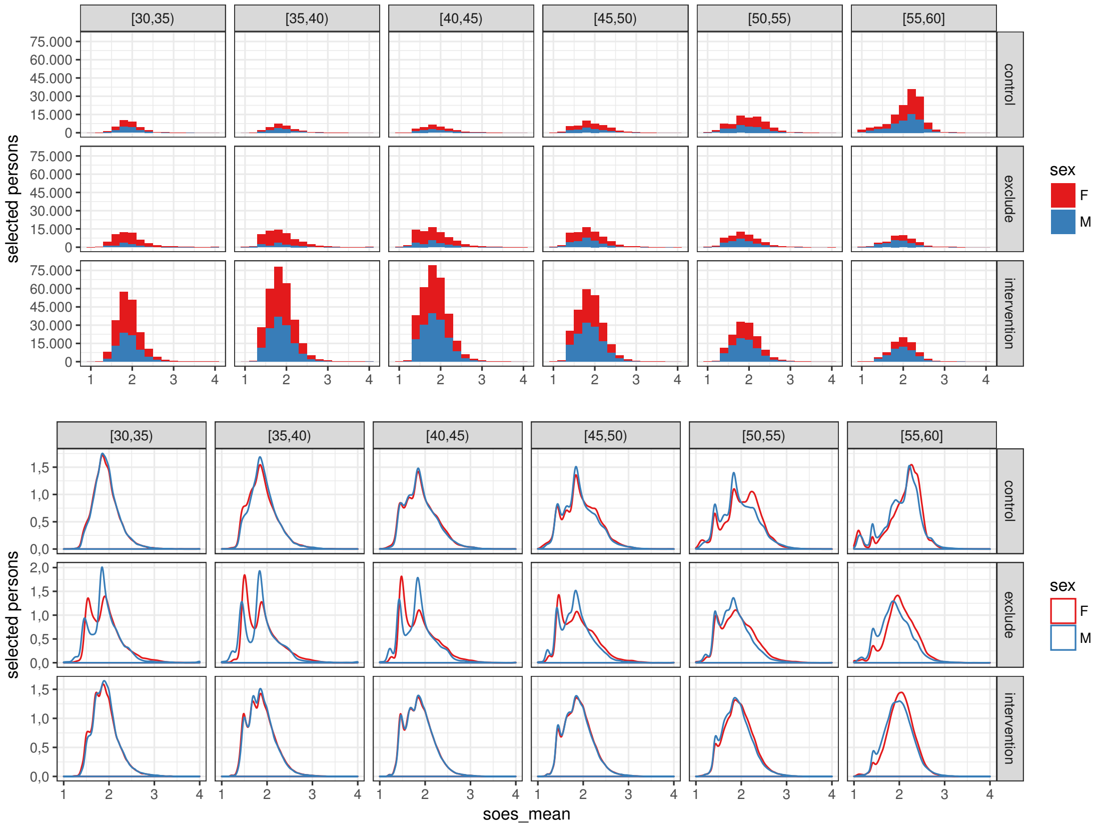


Figure 3.51: Distribution of interpolated SES (variable *soes_mean*): split by age group and cohort

Figure 3.51 plots the interpolated SES values in the same way as the previous graph. For all but the excluded group, a convergence of the density curves of males and females is observed. Since the excluded individuals lack an adult partner (as defined by the study protocol) and therefore in all likelihood do not have additional SES values in their individual networks, this difference is plausible. For the interpolated values, the transition from lower (better) to higher (worse) SES with age is quite strong for the control group, whereas there appears to be only a slight shift for the parents (cohort intervention). There are some spikes in the density curves that are not interpretable and could be artifacts of the original derivation of the SÖS.

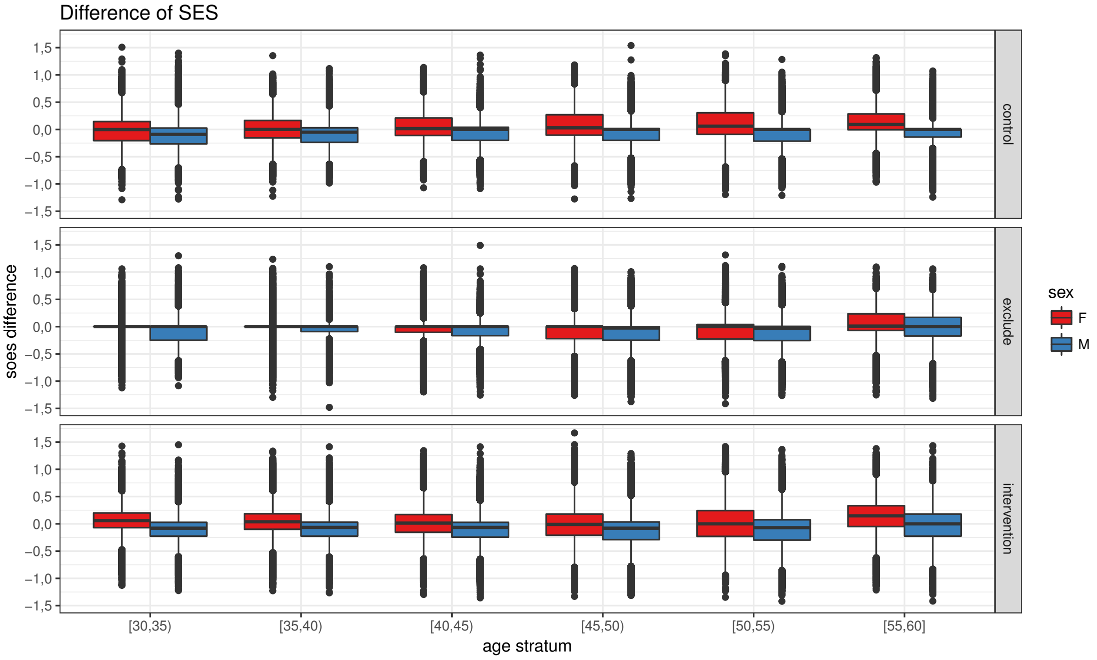


Figure 3.52: Difference between reported and interpolated SES

Figure 3.52 visualizes the difference between the original (variable *soes*) and interpolated (variable *soes_mean*) SES as a boxplot. As with the other SES representations, individuals are grouped by their assigned cohort, age group, and gender.

It can be observed that the total difference of the reported and interpolated SES tightly embraces 0 with a rather small (≤0,5) IQR[25]. Even the outliers stay in a rather narrow region, with most of the extreme values still smaller than ±1. This is most likely a direct result of the comparable small total number of partners per person in these final cohorts, reported in 3.71 on page 121, causing only minor differences.

Although the median difference of about 0 between reported and interpolated SES seems negligible, a trend correlated with gender can be observed. In general, the interpolated SES of women increases, whereas the SES of men tends to decrease. Slight variations in the position of the boxes substantiate this hypothesis. While all boxes cover the value 0 and the medians are almost exactly at 0, the boxes representing men often hang below the 0 line. On the other hand, the boxes representing females appear to be slightly elevated in comparison. This is most likely a result of the fact that females are more often

[25] Interquartile range (IQR) is a robust measure for the spread of a distribution and encloses half of the data including the median.

dependent on males, as shown in Figure 3.1 on page 42 (and following). It cannot be depicted within this analysis whether the special situation of females is blurred incorrectly by considering the SES of their partners, or whether it is appreciated more precisely.

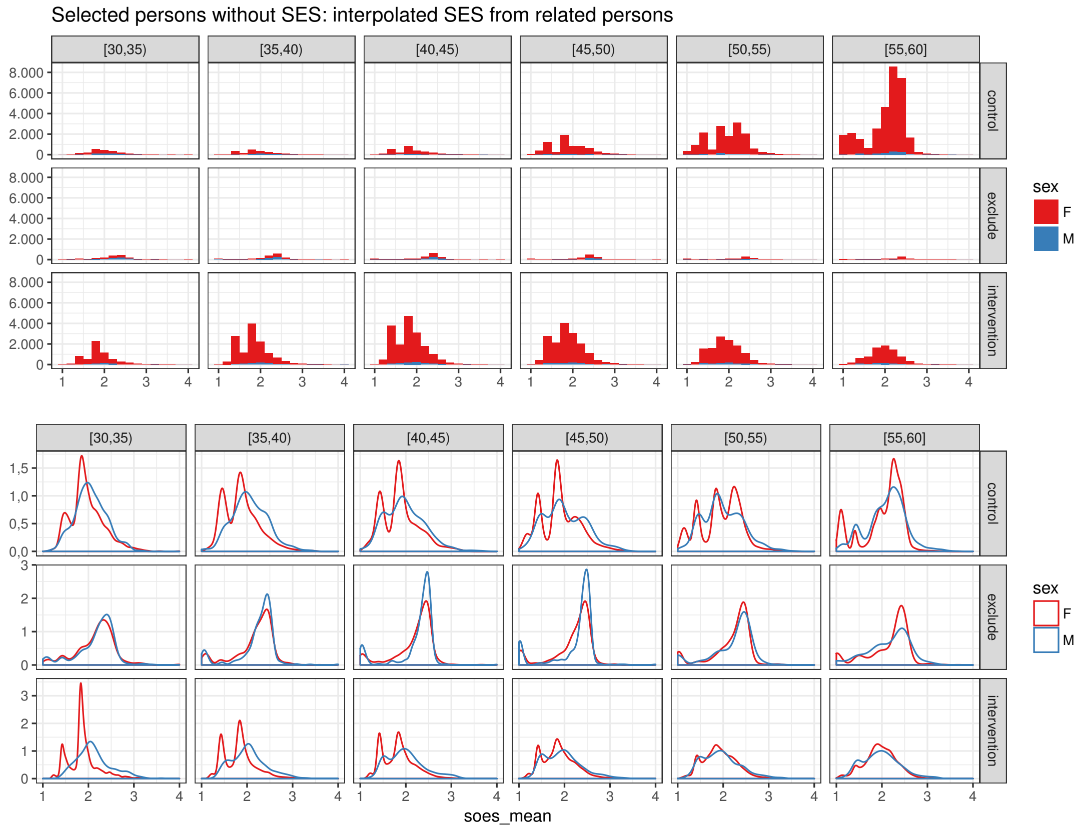


Figure 3.53: Interpolated SES for persons without reported information

Figure 3.53 shows the interpolated SES for all persons without any other information from SÖS, structured according to figures 3.50 and 3.51. It can be clearly seen that the number of people affected is not evenly distributed. It is striking that women are most frequently affected. While most of the cases in the parents (cohort intervention) are between the ages of 35 and 50, almost everyone in the control group is in the highest age bracket. This distribution overstates the overall distribution of individuals. The peaks evident in the density curves result from the comparatively small number of individual cases in each subgroup. Furthermore, in case no personal SES is available, also no statement of facts concerning social insurance[26] is recorded, which pertains to a specific, not randomly selected group of persons.

To eliminate bias due to different states of social and economic well-being, the study

[26] *sozialversicherungsrechtlichen Tatbeständen* in German

protocol suggests equalizing (matching) cohorts according to SES. Therefore, this variable should be included in the analysis.

Figure 3.53 on the preceding page shows that the reported (corrected) SES is not randomly missing in association with age, sex, and assigned cohort. Therefore, some form of imputation is required to compensate for missing values and associated bias. The approach presented summarizes all known SES values from each person's direct relationship network after rounding up as provided in the original data. The difference between the original values and the interpolated values scatters closely around 0, but is not evenly distributed between men and women. It is unclear whether the apparent equalization better describes the particular situation of married couples and families or whether it dilutes the actual differences. Nevertheless, the interpolated SES can be expected to give an accurate view of a person's subjective status by including the social network and, in particular, by compensating for missing information. Since the average SES of a person's social network (of co-insureds) does not describe the same information as the reported SES, the two values cannot be mixed, for example, by using only the calculated information when the reported value is missing.

In summary, socioeconomic status is available as an addendum to the *GAP-DRG* and is added to the extracted cohorts. An error in the original dataset is corrected and missing values are handled by averaging the information in each person's individual relationship network. The resulting variable *soes_mean* is missing for a much smaller proportion of the entire cohort compared to the (corrected) original and is therefore used for further analysis, although open questions remain.

3.3.2 Comorbidity and morbidity score

Three morbidity scores are calculated based on ICD-9 diagnoses from ATC→ICD:

Charlson original Charlson score [Charlson et al., 1987]

Charlson_quan Charlson score is calculated on the basis of the Quan revision of Deyo's ICD-9 mapping, according to [Deyo et al., 1992], [Quan et al., 2011] and [Wasey, 2016]

vanWalraven van Walraven score of the Elixhauser index is calculated from the Quan revision of Elixhauser's ICD-9 mapping according to [Elixhauser et al., 1998], [van Walraven et al., 2009] and [Wasey, 2016]

In case no diagnoses are available, a score of 0 is assumed by default. Thus, there is exactly one of each score per person. Scores are expected to range from 0 to approximately 50, represented as positive (i.e., nonnegative) integers. In the following consideration of the calculated morbidity indices, only the final cohorts used without excluded individuals are included.

First, the pattern of 0s, which are interpreted as missing values, is presented in table 3.22. It can be concluded that a morbidity score (different from 0) is missing in about one third of all individuals, whereas all three of the calculated measures are present in about one in two individuals. The *Charlson_quan* variable is present in every case, and the classic Charlson index is also greater than 0. Overall, the (classic) Charlson index describes the largest proportion of the entire cohort (nearly 62.5%), followed by the *van Walraven* variable with about 55% coverage.

Table 3.22: Combination of missing values (0) for morbidity scores

$\sum$	$\sum$ %	Charlson	vanWalraven	Charlson_quan	missing
791.853	50,20 %	1	1	1	0
28.488	1,81 %	1	1	0	1
55.479	3,52 %	1	0	1	1
45.404	2,88 %	0	1	0	2
108.980	6,91 %	1	0	0	2
547.088	34,69 %	0	0	0	3
$\sum$ NA		592.492 37,6 %	711.547 45,1 %	729.960 46,3 %	2.033.999

Next, the distributions of all three scores are presented, split by 5-year age group, sex, and the assigned cohort in figures 3.54 to 3.57 on page 100. Although a score of exactly 0 has an important interpretation, describing persons without a relevant morbidity, these cases are left out to emphasize persons with a medical condition. Overall, illustrations with and without persons lacking a positive comorbidity measure appear to be very similar. The most apparent differences can be found for persons aged 50 years and above, where a marginally higher 3$^{\text{rd}}$ quartile can be observed.

Figure 3.54 on the next page shows that the classic Charlson index is mostly located below a score of 5. Younger persons tend to have a score between 1 and 2 with an increasing spread towards higher scores with rising age. Especially the highest age group seems to perform differently in both cohorts. Several outliers with rather high scores up to nearly 25 can be observed. These are the effects of multiple medications which are translated to a broad range of ICD-9 codes by ATC→ICD. As these extreme values are a manifold of the median and modal scores, they are not supposed to be integrated into the statistical analysis directly to omit leverage points.

In contrast, Charlson score calculated based on the Quan revision of Deyo's ICD-9 mapping shown in figure 3.55 on the following page gives a more condensed impression. There are fewer extreme outliers and the median values and 3$^{\text{rd}}$ quartiles are more often exactly 1. This might be a direct result of the significantly higher number of persons having a score equal to 0.

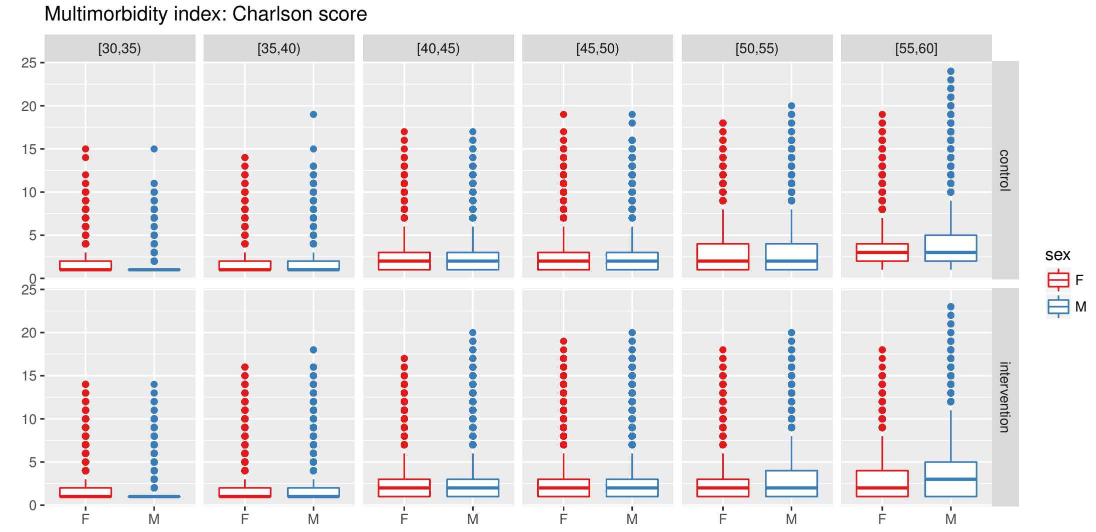


Figure 3.54: Multimorbidity: classic Charlson index by 5-year age groups, sex and cohort

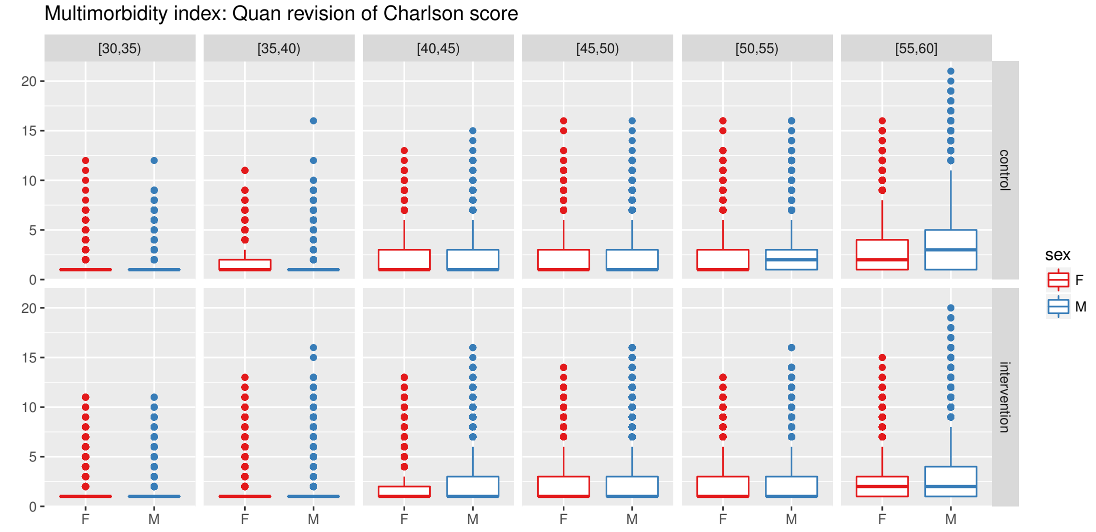


Figure 3.55: Multimorbidity: Quan revision of Charlson index by 5-year age groups, sex and cohort

The difference between both implementations of the Charlson comorbidity index is presented in figure 3.56 on the next page. All persons where the Charlson/Quan index is larger than 0 also have a positive original Charlson index. As a result, only cases where the original index is missing (equals 0) are left out.

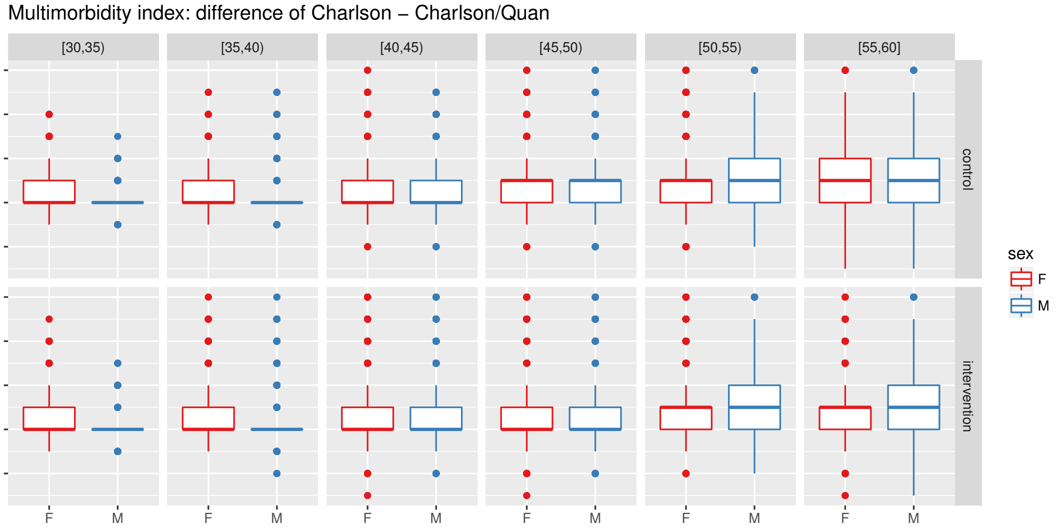


Figure 3.56: Multimorbidity: Difference between two implementations of the Charlson comorbidity index by 5-year age groups, sex, and cohort

Figure 3.56 clearly shows that the difference is mostly located between 0 (no difference) and 1 with a few outliers going up to 6. These differences and their spread grow larger with increasing age. Regarding the different number of zero ratings for both scores, the difference does not seem to be very high, although a score of 1 is often interpreted as drastically different in comparison to 0. Furthermore, the original Charlson index tends to be higher than its newer implementation.

Figure 3.57 on the next page gives a surprising picture of the van Walraven score for the Elixhauser morbidity measure. Despite the seemingly high number of outliers and extreme values above 50, there are also negative values present which are not expected. This issue is also discussed briefly in the original publication of this score:

> Second, our scoring system was derived with, and created for, administrative data. Idiosyncrasies of administrative data — such as those resulting in negative points for some comorbidities — likely influenced the final scoring system. If these idiosyncrasies exist in all administrative systems, our index should—after external validation— be applicable for administrative database research elsewhere. However, a point system derived from primary data is required for studies having primary data collection.
>
> ([van Walraven et al., 2009])

This score has been favored originally by the author due to reports of advantageous performance (e.g., [Sharabiani et al., 2012]) and its more recent publication in 2009.

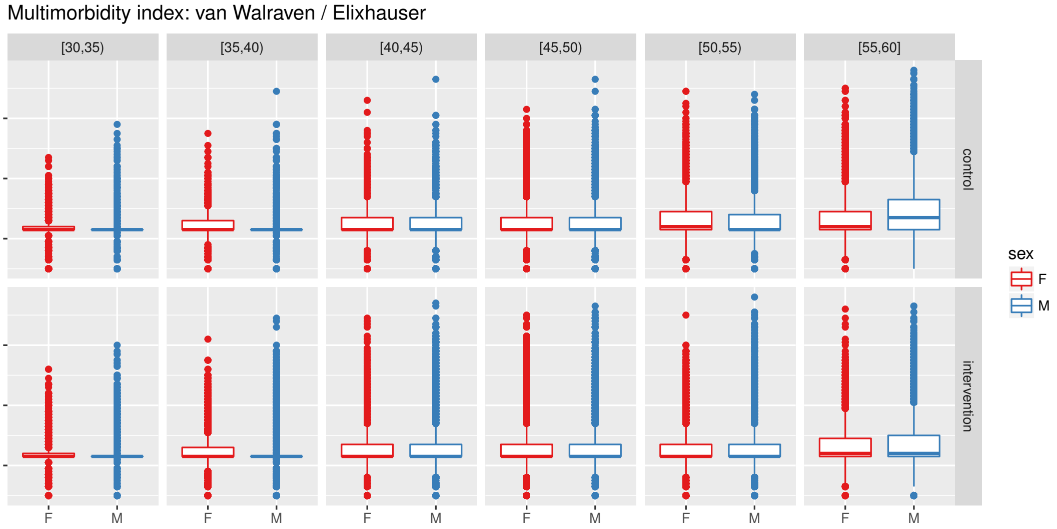


Figure 3.57: Multimorbidity: van Walraven score of the Elixhauser index by 5-year age groups, sex and cohort

Because of these apparent errors and shortcomings most likely caused by the algorithm itself and the inability to test and enhance it in the context of this study, it must be discarded, leaving two implementations of the Charlson index.

Both remaining morbidity indices are rather similar and appear to have differing advantages and drawbacks. While the classic Charlson comorbidity index covers more persons than the revised version, a larger spread and more extreme outliers can be detected. In figure 3.58 on the facing page the distributions of both scores are aligned with each other, split by the assigned cohort, and cut at a score of 15.

Both Charlson scores show a fairly similar distribution with slight differences. In addition to different proportions of zero scores, the Charlson/Quan index also shows significantly larger proportions of individuals with a score of 1. While for the classic Charlson comorbidity index the proportion of individuals decreases uniformly with increasing scores, on the right side of the plot a different, rather discontinuous relationship can be observed, which is mostly unexpected.

Moreover, the number and especially the size of the outliers in both cases is of great concern. To compensate for this, the literature often aggregates the indices into groups, where different, arbitrarily chosen cut points seem to be nothing unusual, e.g., [Johnston et al., 2015] and [Logue et al., 2016]. Exploratory experiments and the interpretation of the indices in the original publications led to three classes according to [Johnston et al., 2015]:

0 no morbidity score

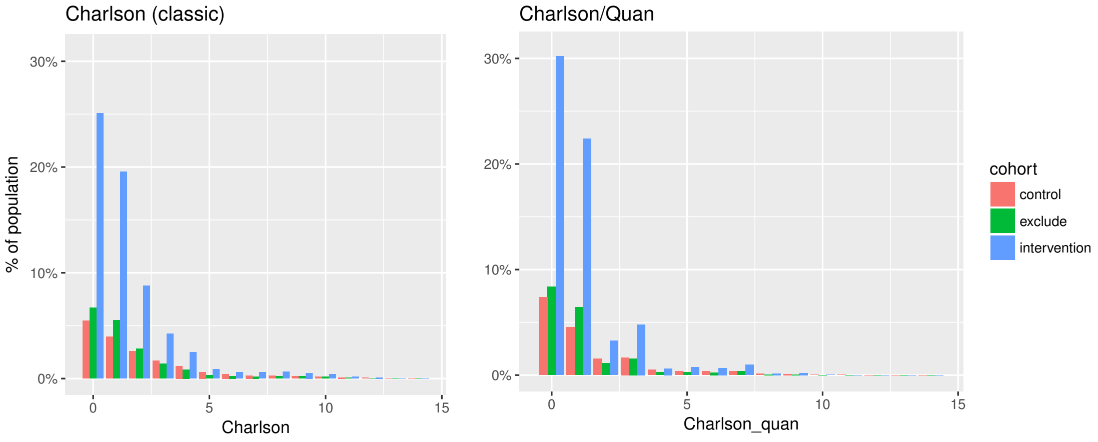


Figure 3.58: Multimorbidity: direct comparison of Charlson scores

1-3 medium occurrence of medical conditions

4+ high to intense appearance of diseases

They are calculated for both implemented Charlson indices. The allocation of these three classes (colors) as a proportion (ordinate) of the entire population split by sex (horizontal split), age groups (vertical split), and cohorts (abscissa) are plotted in figure 3.59 on the next page.[27]

For both scores, a steady increase in the proportion of individuals assigned to higher (worse) morbidity classes is prevalent with increasing age. Most importantly, a shift in the difference between cohorts exists, starting from a slight betterment of the control group for the first two age groups to an inverse relationship for persons older than 50 years. Additionally, this shift appears to be different for each index group. While the 4+ group has a higher proportion in the control cohort of the total population, the 0 group performs as described above.

A distinction can be seen between men and women. Women tend to be found less frequently in the best morbidity class with a score of 0 in younger age groups and more frequently in the other classes. With increasing age, the relative number of individuals in the best morbidity class nearly equalizes between the sexes, and a higher proportion of men are in the 4+ index group. These differences are more pronounced in the morbidity index groups calculated from the classic Charlson index. Moreover, the evolution of these trends appears to be steadier on the left-hand side, showing disjunct for the Charlson/Quan derived classes.[28]

[27] These pairwise bars combined with the SES variable are the subjects of matching in chapter 3.5.

[28] e.g., the growth of the worst class 4+ between the second to last (i.e., *[50,55)*) and highest age group (i.e., *[55,60]*) for Charlson/Quan is disproportional in comparison to the progression in preceding age

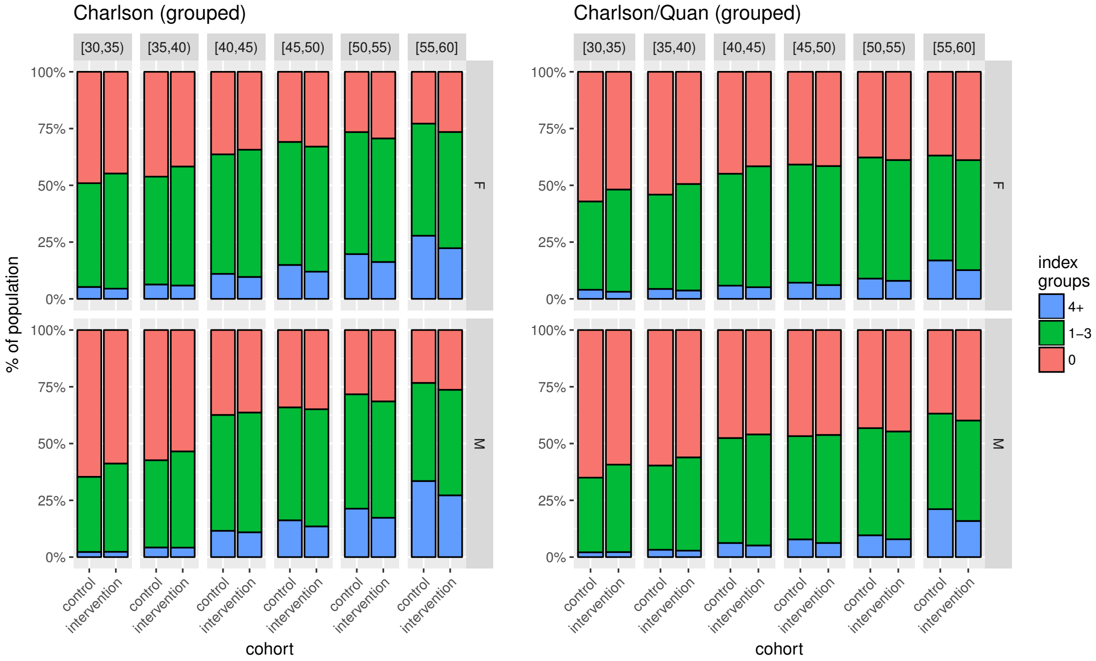


Figure 3.59: Multimorbidity: proportions for classes calculated from Charlson scores

Therefore, the different evolution of the expected overall health status of the compared cohorts is a possible source of bias, although the absolute differences do not seem too serious. Contrary to the author's expectations, the index groups derived from the classic Charlson comorbidity index seem to fit best in terms of overall population coverage and subgroup trends. This finding might apply only to the present application and is not generalizable on the basis of the available evidence.

Finally, a *tabplot* for the proposed morbidity indices and their aggregations is shown in figure 3.60 on the facing page. The correlation of the morbidity indices with each other and with cohort assignment can be clearly seen. Variables such as gender, socioeconomic status, and number of identified partners and children do not appear to be closely related to the Charlson index by which the plot is arranged. It is more likely that the slight similarity is caused by unbalanced cohorts.

In summary, groups representing a morbidity index are derived for the entire study population. Based on ICD-9 diagnoses provided by the ATC→ICD project, three morbidity scores are selected and calculated based on evaluations in the mainstream literature. A closer examination of the results reveals that the classic Charlson comorbidity index, divided into three groups, provides the best fitting variable to approximate health status at the person level. It is therefore used in further analysis.

groups

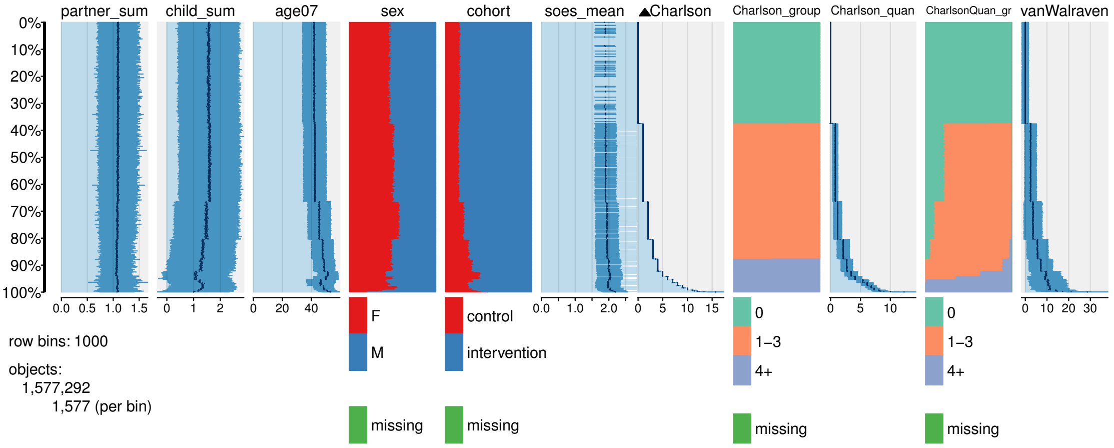


Figure 3.60: Multimorbidity: tabplot of morbidity indices, derived classes, and personal information, arranged by the classic Charlson comorbidity index

3.3.3 Outcome criteria: myocardial infarction

Finally, the outcome event *myocardial infarction* is integrated. According to the study protocol, the following ICD-10 main or additional diagnoses recorded at hospital discharges[29] for the entire dataset from 2006 and 2007 are linked:

I21 ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction[30]

I22 Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction[31]

Altogether, these diagnoses can be found for 3.138 (0,16%) individuals in the entire study population, including *excluded* persons. This appears to be only a minor proportion, but still covers a large number of all cases registered in the GAP-DRG database. There are 9.547 episodes in 2006 and 2007 with one of the defined main diagnoses affecting patients aged between 30 and 60.

The number of episodes per year, ICD-10 diagnosis, type of diagnosis (i.e., main or additional), and age group are listed in table 3.23 for the entire database. It can be clearly seen that I21 is more common than I22. In addition, the number of episodes for the *other* age group, which is almost identical to persons over 60 years of age, is a

[29] Hospital discharge and separation are terms that describe complex administrative processes at the end of a patient's hospital episode. It includes completion of required documentation and reimbursement information. There are several codes that describe the end of an inpatient episode, from normal discharge to various types of inter- and intra-hospital transfers to death. This information is not addressed in the study protocol and therefore is not explored or integrated.

[30] icd10data.com/ICD10CM/Codes/I00-I99/I20-I25/I21

[31] icd10data.com/ICD10CM/Codes/I00-I99/I20-I25/I22

multiple of the group between 30 and 60 years of age. There are approximately three to four episodes with an ICD-10 I21 main diagnosis compared to episodes with the same additional diagnosis. This difference cannot be determined for ICD-10 I22. The total number of episodes is not included in the table because the same episode may be counted twice.

Table 3.23: Total number of episodes with myocardial infarctions in GAP-DRG

year	ICD 10	age group	diagnose	episodes
2006	I21	30-60	main	4.681
2006	I21	30-60	additional	935
2006	I21	other	main	12.420
2006	I21	other	additional	4.657
2006	I22	30-60	main	44
2006	I22	30-60	additional	41
2006	I22	other	main	158
2006	I22	other	additional	165
2007	I21	30-60	main	4.777
2007	I21	30-60	additional	1.047
2007	I21	other	main	12.503
2007	I21	other	additional	4.832
2007	I22	30-60	main	45
2007	I22	30-60	additional	21
2007	I22	other	main	186
2007	I22	other	additional	113

Table 3.24 shows the same information for individual patients instead of inpatient episodes. Again, an individual patient might be counted multiple times. About half of all patients aged between 30 and 60 with one of the defined main diagnoses in 2006 or 2007 are included in the study population. In respect to the defined preconditions concerning relationships, the approximate derivation of this information, and the (unknown) proportion of the entire insured population in a relationship, this share appears to be plausible.

The pattern of missing information for both diagnoses, which equals no recorded indication, is summarized in table 3.25 on the next page for persons in the control or intervention cohort. There are only 28 cases where both conditions are present and 15 persons where an ICD-10 I22 but no I21 main diagnose is recorded. As a result, the indication I22 does not hold much additional information. In the subsequent analysis, both diagnoses are summarized in a single dichotomous variable *mi*.

Tables 3.26 on page 106 to 3.29 on page 106 hold absolute and relative numbers of persons split by the individual observation of a myocardial infarction. All three cohorts,

Table 3.24: Total number of patients with myocardial infarctions in GAP-DRG

year	ICD 10	age group	diagnose	patients
2006	I21	30-60	main	3.040
2006	I21	30-60	additional	769
2006	I21	other	main	8.864
2006	I21	other	additional	3.995
2006	I22	30-60	main	41
2006	I22	30-60	additional	35
2006	I22	other	main	144
2006	I22	other	additional	143
2007	I21	30-60	main	3.085
2007	I21	30-60	additional	797
2007	I21	other	main	9.094
2007	I21	other	additional	4.121
2007	I22	30-60	main	42
2007	I22	30-60	additional	20
2007	I22	other	main	171
2007	I22	other	additional	102

Table 3.25: Combination of missing values (0) for main diagnoses: ICD-10 I21, I22

∑	∑ %	I21	I22	number missing
28	0,00 %	1	1	0
15	0,00 %	0	1	1
2.568	0,16 %	1	0	1
1.574.681	99,83 %	0	0	2
∑ NA		1.574.696 99,8 %	1.577.249 100 %	3.151.945

intervention, control, and exclude, are tabulated to spot potential differences and data quality issues. The cohort *exclude* will be omitted in the following analysis.

Starting with table 3.26, the total number of affected individuals per cohort is summarized. Most cases are registered for the cohort intervention, followed by cohort control. Thus, neglecting the proportions and the different dispersion of the cohorts with respect to covariates such as age and sex, the cohort *interventions* appears to contain the most observed events.

Next, table 3.27 gives the relative number of persons as proportion to the entire study

Table 3.26: Cross table of variables *cohort* and *mi*: absolute number

	mi	no mi	Sum
control	980	331.327	332.307
exclude	527	364.290	364.817
intervention	1.631	1.243.354	1.244.985
Sum	3.138	1.938.971	1.942.109

Table 3.27: Cross table of variables *cohort* and *mi*: proportion of entire population

	mi	no mi	Sum
control	0,05	17,06	17,11
exclude	0,03	18,76	18,78
intervention	0,08	64,02	64,10
Sum	0,16	99,84	100,00

Table 3.28: Cross table of variables *cohort* and *mi*: row %

	mi	no mi	Sum
control	0,29	99,71	100,00
exclude	0,14	99,86	100,00
intervention	0,13	99,87	100,00
Sum	0,16	99,84	100,00

Table 3.29: Cross table of variables *cohort* and *mi*: column %

	mi	no mi	Sum
control	31,2	17,1	17,1
exclude	16,8	18,8	18,8
intervention	52,0	64,1	64,1
Sum	100,0	100,0	100,0

population (with cohort *exclude*) of 1.942.109 individuals. Nearly two-thirds of all persons are assigned to the cohort *intervention*, but only about half of all observed events are in this cohort.

Finally, row- and column percentages are listed in table 3.28 and 3.28. While about 52% of all cases are in cohort *intervention*, the proportion of affected persons per cohort is distributed differently. An event can be observed for 0,29% in cohort *control* but only for 0,14% and 0,13% for the two other groups.

In summary, the absolute and relative figures give a different impression of the actual distribution of those affected. However, relevant covariates such as gender and age are not included. Therefore, no premature conclusion can be drawn yet.

In figure 3.61, the absolute number of individuals with and without a recorded event is shown, split by assigned cohort, age stratum, and gender. Due to the very small proportion of events (i.e., variable *mi* equals TRUE), log-10 scaling is chosen. While this allows the absolute number of individuals with a recorded condition to be seen, it partially obscures the large differences between the two cohorts. In addition to the single bars, bold transparent lines are inserted per group to point out the overall trend with increasing age.

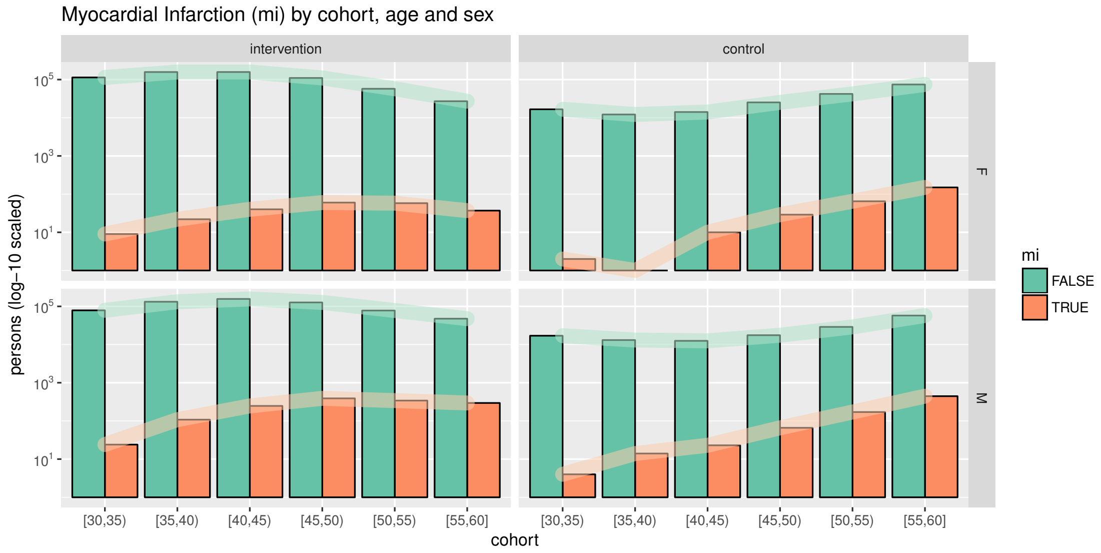


Figure 3.61: **Absolute number** of patients with and without a diagnosed myocardial infarction per cohort, age-stratum, and sex

Even though the comparison of raw absolute numbers might give a false impression of the actual correlations, initial conclusions can be drawn with caution. First, there is a different trend in the individual cohorts. While the number of individuals increases only slightly up to the age between 40 and 45 and decreases thereafter for the cohort *intervention*, the opposite trend can be found on the right-hand side of the graph. In addition, there is a steady, exponential increase in the number of affected patients, especially among men in the control group, which grows faster than the overall subpopulation. Moreover, the very small number of younger women (first two age strata) and young men (first age stratum) in the control cohort with a recorded myocardial infarction are especially noteworthy.

In figure 3.62 on the next page, the relative number of affected patients is shown as a proportion of the respective population. Due to the small proportion of patients with an observed event and its complementary nature, the proportion of individuals without one of the selected diagnoses is not included. Therefore, a more direct comparison of the cohorts per age and sex is appropriate.

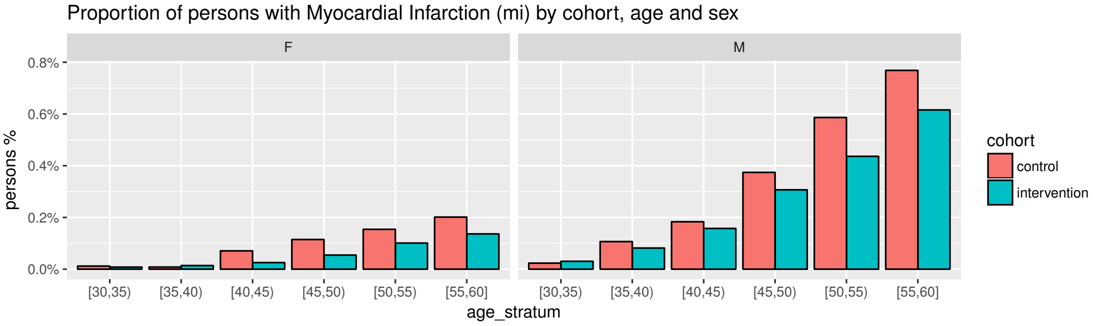


Figure 3.62: **Relative number** of patients with a diagnosed myocardial infarction per cohort, sex and age-stratum

At first sight, figure 3.62 reveals a large difference between males and females as well as growing rates of affected persons with increasing age. Solely in the first age stratum for males and the second one for females, persons associated with the intervention cohort appear to be more likely to suffer from a myocardial infarction. As documented in figure 3.61, there are generally very few cases in the first two age strata, especially for the control group. In all other subsets, the proportion of diseased parents is smaller than for couples without children. Although there are generally fewer females affected, the difference between the cohorts appears larger in the left column of the plot.

This representation contradicts the initial hypothesis in the study protocol and even shows a converse trend. Nevertheless, further covariates must be included and a more profound analysis is required.

Paradigmatically, the same information split by the grouped Charlson comorbidity index, which has been introduced in chapter 3.3.2 on page 96, is presented in figure 3.63. Two main conclusions can be drawn.

First, there appears to be a very strong correlation between the (grouped) Charlson index and the observed events. While the proportion of persons with a comorbidity score of 0 experience a myocardial infarction is mostly 0, up to 1,5% are affected by the oldest subset with high morbidity scores.

Second, although individuals in the control group still show a higher share of relevant events, the differences are unevenly distributed. Especially males aged between 55 and 60 with a high Charlson index appear to have nearly the same rate in both cohorts.

In summary, information on myocardial infarction from inpatient episodes in 2006 and 2007 with a defined main diagnosis at discharge is obtained. The specified ICD-10 diagnoses are evaluated separately and the coverage of the final cohorts are compared

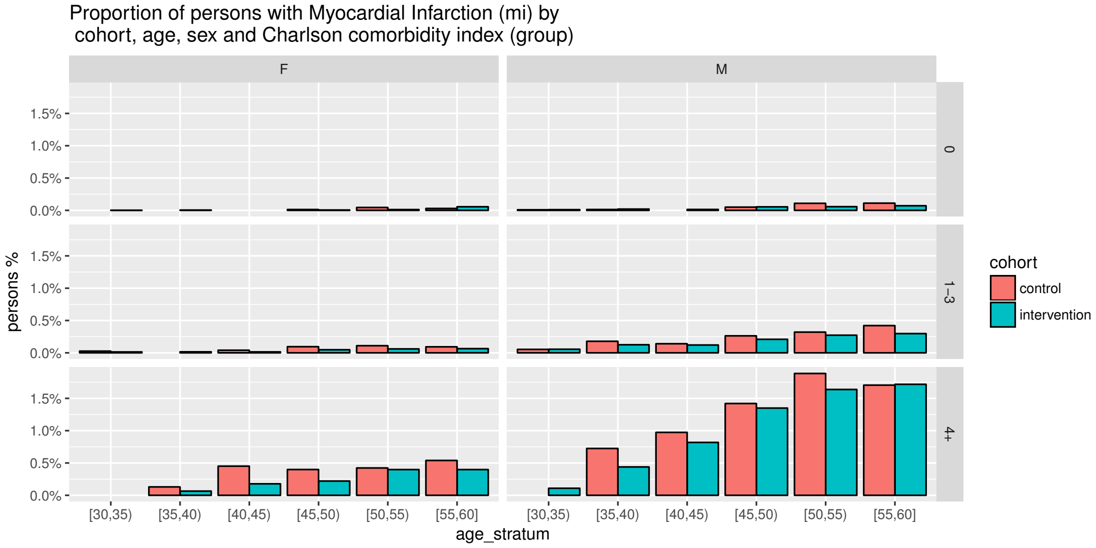


Figure 3.63: **Relative number** of patients with a diagnosed myocardial infarction per cohort, sex, age-stratum and Charlson index groups

with all cases available in the data source. It can be concluded that approximately one out of every two persons aged 30 to 60 years experiencing a myocardial infarction is covered in this study. In addition, one of the two defined ICD-10 diagnoses contains almost all relevant information because of the large difference in frequencies. As a result, a summary variable *mi* is introduced.

Basic explanatory data analysis yielded mixed impressions. On the one hand, the raw frequencies seemed to vaguely support the study's hypothesis. On the other hand, the analysis of proportions in subgroups in relation to additional covariates ended in an opposite preliminary conclusion.

3.3.4 Final variables

The following variables are available in both resulting datasets:

pers_id unique, project-dependent identifier of the person

birthyear year of birth and sex

sex binary biological gender as defined by the social security institutions, coded as *F* for female and *M* for male

age07 age in the year 2007

age07_10th variable *age07* divided by 10 to ease interpretation in e.g., logistic regression models

partner_sum the number of partners within the personal network

partner whether a partner is associated with this person, logical opposite of cohort *exclude*

child_sum the number of children with a maximum distance of two

soes, soes_mean socio economic status, described in chapter 3.3.1 on page 91

intervention, control, exclude affiliation with the corresponding cohort as Boolean (true, false)

cohort associated cohort as text (*intervention*, *control*, *exclude*)

age_stratum 5-years age group as suggested by the study protocol.

- Starting with 30, each group covers 5 years. i.e., 30-34, 35-39, ...
- These groups are also labelled as intervals according to ISO 31-11. i.e., [30-35), [35-40)
- The last group covers 6 years (55 till 60 inclusive, [55-60]).

I21, I22 diagnosis of myocardial infarction, described in chapter 3.3.3 on page 103

mi whether a myocardial infarction has occurred during the observation period (*boolean: TRUE / FALSE)*

Charlson, Charlson_quan, vanWalraven morbidity score, described in chapter 3.3.2 on page 96

Charlson_group acquired grouped comorbidity estimation

3.4 Data quality assessment

This subsection is intended to present the extracted datasets and cohorts by documenting univariate distributions, content, and selected multivariate relationships. Further elaboration on additional content can be found in specific subsections.

Based upon the presented univariate data profiles, the selected multivariate analysis are discussed in the following subsection.

3.4.1 Gender and age

Crosstabulation for gender and age strata of 5 years each[32] are presented. Absolute numbers for both cohorts *full* and *no-NA* are introduced in tables 3.30 and 3.31. Differences between these cohorts as absolute numbers and relative to the cohort *full* are presented in tables 3.32 and 3.33. Finally, row, column and absolute percentage are listed for the cohort *no-NA* in tables 3.34 to 3.36 on page 113.

The absolute numbers of persons in each cohort appear equally distributed at first glance. As expected, the reduced cohort *no-NA*, which contains only persons with no missing information, is smaller than the full cohort in each subgroup. In both tables, the 40-44 age stratum is the largest. There are more women than men up to age 50, with inverse proportions in the older groups. Overall, there are significantly more women than men.

Table 3.30: Crosstabulation of sex and age (5-year groups) for cohort *full*

	[30,35)	[35,40)	[40,45)	[45,50)	[50,55)	[55,60]	Sum
F	180.150	232.593	234.399	182.023	128.588	119.462	1.077.215
M	115.006	166.498	197.435	178.852	137.082	131.619	926.492
Sum	295.156	399.091	431.834	360.875	265.670	251.081	2.003.707

Table 3.31: Crosstabulation of sex and age (5-year groups) for cohort *no-NA*

	[30,35)	[35,40)	[40,45)	[45,50)	[50,55)	[55,60]	Sum
F	170.910	222.852	227.440	178.579	127.250	118.202	1.045.233
M	108.291	158.416	190.476	174.291	135.163	130.239	896.876
Sum	279.201	381.268	417.916	352.870	262.413	248.441	1.942.109

The absolute and relative differences are listed in tables 3.32 and 3.33.

While the largest absolute differences are observed for age groups in the middle of the tables, there are relatively more people excluded due to lack of information for younger groups. There is also a difference between men and women that correlates with age. More men are excluded in younger cohorts. This gender difference only increases with age and decreases for the oldest groups.

In tables 3.34 to 3.36, row, column and absolute percentages for the cohort *no-NA* are listed. Previously described varieties between genders and age groups can be observed in more detail. There appears to be a large difference between females and males for younger age groups, and there are more than 60% females in the youngest age group, located between 30 to 34 years, giving a relative difference of 23% . Moreover, there

[32] despite the highest age group ranging from 55 years to 60 years

Table 3.32: Crosstabulation of difference between cohorts for sex and age (5-year groups)

	[30,35)	[35,40)	[40,45)	[45,50)	[50,55)	[55,60]	Sum
F	9.240	9.741	6.959	3.444	1.338	1.260	31.982
M	6.715	8.082	6.959	4.561	1.919	1.380	29.616
Sum	15.955	17.823	13.918	8.005	3.257	2.640	61.598

Table 3.33: Crosstabulation of difference between cohorts for sex and age (5 year groups): % relative to cohort *full*

	[30,35)	[35,40)	[40,45)	[45,50)	[50,55)	[55,60]	Sum
F	5,13	4,19	2,97	1,89	1,04	1,05	2,97
M	5,84	4,85	3,52	2,55	1,40	1,05	3,20
Sum	5,41	4,47	3,22	2,22	1,23	1,05	3,07

is also a difference of about 17% and 14% percentage points in the subsequent groups between females and males.

Table 3.34: Crosstabulation of sex and age (5-year groups) for cohort *no-NA*: row percentages

	[30,35)	[35,40)	[40,45)	[45,50)	[50,55)	[55,60]	Sum
F	16,4	21,3	21,8	17,1	12,2	11,3	100,0
M	12,1	17,7	21,2	19,4	15,1	14,5	100,0
Sum	14,4	19,6	21,5	18,2	13,5	12,8	100,0

Table 3.35: Crosstabulation of sex and age (5-year groups) for cohort *no-NA*: column percentages

	[30,35)	[35,40)	[40,45)	[45,50)	[50,55)	[55,60]	Sum
F	61,2	58,5	54,4	50,6	48,5	47,6	53,8
M	38,8	41,5	45,6	49,4	51,5	52,4	46,2
Sum	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Figure 3.64 on the next page visualized the absolute number of persons per age stratum, gender, and cohort assignment for the dataset *no-NA*. In contrast to the preceding tables, the entire study population is split into two defined cohorts and the group of excluded persons. While the corresponding bars are aligned one underneath the other, the ordinate is not scaled equally. As a result, the absolute height of bars cannot be compared directly.

Table 3.36: Crosstabulation of sex and age (5-year groups) for cohort *no-NA*: total percentages

	[30,35)	[35,40)	[40,45)	[45,50)	[50,55)	[55,60]	Sum
F	8,80	11,47	11,71	9,20	6,55	6,09	53,82
M	5,58	8,16	9,81	8,97	6,96	6,71	46,18
Sum	14,38	19,63	21,52	18,17	13,51	12,79	100,00

In general, most of the results extracted from the following tables in this chapter appear to be dominated by the largest cohort intervention. Thus, the distribution of this cohort closely resembles the distribution of the overall study population. Different patterns are observed for the remaining cohorts.

The absolute number of individuals per cohort varies widely. While there are less than 20.000 men and women for the first three age groups in the control cohort, the intervention cohort has more than 100.000 individuals in (almost) every aligned subgroup.

Furthermore, a trend of the ratios of women and men can be observed. While they correspond to the previous description for the cohort intervention, different conclusions can be drawn for the other cohorts. For the cohort control, women are the largest group in almost all age strata. This difference increases with age. Cohort *exclude*, in which parents without a suitable partner are found, consists predominantly of women. More than a quarter of all younger mothers are excluded from the cohort intervention. For older age groups, more men are excluded than women. Overall, more individuals were excluded than assigned to the control group. This ratio is strong among the youngest groups, where most identified parents are also found.

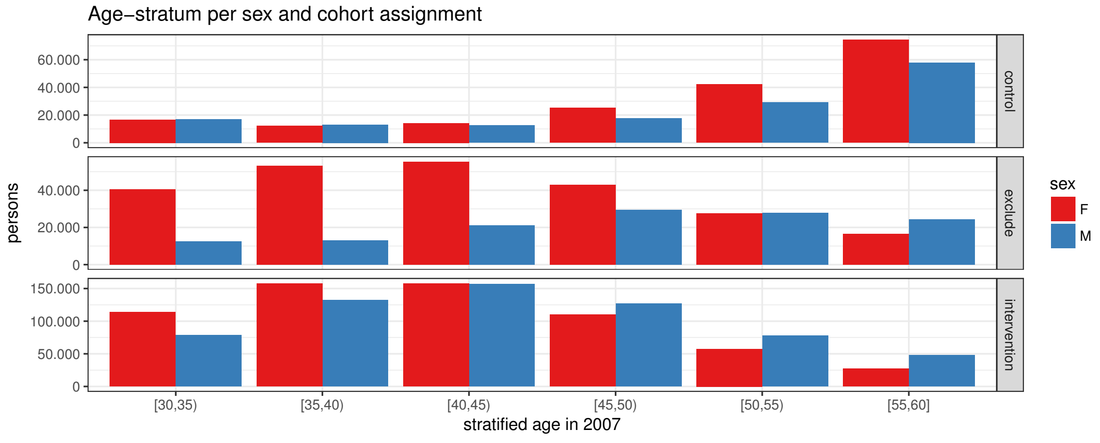


Figure 3.64: Age-stratum per sex and cohort assignment

Figure 3.65 shows the number of persons per age in (single) years per gender.

The range between ages 40 and 50 is zoomed in to highlight the tipping point between a higher number of women compared to men. The tipping point is between ages 46 and 47 for the entire study population. The excess of women compared with men up to age 45 is a result of the allocation in cohort *intervention* and cohort *exclusion*, but contradicts the situation in cohort *control*.

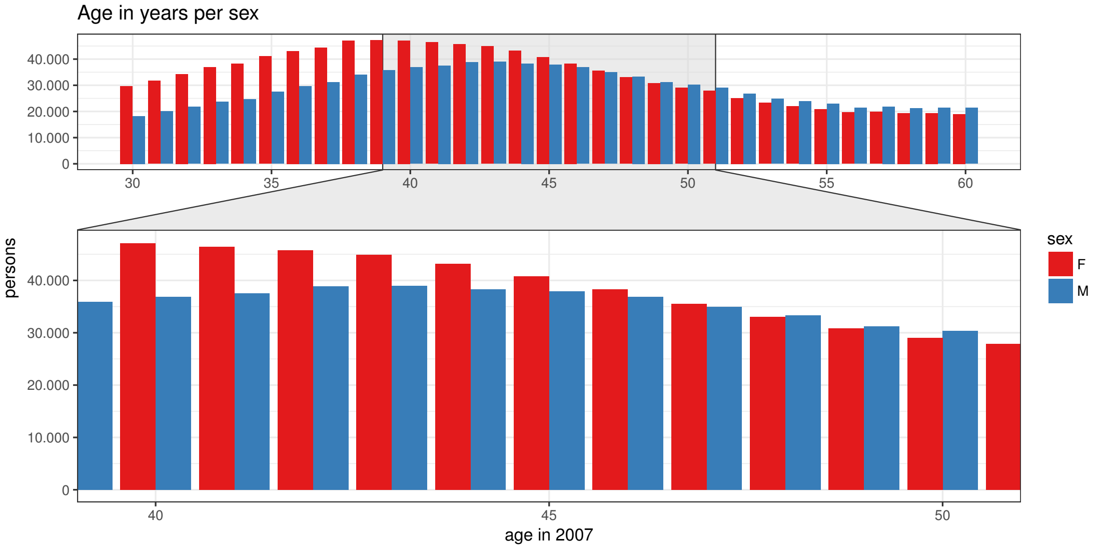


Figure 3.65: Age in years by gender with emphasized region

Another change in trend is highlighted in figure 3.66.

The graph shows the absolute number of persons per age in years, colored by cohort affiliation. The overall distribution of persons per cohort is shown in the top plot. The absolute excess of cohort *control* for most of the entire age spectrum is shown. The cohort *control* begins at a comparatively low level and appears to decline into the 40-45 age range. In contrast, the excluded groups begin with a slightly higher number of individuals and show a relative trend similar to the cohort *intervention*.

In the second lower plot, the region at age 45 is highlighted. Two important changes in the trend can be located. First, at age 51, the number of individuals in the cohort *control* overtakes the excluded group. Their upswing gains pace, while the number of persons in the cohort *intervention* steadily declines. The second major breaking point is at age 56, where the cohort *control* finally overtakes the cohort *intervention* and becomes the dominant group.

Finally, the relation of males and females per age in years and cohort assignment is presented in figure 3.67. The overall development with increasing age of the individual

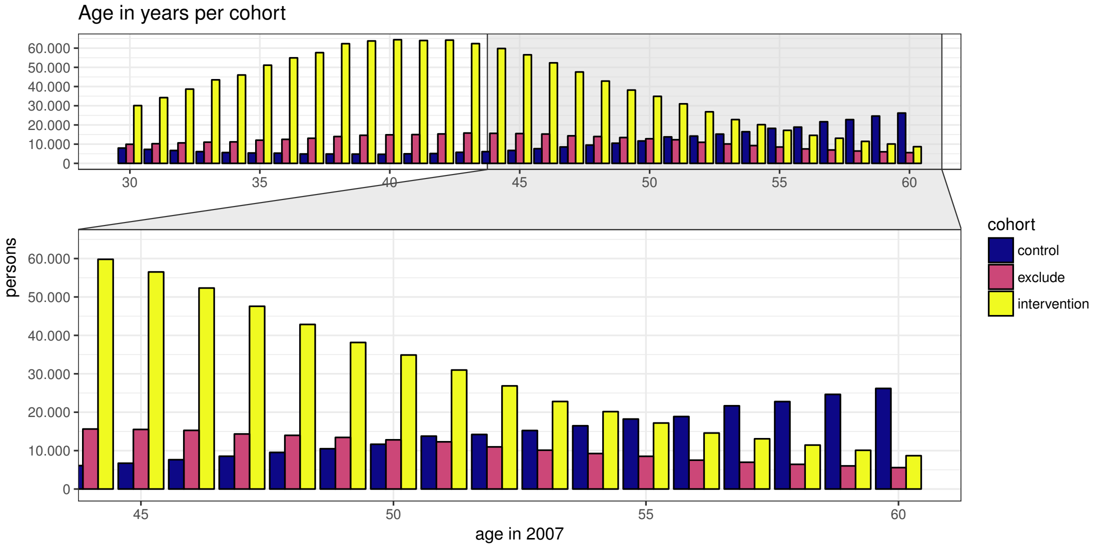


Figure 3.66: Age in years by cohort with emphasized region

cohorts generally does not change. Nevertheless, there are clear differences regarding gender.

In the cohort *intervention* and excluded individuals, there are more females than males in the younger age groups. The similarity between these two groups is most likely determined by the fact that both contain parents. There is a striking difference between women and men for excluded individuals up to about age 50. The shift in sex for the intervention and control cohorts could also be a source of bias. For example, if it can be assumed that older men are more likely to have myocardial infarction, the shift in sex ratio could give a false impression when absolute numbers are compared. In addition, the individuals in the excluded group might not have been randomly selected.

3.4.2 Child and children

The number of children per identified parent is discussed in this section. This aspect is compared for both data sets, *full* and *No-NA*.

The following figures include both cohorts studied, i.e., intervention and control, without excluded individuals. Each figure is divided into two sections. On the left side, the absolute numbers of children per person are shown as a stacked bar chart, subdivided by gender and limited to a maximum of 9 children. Since this plot is dominated by individuals with few children and the entire range is hidden, a second plot is presented on the right. It contains essentially the same information without distinction by gender, including a log-scaled ordinate and individuals with up to 22 associated children.

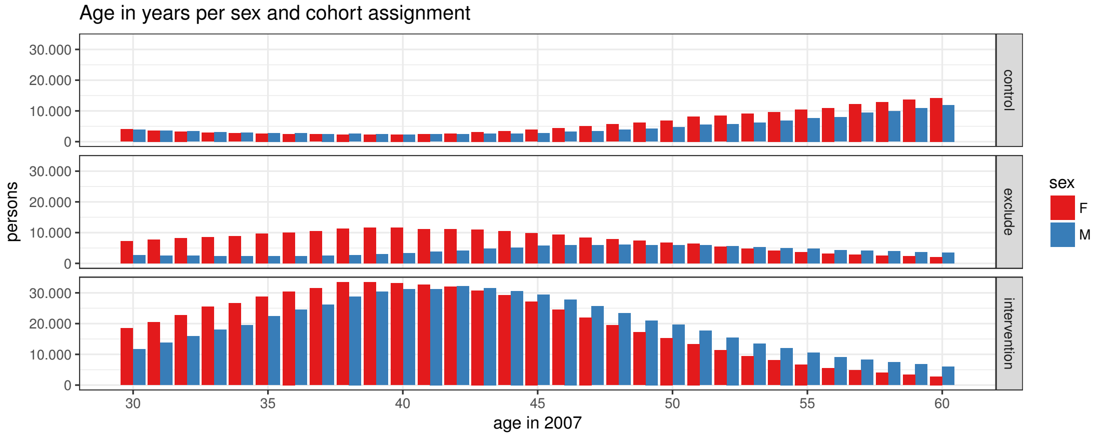


Figure 3.67: Age in years by gender and cohort

Cohort assignment can be indirectly determined in these graphs. All persons without any children (i.e., the number of children equals 0) are part of cohort *control*.

In figure 3.68 on the next page, the complete data set is presented without restriction due to data quality considerations (i.e., completeness of information). The left plot clearly shows that in most cases, less than 3 children per person can be identified. Parents with two children form the largest group, followed closely by one child, (no children), and three children. There is a clear break between two and three, and three and four children. Overall, the distribution seems probable.

On the right side is the long tail of parents associated with a higher number of children. 10,000 individuals have between five and six and about 1,000 individuals are associated with eight to nine children. Although there are about 100 parents with 12 children and some individuals with an even larger number of offspring, there are not many individuals with uncommonly sized families.

Figure 3.69 on the facing page holds the same information for the restricted dataset *no-NA*.

While the overall outlines of the recognizable distributions are very similar to the previous figure, some differences can be observed. The long tail, which concerns parents with a fairly high number of children, is much shorter. In addition, the number of individuals with no or one child is almost the same for both data sets, while it is much smaller for individuals with two or more children for the clean version.

The group of excluded people shows a different distribution in terms of the number of children. Most parents are connected to only one child, while higher numbers are very rare. This could be a consequence of a lack of indirect connections to children through an adult partner or a bias introduced by the different distribution of this group in terms of age and gender. The exact reason for this disparity cannot be determined and is not

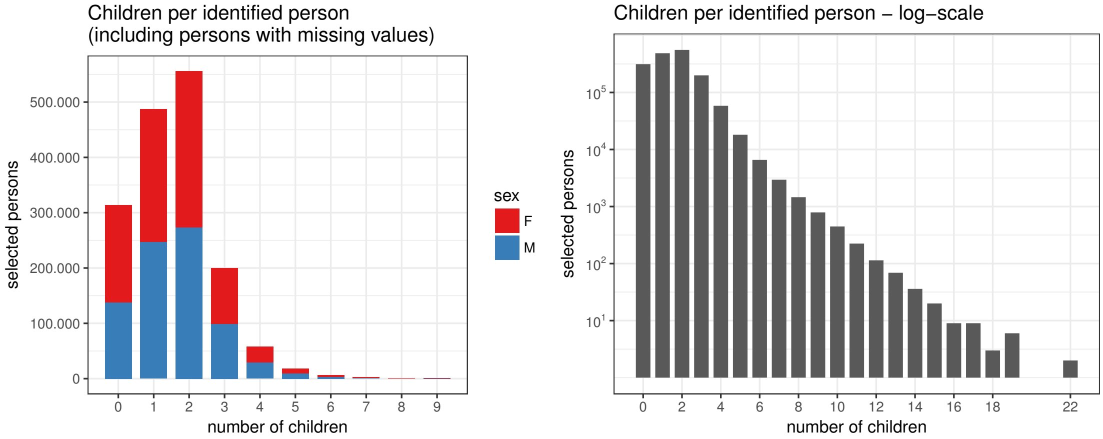


Figure 3.68: Number of children per selected person: cohort from networks with missing gender or age

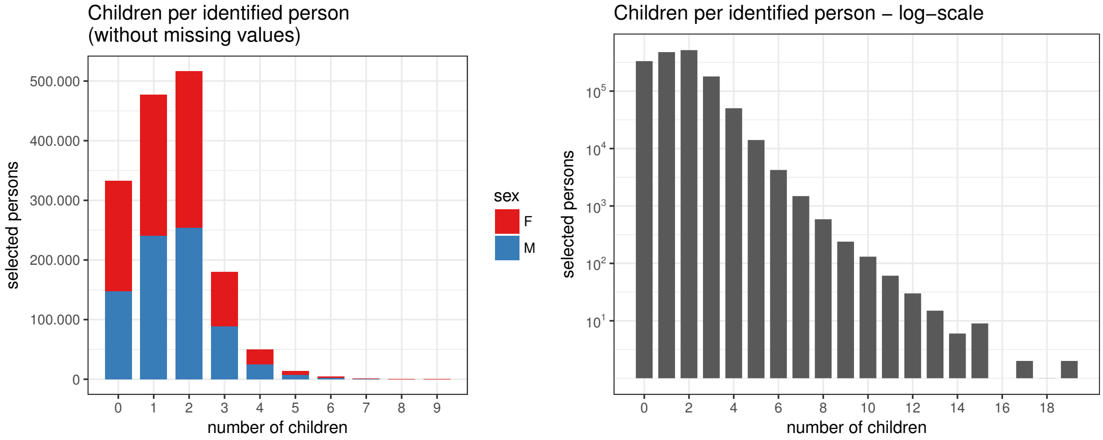


Figure 3.69: Number of children per selected person: cohort from networks without missing gender or age

investigated further, although it could contribute as an important source of bias in the study.

Overall, the number of children per parent can be considered promising. Despite long tails consisting of a few parents with a rather high number of children, the clear majority of individuals meet the author's expectations. Furthermore, the decrease in individuals with more than one child and the general reduction of the long tail in the *no-NA* dataset can be interpreted as an increase in data quality.

Excursus: family/children statistics

Various statistics and data sets are provided by Statistics Austria. Two excerpts are presented here to provide a brief reality check of the results obtained. They are taken from the article *Families*, which can be found in the section *Population Censuses, Register-based Census, Register-based Labor Market Statistics* on the official homepage.

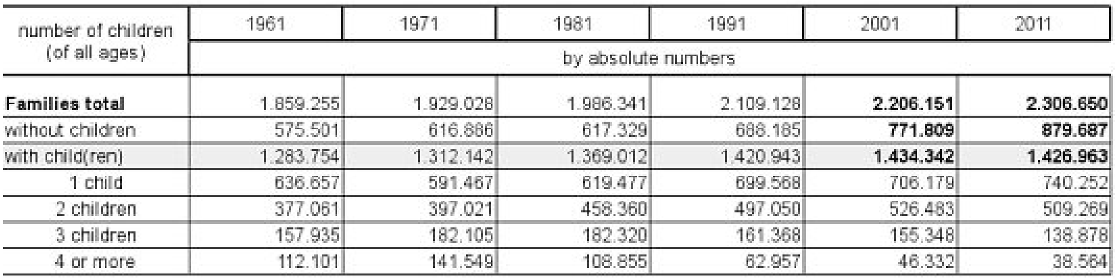
The utilized source is defined as follows:

> STATISTICS AUSTRIA, Population Censuses 1961 to 2001, Register-based Census 2011. Compiled on 4 November 2013. A family nucleus comprises of couples (married or cohabiting) with or without children or lone parents with one or more children living in the household. Children are all biological children, step children, and adopted children, who do not have an own partner and have no child of their own while still living in the household of their parents are regarded as children - without considering their age or marital status. Up to 1991, grown up sons and daughters were considered as children only when they were unmarried. 1) 1961: with children under 14 years.
> (statistik.at: People & Society - Population - Families, last visited 2020-07-16)

Table 3.37 lists the number of families with and without children for several years. About 2.300.000 families in total and about 1.400.000 families with children have been identified by Statistics Austria. In contrast, roughly 1.330.000 parents and 310.000 spouses without children are included in the *full* dataset from GAP-DRG.

However, a direct comparison is not feasible because the definitions of child and family are very different. These numbers would suggest that about every second parent was perceived, but only every fifth to sixth family without children was identified. These figures become worse for the cleaned data set *no-NA*.

Table 3.37: Statistics Austria: families and children per family



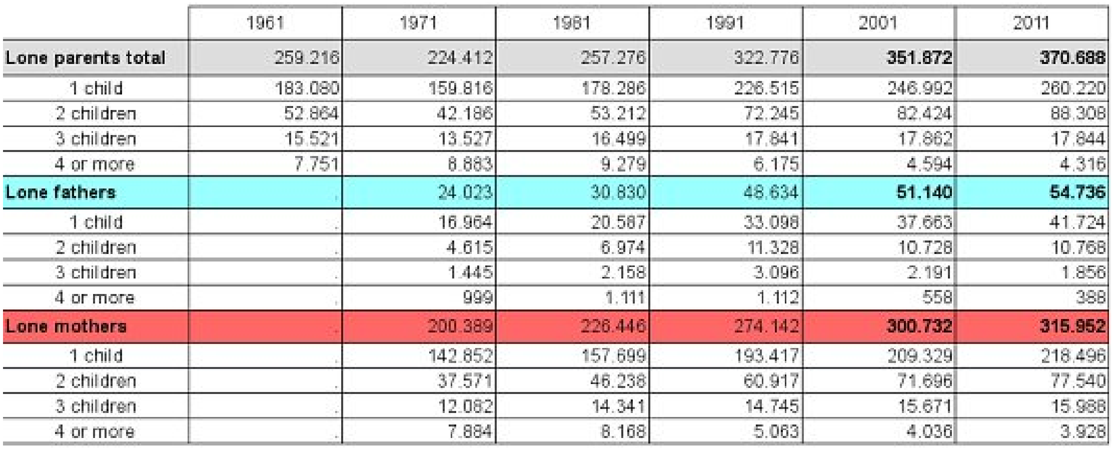
number of children (of all ages)	1961	1971	1981	1991	2001	2011
	by absolute numbers					
Families total	1.859.255	1.929.028	1.986.341	2.109.128	**2.206.151**	**2.306.650**
without children	575.501	616.886	617.329	688.185	**771.809**	**879.687**
with child(ren)	1.283.754	1.312.142	1.369.012	1.420.943	**1.434.342**	**1.426.963**
1 child	636.657	591.467	619.477	699.568	706.179	740.252
2 children	377.061	397.021	458.360	497.050	526.483	509.269
3 children	157.935	182.105	182.320	161.368	155.348	138.878
4 or more	112.101	141.549	108.855	62.957	46.332	38.564

Table 3.38 on the facing page shows the number of lone parents split by gender.

In total, there were between 350.000 and 370.000 single parents in Austria between 2001 and 2011. In this context, single mothers outnumber single fathers by a factor of almost six. In contrast, there are about 317.000 excluded parents corresponding to single parents in the data set *full*, but there are only about twice as many single mothers as fathers.

Since the *exclude* group in the extracted data set contains not only single parents but also persons for whom the partner could not be identified, a comparison of these proportions is not valid at all and therefore cannot be interpreted.

Table 3.38: Statistics Austria: lone parents and children per parent



	1961	1971	1981	1991	2001	2011
Lone parents total	259.216	224.412	257.276	322.776	**351.872**	**370.688**
1 child	183.080	159.816	178.286	226.515	246.992	260.220
2 children	52.864	42.186	53.212	72.245	82.424	88.308
3 children	15.521	13.527	16.499	17.841	17.862	17.844
4 or more	7.751	8.883	9.279	6.175	4.594	4.316
Lone fathers	.	24.023	30.830	48.634	**51.140**	**54.736**
1 child	.	16.964	20.587	33.098	37.663	41.724
2 children	.	4.615	6.974	11.328	10.728	10.768
3 children	.	1.445	2.158	3.096	2.191	1.856
4 or more	.	999	1.111	1.112	558	388
Lone mothers	.	200.389	226.446	274.142	**300.732**	**315.952**
1 child	.	142.852	157.699	193.417	209.329	218.496
2 children	.	37.571	46.238	60.917	71.696	77.540
3 children	.	12.082	14.341	14.745	15.671	15.988
4 or more	.	7.884	8.168	5.063	4.036	3.928

Although the absolute numbers are difficult to compare due to their different origins and definitions, the proportions of individuals identified indicate the completeness of the results obtained. It can be concluded that the control group is underrepresented and biased compared to the cohort intervention. This is a direct consequence of the extraction process and the available data and cannot be corrected. Quantifying these differences could allow adjustment of the available data set through resampling, matching, imputation, or weighting of observations in statistical models.

However, the extracted cohorts and the officially published population numbers are not further compared in this project. Nevertheless, this rough comparison is an important result and can be used for the final interpretation of the result. On the one hand, the breakdown of proportions (based on only roughly similar groups) indicates an underrepresentation of the control cohort, but on the other hand it also shows that a relevant proportion of the target population is covered.

3.4.3 Partners

The number of partners per identified person is summarized in this section. It is elaborated on for both datasets *full* and *no-NA*. Both figures are constructed according to the description in the previous section dealing with the number of children per adult.

The absolute number of partners meeting the selection criteria per person, split by gender, are shown in figure 3.70 on the next page. It can be concluded that the clear majority of individuals have exactly one partner, followed by individuals (from the excluded cohort) without any associated adult. Individuals with two or more partners are rare in comparison.

The large drop in individuals with more than one associated adult can be examined on the left side of the plot, e.g., there are fewer than 10.000 individuals with 4 co-insured partners.

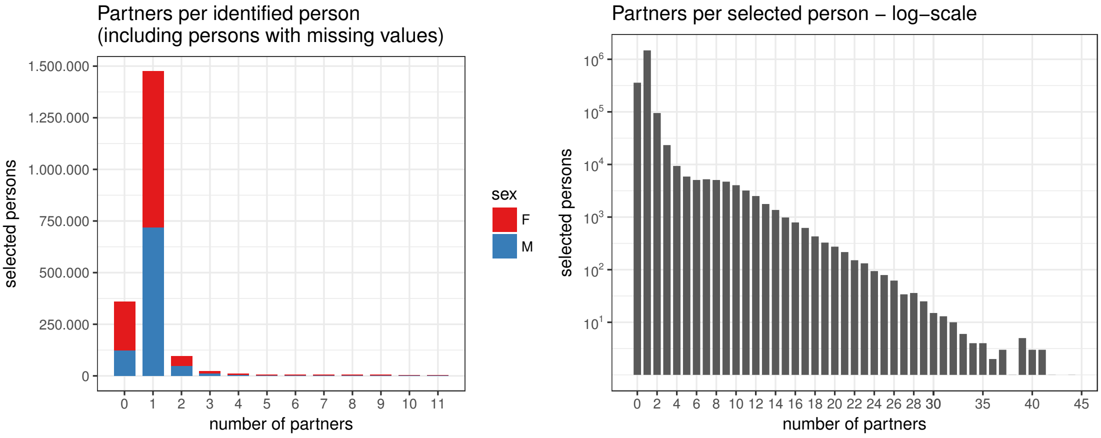


Figure 3.70: Number of partners per selected person:
cohort from networks with missing gender or age

In table 3.39 on the facing page, the absolute and relative number of partners per individual, ordered by the number of affected individuals, are listed for the 10 most frequent cases. Column "1 - % cumulative" contains the reciprocal of the cumulative percentage of the total study population, i.e., the ratio of the population at the bottom of the table. It can be noted that less than 3% of all individuals have more than two associated partners.

The same information is presented for the *no-NA* data set in figure 3.71. Overall, a very similar conclusion can be drawn from the plot on the left. There are slightly more individuals who do not have a partner and are therefore marked as excluded. There is also a slight decrease in the number of individuals with 2 or more associated adults.

The log-scaled plot shows that there are significantly fewer individuals with more than two partners. The maximum number of associated adults is also significantly lower in comparison.

More details can be found in table 3.40, analogous to table 3.39. The entire table holds 15 entries with a maximum of 16 partners per individual in a single case. Altogether, only about 1% of all persons have more than two partners.

Finally, the coherence between the total number of partners and children per person in the reduced study population *no-NA* is analyzed in figure 3.72. Each of the 36 facets is

Table 3.39: Number of individuals per total number of associated adult partners (truncated)

partner_sum	count partners	relative %	1 - % cumulative
1	1.474.433	73,59	26,41
0	358.315	17,88	8,53
2	95.076	4,75	3,78
3	23.346	1,17	2,61
4	9.371	0,47	2,14
5	5.886	0,29	1,85
7	5.235	0,26	1,59
6	5.070	0,25	1,34
8	5.063	0,25	1,09
9	4.709	0,24	0,85

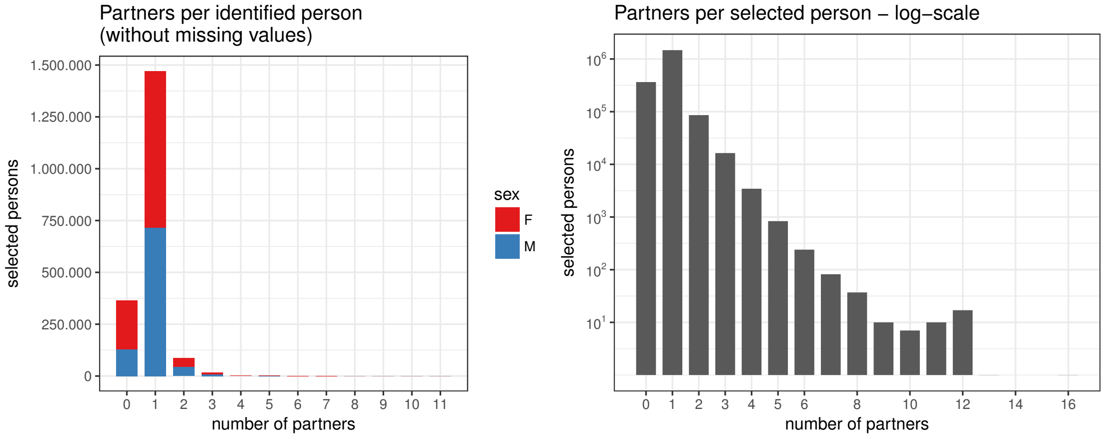


Figure 3.71: Number of partners per selected person:
cohort from networks without missing gender or age

limited to a maximum of 5 partners and 5 children to focus on the relevant content. The entire range of the distribution and potential outliers have already been presented above.

Cohorts *control* and excluded individuals extend only along one axis because children or partners are not present by design. With the cohort *intervention*, any combinations are possible despite individuals without a partner or child.

It can be concluded that all facets appear plausible. Most individuals have a single partner and individuals with more than three children are not widespread. An increase in observations with a partner in the cohort *control* with increasing age can be explained by the overall distribution itself.

Table 3.40: Number of individuals per total number of associated adult partners (truncated)

partner_sum	count partners	relative %	1 - % cumulative
1	1.470.292	75,71	24,29
0	364.817	18,78	5,51
2	85.956	4,43	1,08
3	16.359	0,84	0,24
4	3.445	0,18	0,06
5	835	0,04	0,02
6	240	0,01	0,01
7	82	0,00	0,01
8	37	0,00	0,01
12	17	0,00	0,01

For the cohort *intervention*, there is an early increase followed by a decrease in those with children. In a direct comparison, men associated with children are in a higher age stratum than women. Since the age of parents in 2007 is used, no direct statement can be made about the age of (first) parenthood.

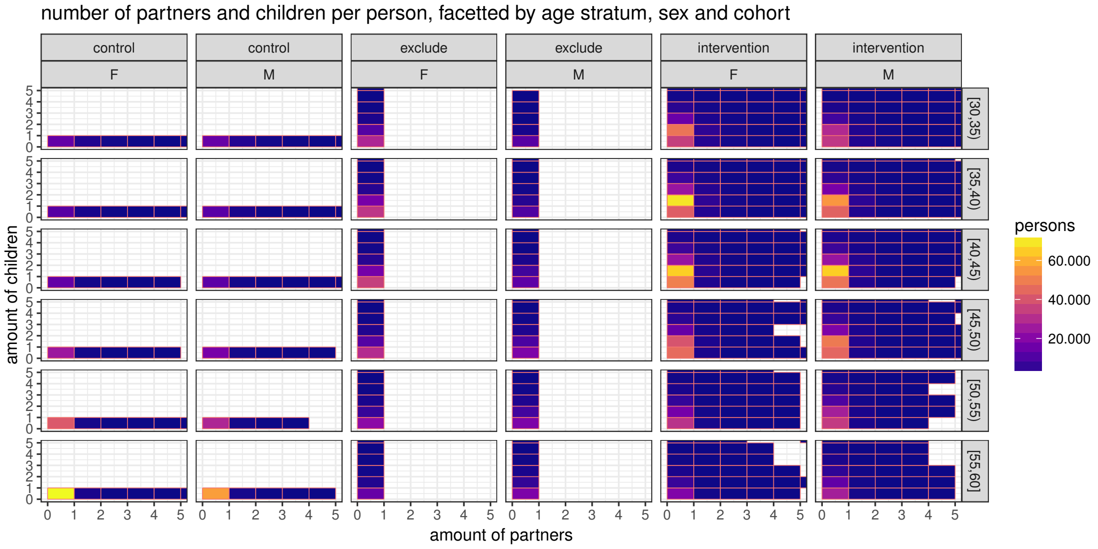


Figure 3.72: Number of partners and children per person, split by age stratum, sex and cohort assignment

Summarizing, the number of associated partners and children per person appear to be reasonable and correspond with the author's expectations. The most likely and presumably most common situations are by far the most frequent cases in the presented cohort.

A significant increase in data quality, expressed by a higher proportion of common cases and a lower spread of distributions describing the number of partners and children per person, is found in the reduced data set *no-NA*. This observation is a direct effect of the introduced restrictions. Thus, the applied filters increase the quality of the dataset even for variables that do not originate from them. Overall, this increases the trustworthiness of these measures and the resulting dataset. Consequently, the filtered dataset *no-NA* is used for statistical analysis.

3.5 Statistical analysis

Building on data extraction, data transformation and preparation, exploratory analysis, and data quality assessment, this chapter presents the resulting data set, followed by various ways of statistical analysis and interpretation of the results.

First, the final study cohort, split by the outcome criteria, is summarized, including the univariate tests in chapter 3.5.1. Next, as suggested in the study protocol, crosstabulations are listed in chapter 3.5.2 on page 126. More complex methods such as decision trees in chapter 3.5.4 on page 137, regression models in chapter 3.5.3 on page 128 and gradient boosting machine to optimize out-of-sample prediction for this highly unbalanced dataset in chapter 3.5.5 on page 145 are applied subsequently. In addition, propensity score matching utilizing the numeric socioeconomic status (SES) and multimorbidity classification (i.e., *Charlson (group)*) is evaluated to obtain balanced groups.

Resampling methods are applied to optimize model results and predictions where appropriate and computationally feasible.

3.5.1 Univariate description

The final data set is outlined in this chapter. Based on the previous analysis, the dataset without missing information *no-NA* with 1.942.109 entries is used as the basis. All excluded individuals [33] and individuals with missing values in the variable *soes_mean*[34] are removed[35], resulting in 1.576.061 observations.

The evolution of the final dataset and cohorts is shown in figure 3.73 on the following page. Starting from the entire population of Austria and the content of the database GAP-DRG, the co-insured population down to the selected individuals are summarized as a flowchart, inspired by the PRISMA Statement.

The outline of the entire final dataset and variables used in the subsequent analysis are listed in table 3.41 on page 125 for documentation and reference. There are no more

[33] 364.817 persons or 18,78% of the entire dataset

[34] 2.083 persons or 0,11% of the entire dataset

[35] in total, 366.048 persons or 18,85% of the entire dataset are removed due to these constraints

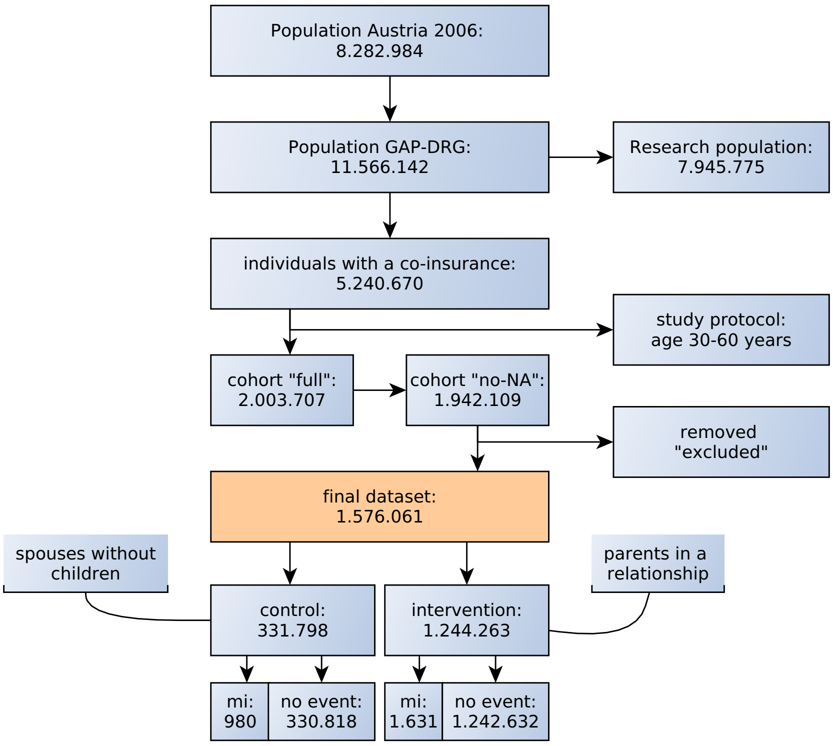


Figure 3.73: Population and cohort flowchart

missing values. This table represents the starting point of a rather fair progression from the entire dataset without any segmentation to the crosstabulation suggested by the study protocol, which are presented in the following chapter 3.5.2 on page 126.

Univariate statistics and tests split by cohort assignment are presented in table 3.42 on page 126. The rightmost column includes test statistics from the Wilcoxon signed-rank test [Wilcoxon, 1945] for continuous variables and the Pearson's chi-squared test [Pearson, 1900] for categorical variables. All univariate tests suggest highly significant differences between the cohorts.[36]

As detailed in chapter 3.4 on page 110, there are apparent imbalances between cohorts in terms of age groups, sex, and, possibly as a correlated variable, comorbidity class.

The same comparison split by outcome variable *mi* is found in table 3.43 on page 127. In total, there are only 2.611 recorded myocardial infarctions in the data set. As is generally known, elderly men are the most affected group. Most strikingly, the $\tilde{\chi}^2$ statistics of cohort assignment and the essentially identical variable *parent*, although highly significant, are at a comparatively lower level than the differences in age and sex.asse.

[36] The level of significance is a result of the rather large cohort sizes, which can be observed in, e.g., variable *partner_sum*. Nevertheless, there are actual differences between the cohorts.

Table 3.41: Baseline characteristics. $_a$ b $_c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD. Numbers after proportions are frequencies.

	$N = 1576061$	
age07	$_{37}$ 43 $_{50}$	(44 ± 8)
age_stratum : [30,35)	14%	(225909)
[35,40)	20%	(314793)
[40,45)	22%	(341114)
[45,50)	18%	(280142)
[50,55)	13%	(206841)
[55,60]	13%	(207262)
sex : F	51%	(808278)
M	49%	(767783)
partner_sum	$_{1,0}$ 1,0 $_{1,0}$	(1,1 ±0,4)
soes_mean	$_{1,7}$ 1,9 $_{2,1}$	(1,9 ±0,3)
Charlson_group : 0	38%	(591673)
1-3	50%	(791138)
4+	12%	(193250)
mi : FALSE	100%	(1573450)
TRUE	0%	(2611)
cohort : control	21%	(331798)
intervention	79%	(1244263)

Finally, the equivalent univariate overview is presented in table 3.45 on page 129 for both cohorts, each split by variable *mi*.

Although direct comparisons are not possible at this stage, it appears that there is a different pattern concerning the relative number of myocardial infarctions per age group between the two cohorts. Whereas in the control cohort the most affected patients are in the highest age group, in the intervention cohort they are more evenly distributed. As already documented in figure 3.63 on page 109, this difference is most likely an artifact of the overall distribution of age and sex in each cohort.

Table 3.42: Baseline characteristics by cohort. $_a$ b $_c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD. Numbers after proportions are frequencies. Tests used: ^{1}Wilcoxon test; ^{2}Pearson test .

	control $N = 331798$	intervention $N = 1244263$	Test Statistic
age07	$_{44}$ 52 $_{57}$ (50 ± 9)	$_{37}$ 42 $_{47}$ (42 ± 7)	$F_{1,1576059}$=2e+05, P<0.001^1
age_stratum : [30,35)	10% (33594)	15% (192315)	χ^2_5=3e+05, P<0.001^2
[35,40)	8% (25232)	23% (289561)	
[40,45)	8% (26626)	25% (314488)	
[45,50)	13% (42856)	19% (237286)	
[50,55)	21% (71242)	11% (135599)	
[55,60]	40% (132248)	6% (75014)	
sex : F	56% (184646)	50% (623632)	χ^2_1=3206, P<0.001^2
M	44% (147152)	50% (620631)	
partner_sum	$_{1,0}$ 1,0 $_{1,0}$ (1,1 ±0,3)	$_{1,0}$ 1,0 $_{1,0}$ (1,1 ±0,4)	$F_{1,1576059}$=2022, P<0.001^1
soes_mean	$_{1,8}$ 2,0 $_{2,3}$ (2,0 ±0,4)	$_{1,7}$ 1,9 $_{2,1}$ (1,9 ±0,3)	$F_{1,1576059}$=34713, P<0.001^1
Charlson_group : 0	32% (105561)	39% (486112)	χ^2_2=25342, P<0.001^2
1-3	48% (159408)	51% (631730)	
4+	20% (66829)	10% (126421)	
mi : FALSE	100% (330818)	100% (1242632)	χ^2_1=427, P<0.001^2
TRUE	0% (980)	0% (1631)	

In summary, the univariate listings and tests presented, while providing important background information and references, tempt to jump to conclusions by hiding multivariate relationships. They are formatted analogously to the first summary table, i.e., *table one*, in common scientific publications and therefore provide an important basis for the following analysis.

3.5.2 Cross tabulation

Results are presented in tables 3.46 on page 129 and 3.47 on page 130 for both cohorts. A visualization of the same data can be found in figure 3.74 on page 130.

The absolute and relative numbers of persons suffering myocardial infarction increase with age for both cohorts and gender. It is striking that there are very few cases in the youngest age groups. For example, there is only one woman between 35 and 39 years of age in the control cohort with a recorded event. The same rate (0.08‰) is found for mothers in the youngest group. Overall, crude prevalence rates appear to increase more rapidly and are generally higher in the cohort control, contradicting the original assumption of the study.

Table 3.43: Baseline characteristics by outcome. $_a$ b $_c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD. Numbers after proportions are frequencies. Tests used: [1]Wilcoxon test; [2]Pearson test .

	myocardial infarction	no event	Test Statistic
	$N = 2611$	$N = 1573450$	
age07	$_{46}$ 51 $_{57}$ (51 ± 7)	$_{37}$ 43 $_{50}$ (44 ± 8)	$F_{1,1576059}$=1784, P<0.001[1]
age_stratum : [30,35)	1% (39)	14% (225870)	χ^2_5=1937, P<0.001[2]
[35,40)	6% (145)	20% (314648)	
[40,45)	12% (320)	22% (340794)	
[45,50)	21% (545)	18% (279597)	
[50,55)	24% (635)	13% (206206)	
[55,60]	36% (927)	13% (206335)	
sex : F	18% (483)	51% (807795)	χ^2_1=1125, P<0.001[2]
M	82% (2128)	49% (765655)	
partner_sum	$_{1,0}$ 1,0 $_{1,0}$ (1,1 ±0,3)	$_{1,0}$ 1,0 $_{1,0}$ (1,1 ±0,4)	$F_{1,1576059}$=2, P=0.2[1]
parent : FALSE	38% (980)	21% (330818)	χ^2_1=427, P<0.001[2]
TRUE	62% (1631)	79% (1242632)	
child_sum	$_0$ 1 $_2$ (1 ±1)	$_1$ 1 $_2$ (2 ±1)	$F_{1,1576059}$=304, P<0.001[1]
soes_mean	$_{1,8}$ 2,0 $_{2,3}$ (2,0 ±0,4)	$_{1,7}$ 1,9 $_{2,1}$ (1,9 ±0,3)	$F_{1,1576059}$=325, P<0.001[1]
Charlson_group : 0	5% (126)	38% (591547)	χ^2_2=6201, P<0.001[2]
1-3	33% (863)	50% (790275)	
4+	62% (1622)	12% (191628)	
cohort : control	38% (980)	21% (330818)	χ^2_1=427, P<0.001[2]
intervention	62% (1631)	79% (1242632)	

Both tables are summarized in figure 3.74. In the upper part, absolute numbers are plotted by gender (columns), cohort (rows), age stratum (abscissa), and outcome (color). Transparent lines are added to emphasize the overall trend. The absolute number of individuals is scaled log-10. Overall, the same trend as in the previous tables can be observed more directly, although it is still not clear whether the differences result from an overall trend in each cohort or whether there is a specific distinction.

Rates per 1.000 population, sex, age stratum, and cohort are presented in the lower part of figure 3.74 for individuals with myocardial infarction. It is clear that men are more frequently affected than women and that there is a marked increase with increasing age. In addition, the relative number of affected individuals in the control cohort exceeds that of the intervention cohort in almost all subgroups shown.

In summary, despite the required matching of individuals by socioeconomic status (SES) and the proposed additional variables, these cross-tabulations resolve the main part of the study protocol. No significant evidence for the initial assumption can be identified.

Table 3.44: Baseline characteristics by cohort and outcome

control

	myocardial infarction	no event	Test Statistic
	$N = 980$	$N = 330818$	
age07	52 56 59 (55 ± 5)	44 52 57 (50 ± 9)	$F_{1,331796}$=286, P<0.001[1]
age_stratum : [30,35)	1% (6)	10% (33588)	χ^2_5=280, P<0.001[2]
[35,40)	2% (15)	8% (25217)	
[40,45)	3% (33)	8% (26593)	
[45,50)	10% (95)	13% (42761)	
[50,55)	24% (236)	21% (71006)	
[55,60]	61% (595)	40% (131653)	
sex : F	26% (257)	56% (184389)	χ^2_1=345, P<0.001[2]
M	74% (723)	44% (146429)	
partner_sum	1,0 1,0 1,0 (1,0 ±0,2)	1,0 1,0 1,0 (1,1 ±0,3)	$F_{1,331796}$=8, P=0.004[1]
soes_mean	1,8 2,1 2,4 (2,1 ±0,4)	1,8 2,0 2,3 (2,0 ±0,4)	$F_{1,331796}$=52, P<0.001[1]
Charlson_group : 0	4% (40)	32% (105521)	χ^2_2=1442, P<0.001[2]
1-3	28% (273)	48% (159135)	
4+	68% (667)	20% (66162)	

In subsequent chapters, more sophisticated methods and additional variables are used to understand and quantify the influence of covariates and their association.

3.5.3 Logistic regression

In this chapter, results from two multivariate logistic regression models are presented as described in the corresponding methods section. In addition, interactions between variables (e.g., age and gender), a direct comparison of related models, and confirmation of results by cross-validation are included.

The resulting coefficients (i.e., log-odds) transformed into odds ratios and their 95% confidence intervals are presented as a forest plot in figure 3.75 on page 131 for model 1 and figure 3.76 for model 2. Odds ratios are chosen instead of log-odds to facilitate interpretation of effect sizes.

Both models show similar effects. While the intercept is very low because of the rarity of the observed events, most of the other predictors increase the risk of having a myocardial infarction.

In both models, increasing age can be identified as a risk factor. Model 1, in which age is included as a continuous variable, estimates a risk increase of approximately 75% for every 10 years of age. In model 2, a similar trend is observed with even higher coefficients for different age strata compared with the youngest group of 30- to 34-year-olds. It is

Table 3.45: Baseline characteristics by cohort and outcome. $_a\ b\ _c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD. Numbers after proportions are frequencies. Tests used: ^{1}Wilcoxon test; ^{2}Pearson test .

intervention

	myocardial infarction		no event		Test Statistic
	$N = 1631$		$N = 1242632$		
age07	$_{44}$ 48 $_{54}$	(48 ± 7)	$_{37}$ 42 $_{47}$	(42 ± 7)	$F_{1,1244261}$=1094, P<0.001^1
age_stratum : [30,35)	2%	(33)	15%	(192282)	χ^2_5=1287, P<0.001^2
[35,40)	8%	(130)	23%	(289431)	
[40,45)	18%	(287)	25%	(314201)	
[45,50)	28%	(450)	19%	(236836)	
[50,55)	24%	(399)	11%	(135200)	
[55,60]	20%	(332)	6%	(74682)	
sex : F	14%	(226)	50%	(623406)	χ^2_1=859, P<0.001^2
M	86%	(1405)	50%	(619226)	
partner_sum	$_{1,0}$ 1,0 $_{1,0}$	(1,1 ±0,4)	$_{1,0}$ 1,0 $_{1,0}$	(1,1 ±0,4)	$F_{1,1244261}$=1, P=0.3^1
soes_mean	$_{1,8}$ 2,0 $_{2,2}$	(2,0 ±0,3)	$_{1,7}$ 1,9 $_{2,1}$	(1,9 ±0,3)	$F_{1,1244261}$=188, P<0.001^1
Charlson_group : 0	5%	(86)	39%	(486026)	χ^2_2=4310, P<0.001^2
1-3	36%	(590)	51%	(631140)	
4+	59%	(955)	10%	(125466)	

Table 3.46: Results: crosstabulation for cohort *intervention*

	females			**males**			$\sum$		
age group	N	mi	rate ‰	N	mi	rate ‰	N	mi	rate ‰
[30,35)	113.596	9	0,08	78.719	24	0,3	192.315	33	0,2
[35,40)	157.469	22	0,14	132.092	108	0,8	289.561	130	0,5
[40,45)	157.764	40	0,25	156.724	247	1,6	314.488	287	0,9
[45,50)	110.203	60	0,54	127.083	390	3,1	237.286	450	1,9
[50,55)	57.470	58	1,01	78.129	341	4,4	135.599	399	2,9
[55,60]	27.130	37	1,36	47.884	295	6,2	75.014	332	4,4
$\sum$	623.632	226	0,36	620.631	1.405	2,3	1.244.263	1.631	1,3

likely that model 2 more directly reflects the nonlinear increase in risk with increasing age but may also overestimate the true impact because of the rarity of events in the youngest group.

Being male rather than female shows an almost equal increase in risk in both models.

Table 3.47: Results: crosstabulation for cohort *control*

	females			males			∑		
age group	N	mi	rate ‰	N	mi	rate ‰	N	mi	rate ‰
[30,35)	16.592	2	0,12	17.002	4	0,2	33.594	6	0,2
[35,40)	12.146	1	0,08	13.086	14	1,1	25.232	15	0,6
[40,45)	14.111	10	0,71	12.515	23	1,8	26.626	33	1,2
[45,50)	25.253	29	1,15	17.603	66	3,8	42.856	95	2,2
[50,55)	42.129	65	1,54	29.113	171	5,9	71.242	236	3,3
[55,60]	74.415	150	2,02	57.833	445	7,7	132.248	595	4,5
∑	184.646	257	1,39	147.152	723	4,9	331.798	980	3,0

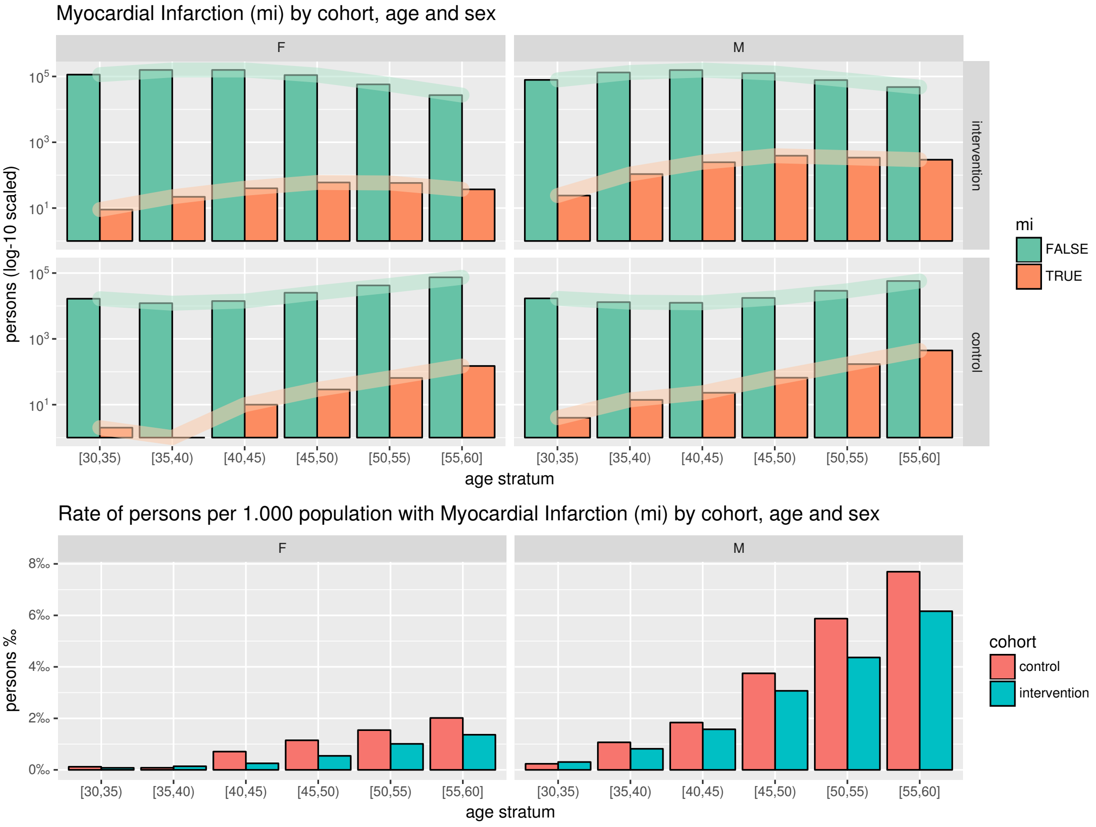
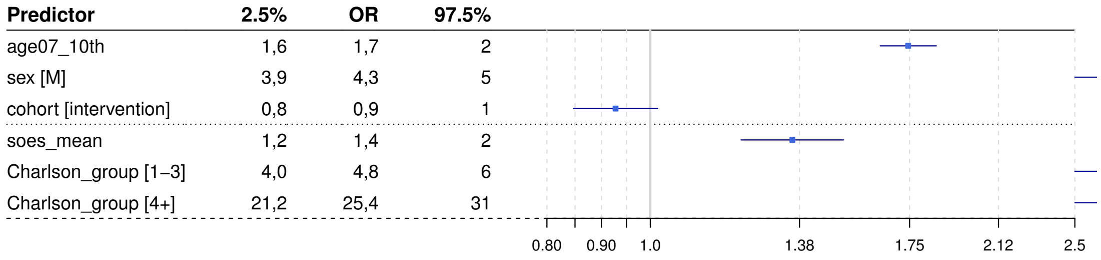


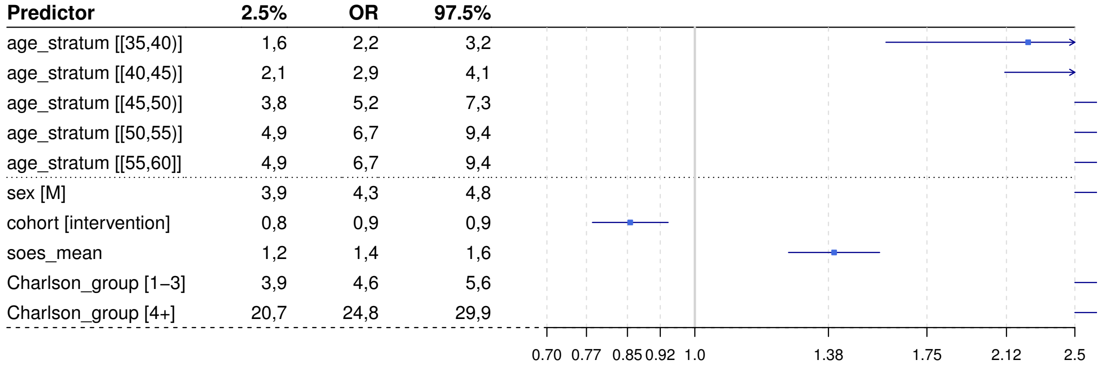
Figure 3.74: Visualization of cross tabulated results: absolute and relative counts

An increase (i.e., worsening) in socioeconomic status and a higher Charlson comorbidity group compared with the lowest (i.e., best) increases the risk of having a myocardial



Predictor	2.5%	OR	97.5%
age07_10th	1,6	1,7	2
sex [M]	3,9	4,3	5
cohort [intervention]	0,8	0,9	1
soes_mean	1,2	1,4	2
Charlson_group [1–3]	4,0	4,8	6
Charlson_group [4+]	21,2	25,4	31

Figure 3.75: Logistic regression: forestplot for model 1



Predictor	2.5%	OR	97.5%
age_stratum [[35,40)]	1,6	2,2	3,2
age_stratum [[40,45)]	2,1	2,9	4,1
age_stratum [[45,50)]	3,8	5,2	7,3
age_stratum [[50,55)]	4,9	6,7	9,4
age_stratum [[55,60]]	4,9	6,7	9,4
sex [M]	3,9	4,3	4,8
cohort [intervention]	0,8	0,9	0,9
soes_mean	1,2	1,4	1,6
Charlson_group [1–3]	3,9	4,6	5,6
Charlson_group [4+]	20,7	24,8	29,9

Figure 3.76: Logistic regression: forestplot for model 2

infarction significantly and similarly in both models.

Most interestingly, individuals assigned to the cohort *intervention* rather than the cohort *control* lower their risk in both models, although the effect is not significant in model 1. Nevertheless, the hypothesis that parents are at higher risk is not supported.

Table 3.48 on the following page shows model statistics, indicators of significance of influence (asterisks), and Wald confidence intervals for odds ratios. In addition, models 1b and 2b are listed with interaction terms between sex and the corresponding age variables.

Considering the model statistic *Akaike Information Criterion* (AIC), shown in the last row of the table 3.48 on the next page, each model listed from left to right performs slightly better than its predecessor (i.e., left neighbor). The confidence intervals differ a bit, but do not change the overall conclusions. It is worth noting that the influence of individual variables appears to be quite strong. In particular, individuals in the highest (i.e., worst) Charlson group 4+ show a significantly higher risk of having a myocardial

Table 3.48: Logistic regression results

	Dependent variable:			
	Myocardial infarction (mi)			
	Model 1a	Model 1b	Model 2a	Model 2b
age07_10th	2,00***	2,00***		
	(2,00–2,00)	(2,00–2,00)		
age_stratum [[35,40)]			2,00***	1,00
			(2,00–3,00)	(0,80–2,00)
age_stratum [[40,45)]			3,00***	3,00***
			(3,00–3,00)	(2,00–3,00)
age_stratum [[45,50)]			5,00***	5,00***
			(5,00–6,00)	(4,00–6,00)
age_stratum [[50,55)]			7,00***	7,00***
			(6,00–7,00)	(7,00–8,00)
age_stratum [[55,60]]			7,00***	8,00***
			(6,00–7,00)	(7,00–9,00)
sex [M]	4,00***	17,00***	4,00***	5,00***
	(4,00–4,00)	(16,00–17,00)	(4,00–4,00)	(4,00–5,00)
cohort [intervention]	0,90	0,90	0,90***	0,90***
	(0,80–1,00)	(0,80–1,00)	(0,80–0,90)	(0,80–1,00)
soes_mean	1,00***	1,00***	1,00***	1,00***
	(1,00–1,00)	(1,00–1,00)	(1,00–2,00)	(1,00–2,00)
Charlson_group [1-3]	5,00***	5,00***	5,00***	5,00***
	(5,00–5,00)	(5,00–5,00)	(4,00–5,00)	(4,00–5,00)
Charlson_group [4+]	25,00***	26,00***	25,00***	25,00***
	(25,00–26,00)	(25,00–26,00)	(25,00–25,00)	(25,00–25,00)
age07_10th:sex [M]		0,80***		
		(0,60–0,90)		
age_stratum [[35,40)]:sex [M]				2,00
				(0,80–2,00)
age_stratum [[40,45)]:sex [M]				1,00
				(0,40–2,00)
age_stratum [[45,50)]:sex [M]				1,00
				(0,30–2,00)
age_stratum [[50,55)]:sex [M]				0,90
				(0,10–2,00)
age_stratum [[55,60]]:sex [M]				0,80
				(0,08–2,00)
Log Likelihood	-16.461,00	-16.453,00	-16.417,00	-16.410,00
Akaike Inf. Crit.	32.936,00	32.922,00	32.857,00	32.853,00
Note:				*p<0,1; **p<0,05; ***p<0,01

infarction compared with Group 0.

The risk increases dramatically with age and for men compared with women. The two models with interactions between age and sex show only slightly different coefficients in most cases. A deviation from this observation is the gender variable in models 1a and 1b, where a significantly worse result is observed for males compared to females when the interaction terms are included. As a trade-off, model 1b suggests that the interaction of increasing age and being male rather than female reduces the risk for a positive outcome, which is not consistent with all other results and the author's expectations.

Stepwise model selection with both forward and backward searches was performed to find the optimal variables, using AIC as the benchmark. In any case, no better model than those presented with the total number of variables can be identified.

To obtain more robust information about the variance and accuracy of these logistic regression models, 100-fold cross-validation is performed for model 1 and model 2.

Nearly identical results and interpretations can be obtained from most of the performance measures presented across both sets of cross-validated models. Therefore, only the common ROC curves and the areas under these curves are presented and discussed here.

Figure 3.77 shows two different but closely related plots. The upper one contains 100 ROC curves for all cross-validation partitions of model 1. The individual curves are transparent to some extent, but it is still not possible to directly derive a valid summary. Nevertheless, it can already be observed that almost all cross-validation sets perform quite well, although there is some variation and even individual outliers can be identified.

The bottom plot of figure 3.77 on the following page summarizes all ROC curves in individual steps as boxplots. In addition, the median of all original ROC curves is highlighted as a red line and the areas under the curves are summarized in the lower right corner. This gives a clear picture of the median model and the dispersion of the quality of a model when 99% subsets of the data are randomly selected for training. Common measures such as the mean and standard deviation *sd*, as well as robust measures (median and interquartile range, i.e., IQR) are presented for the univariate measure *area under the ROC curve* (i.e. AUC or AUROC) calculated using [Robin et al., 2011].

Most useful information can be obtained from the second bottom plot of figure 3.77 on the next page. It can be seen that the median model performs significantly better than the random class assignment and that there are individual outliers. Between a false positive percentage of 0,02 and 0,4, the dispersion of the quality of all models seems to be rather symmetrical. The mean and median of the areas under the ROC curves are both 0,87 with an identical spread.

Figure 3.78 on page 135 shows the distribution of the regression coefficient as an odds ratio for each variable from all 100 cross-validation models as a boxplot and individual values (points). A horizontal random jitter is added to the points and they are colored according to the AUC of the associated model. Several features are expected from this plot. The dispersion, the number of outliers, and the symmetry of each boxplot indicate

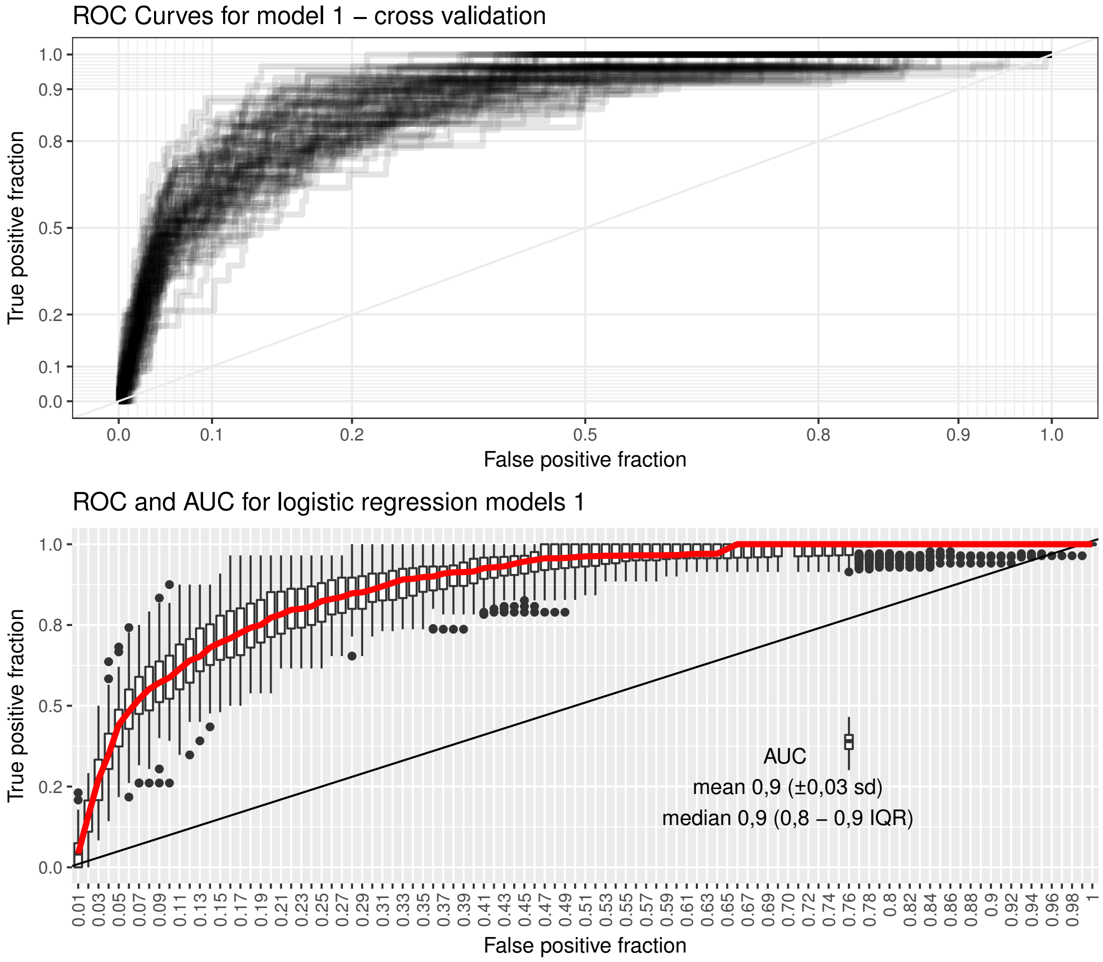


Figure 3.77: Logistic regression cross validation: ROC and AUC for model 1

the reliability and stability of a variable. In addition, the distribution and possibly the clustering of coefficients with similar AUC could indicate a previously unnoticed trend. This method is significantly limited by its focus on single variables and leaves out multivariate relationships. It should also be noted that odds ratios are presented instead of log-odds. This can make conclusions about symmetry misleading.

In Figure 3.78, it can be observed that there is at least one outlier with comparably poor model performance, present for all variables except cohort assignment. In terms of the dispersion of coefficients, it is most striking for the gender. All but the coefficients for the Charlson group show a rather symmetric spread around the median. A slight tendency for models with a higher AUC in lower ranges is present at least for the highest Charlson group. It can be concluded that models that do not overestimate this regressor tend to perform better, although the absolute effect is still very high. A similar association can

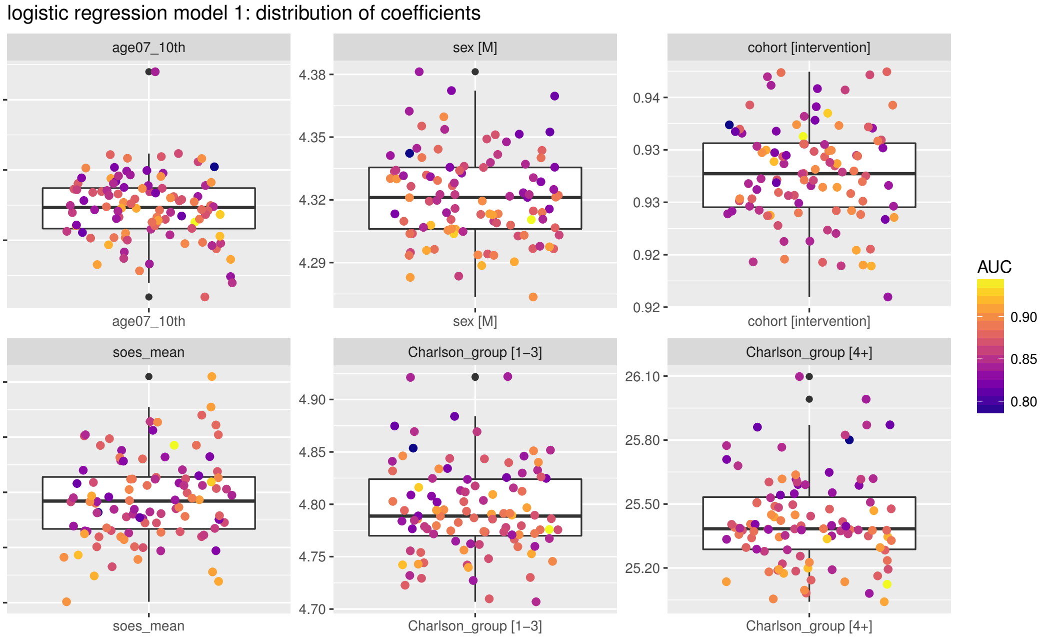


Figure 3.78: Logistic regression cross validation: distribution of coefficients for model 1

be found for the gender variable.

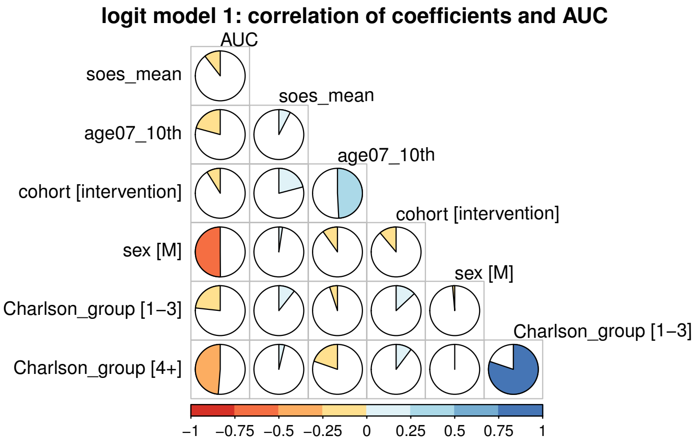


Figure 3.79: Logistic regression cross validation: correlation of coefficients for model 1

Figure 3.79 shows a visualization of the correlation matrix of all coefficients and the AUC of the model for all 100 runs of model 1. It confirms the previously suspected correlation between lower scores for the worst Charlson comorbidity group and gender (i.e., being male rather than female) with the performance of the model measured as AUC. Interestingly, all other coefficients also appear to be slightly negatively correlated with the area under the ROC curve. This finding fits well with expectations, as the outcome is a very rare and thus unlikely event. As a result, the number of false-positive classifications exceeds the number of false-negative assignments by several orders of magnitude.

The same information for a 100-fold cross-validation of model 2 is shown in 3.80, 3.81, and 3.82 on page 138. All randomly selected partitions are not overlapping and not identical to the selection used to validate model 1.

Boxplots summarizing ROC curves and the area under these curves are shown in figure 3.80. Individual ROC curves are omitted compared to 3.77 due to their low significance. The boxplots document a fairly similar distribution of the ROC curves, which is also confirmed by the nearly equal AUC. Only the scatter seems to be higher in contrast to model 1.

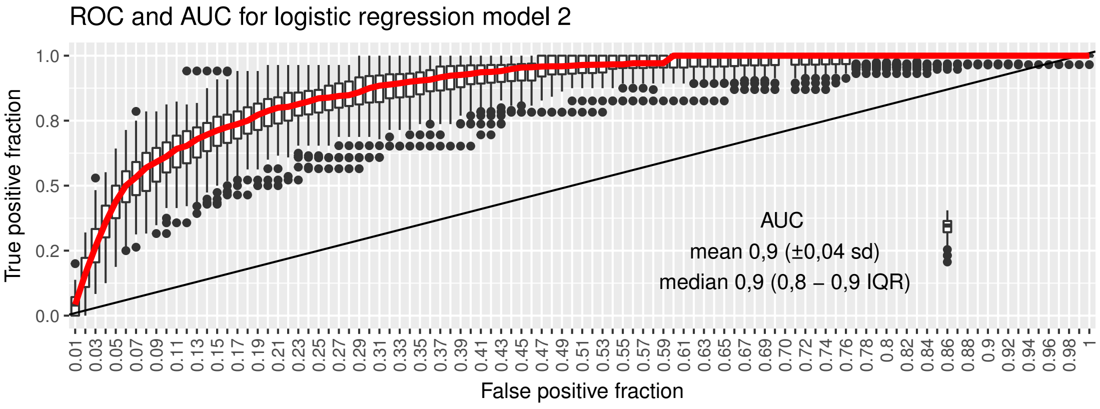


Figure 3.80: Logistic regression cross validation: median ROC and AUC for model 2

Figure 3.81 on the facing page shows the estimated coefficients for all variables. It appears that there are more outliers in general. In particular, the individual levels representing age are not symmetrically distributed. These disparities actually increase with increasing age, showing two clusters of models for the oldest group, one closely around the median and a second slightly above the upper boundary of the box. Moreover, the correlation of model performance, measured as AUC and coded as color, is not as prominent as in model 1.

The same conclusion can be drawn from figure 3.82 on page 138. While the coefficients for each level of age are positively correlated, a negative correlation is observed between age and comorbidity. As in mode 1, the AUC is correlated with both groups of Charlson scores but shows a weaker association with the gender variable.

In summary, the presented models show quite stable performance and are able to classify the training dataset much better than a random guess. Due to the small number of actual cases and their uneven distribution, a high number of false positives is unavoidable. In addition, all models suffer greatly from correlated input data and unbalanced outcomes.

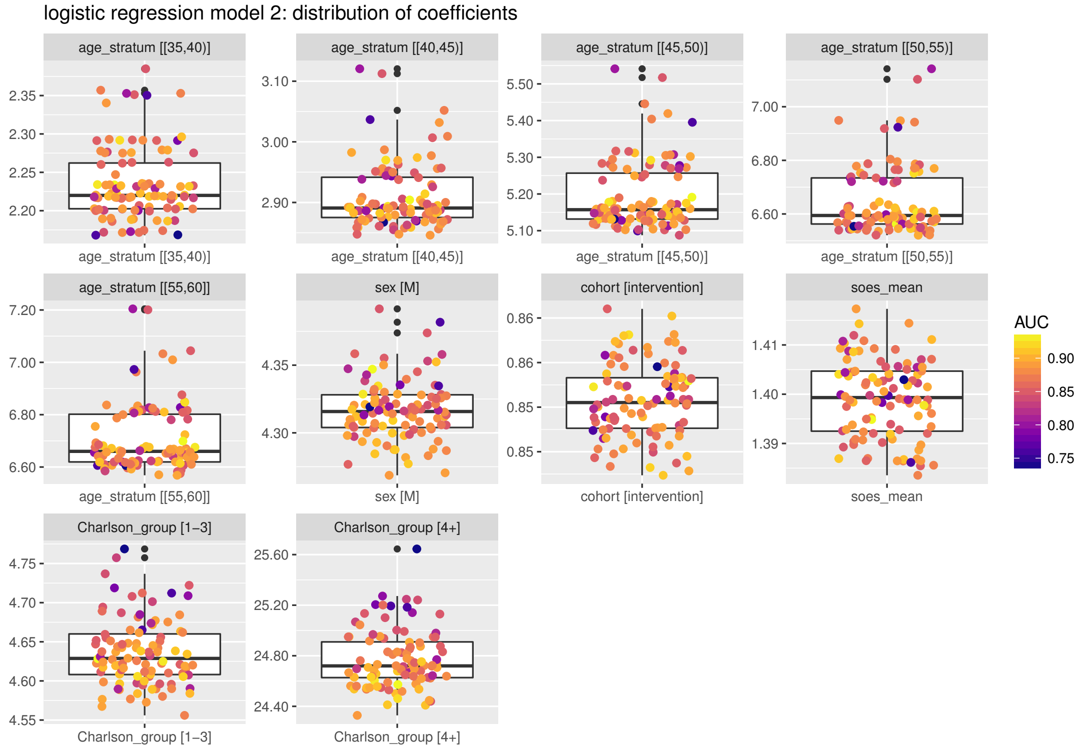


Figure 3.81: Logistic regression cross validation: distribution of coefficients for model 2

Under these circumstances, cohort membership cannot be determined as a relevant influential factor, and being a parent (i.e., cohort intervention) does not increase the risk of myocardial infarction, in contrast to being an adult in a partnership without children (i.e., cohort control).

3.5.4 Decision trees

Four FFTree models are calculated and presented in this chapter.

reference The reference model is a very basic tree, only involving two variables, age and sex.

small For the small tree, the variable of interest *cohort* is added.

full In the full tree also the Charlson groups and the SES (variable *soes_mean*) are included.

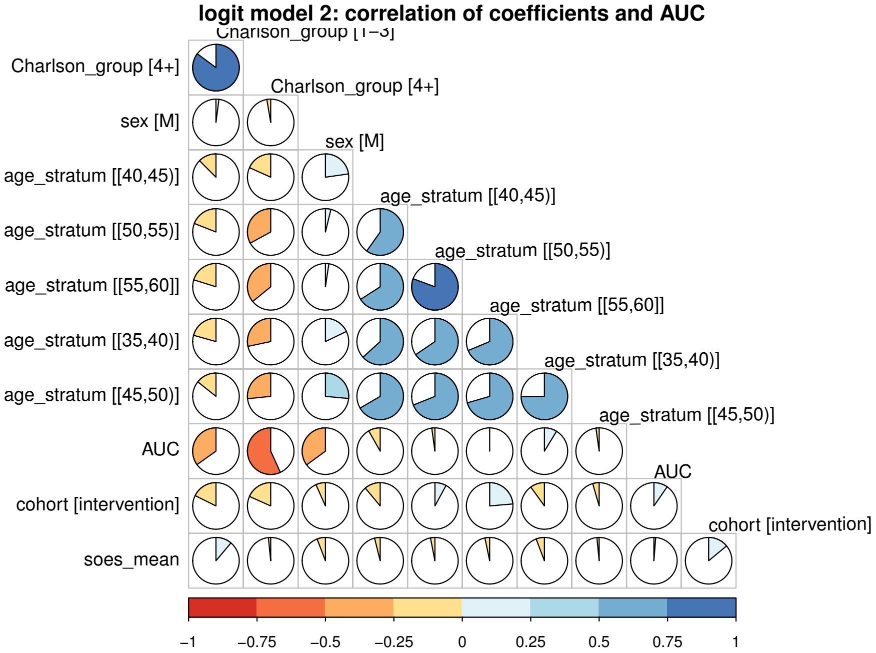


Figure 3.82: Logistic regression cross validation: correlation of coefficients for model 2

full default In comparison to the full model, the default parameters of the FFTree routines are left unchanged. As a result, only the four most important decisions are included, lacking cohort assignment.

Table 3.49 gives the absolute number and table 3.50 holds the percentages per row for the training and test datasets. It can be observed that the samples are stratified equally between both cohorts but that there are slight deviations from 75% for the outcome variable *mi*.[37]

	train: no mi	train: mi	test: no mi	test: mi	$\sum$
control	248.146	702	82.672	278	331.798
intervention	931.946	1.251	310.686	380	1.244.263
$\sum$	1.180.092	1.953	393.358	658	1.576.061

Table 3.49: Sampled data: number of observations

The best of the two possible trees in the reference model, containing only age and gender as predictor variables, is shown in figure 3.83. It states that individuals younger than

[37] Overall, there are 76, 7% of *mi* events from cohort *intervention*, 71, 63% of *mi* event from cohort *control* and 74, 8% for all events in the training dataset.

	train: no mi	train: mi	test: no mi	test: mi	$\sum$
control	75	0,2	25	0,08	100
intervention	75	0,1	25	0,03	100
$\sum$	75	0,1	25	0,04	100

Table 3.50: Sampled data: row percentages

45 years and females are not affected. Conversely, all males older than 44 years are classified as positive outcomes. Given the small number of variables and information, about two-thirds of all real cases are correctly identified.

This tree will be used as a baseline model. Subsequently, more complex decision trees are interpreted not only separately, but usually in comparison with this reference. Sensitivity and specificity are both at a moderate level. The ROC plot shows that there are two trees in this model and that logistic regression and CART perform rather poorly in comparison. Of course, the precision (0,005) is rather low, which is also manifested in a high FPR (0,23) and NPV (0,999).

Further details are listed in table 3.51.

Table 3.51: FFTree reference: performance measures

TP	FP	TN	FN	AUC	sens	spec
431	88.874	304.484	227	0,8	0,7	0,8

PPV	NPV	FPR	FNR	FDR	ACC	F1	κ
0	1	0,2	0,3	1	0,8	0,01	0,8

In contrast to the first tree of the reference model, the second is shown in figure 3.84. All individuals older than 44 years and all males are classified as positive results. The hit rate is quite high, which means that almost all actual cases are detected.[38] The high sensitivity comes at the cost of very low precision and generally worse model measures as tree *#1*.

Performance measures of this tree are listed in table 3.52.

This example also illustrates the basic considerations and assumptions of this study. The cross-tabulations and models presented describe demonstrably low rates of myocardial infarction in a fairly large population. On the one hand, every individual is at risk and therefore could suffer an event; on the other hand, most individuals will not have a positive finding or doubt a heart attack by, for example, showing up at the hospital with severe chest pain. Consequently, the superficial question that these decision trees can

[38] Only the few cases where young females are suffering from myocardial infarction are not recognized.

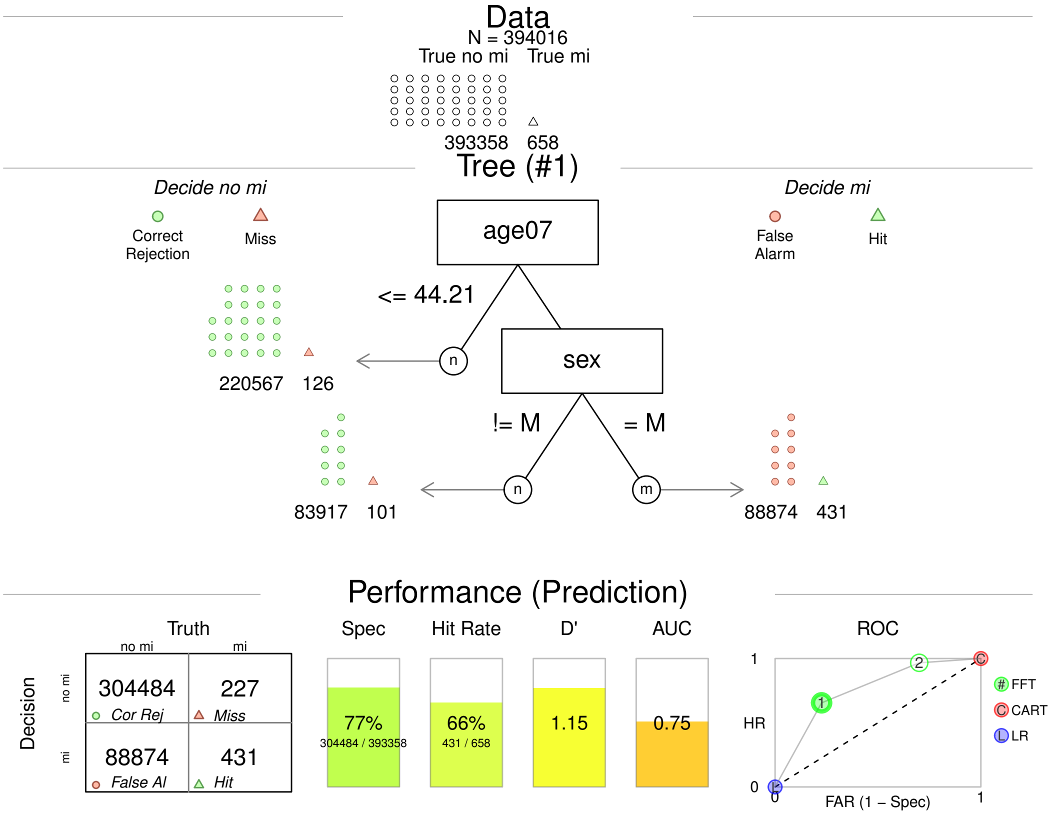


Figure 3.83: Fast and Frugal Tree: reference model, best tree

answer is "Does anyone in the entire selected population have a heart attack during 2006 or 2007?" Hence, the high false positive rate and low precision. Given the underlying research question of whether being a parent affects the risk of a positive outcome, it is not the crude predictive power of the data, variables, and models for the entire population that is of interest but the relative improvement when the distinction between the two defined cohorts of intervention and control is included.

Table 3.52: FFTree reference - tree #2: performance measures

TP	FP	TN	FN	AUC	sens	spec
636	275.484	117.874	22	0,8	1	0,3

PPV	NPV	FPR	FNR	FDR	ACC	F1	κ
0	1	0,7	0,03	1	0,3	0	0,3

Figure 3.85 shows a small decision tree with all variables from the reference model and the

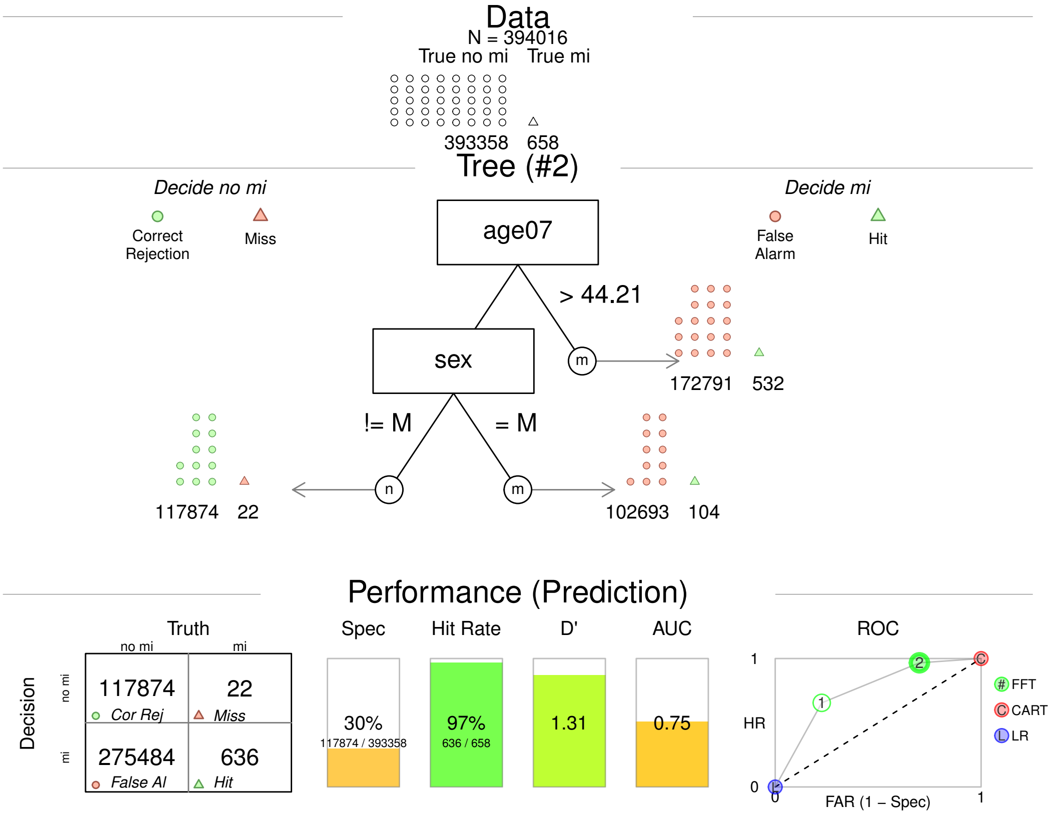


Figure 3.84: Fast and Frugal Tree: *reference* model, 2^{nd} tree

cohort assignment. The first two nodes or decisions correspond to tree #1 of the reference model in figure 3.83. In summary, the presented tree performs little better than the reference model, although a slightly higher AUC of 0,76 might indicate some improvement. The recall (0,76) is higher, although the number of false-positive predictions increases sharply.

The overall benefit of adding cohort assignment information compared to the reference model is not clear in this case.

Performance measures are listed in table 3.53.

Table 3.53: FFTree small: performance measures

TP	FP	TN	FN	AUC	sens	spec
502	124.298	269.060	156	0,8	0,8	0,7

Next, all available variables, including Charlson comorbidity group and socioeconomic status, are used with default FFTrees settings. The best resulting tree is visualized

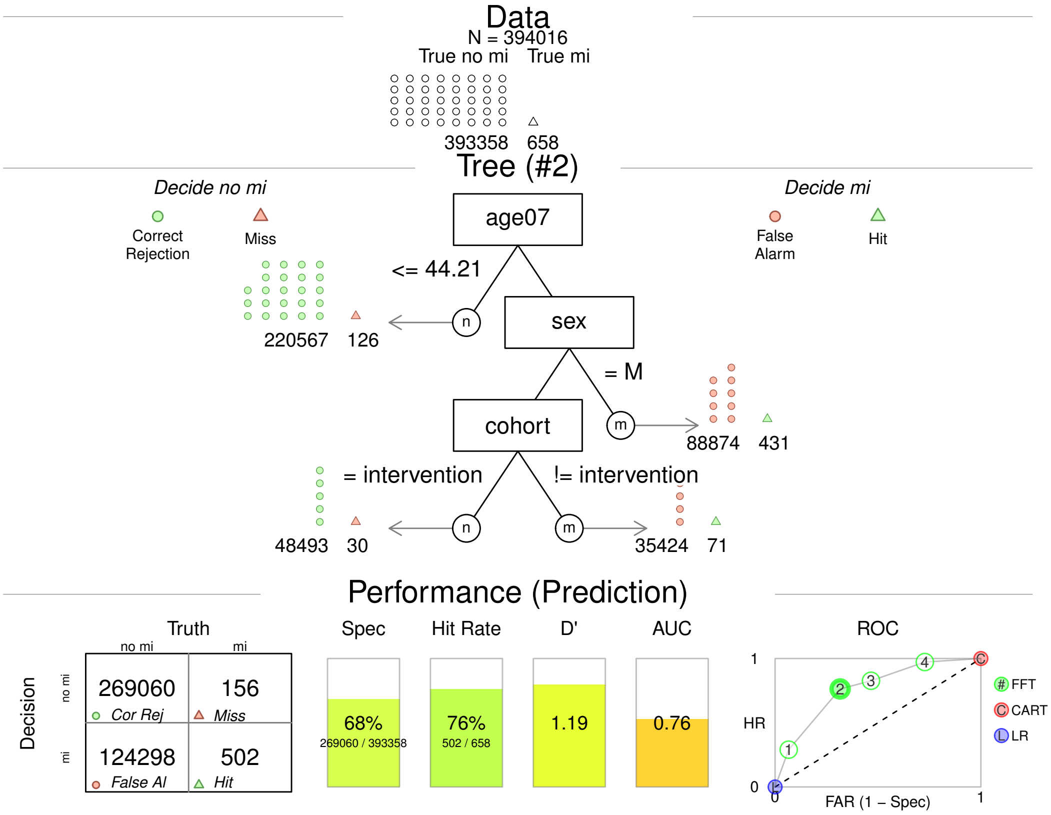


Figure 3.85: Fast and Frugal Tree: *small* model

PPV	NPV	FPR	FNR	FDR	ACC	F1	κ
0	1	0,3	0,2	1	0,7	0,01	0,7

in figure 3.86. It is noteworthy that only the top four levels are included and the cohort variable is missing, as it contributes the least additional discriminatory power to the model. Most performance measures listed in Table 3.54 are significantly improved compared with the reference model. In particular, AUC (0,84), ACC (0,83), and κ (0,83) improved. Other features that affect false positives, such as Precision (0,01), are slightly better but still show very low overall performance. This is most likely a direct result of the unbalanced groups and the rarity of the events being observed.

Table 3.54: FFTree full default: performance measures

TP	FP	TN	FN	AUC	sens	spec
464	67.020	326.338	194	0,8	0,7	0,8

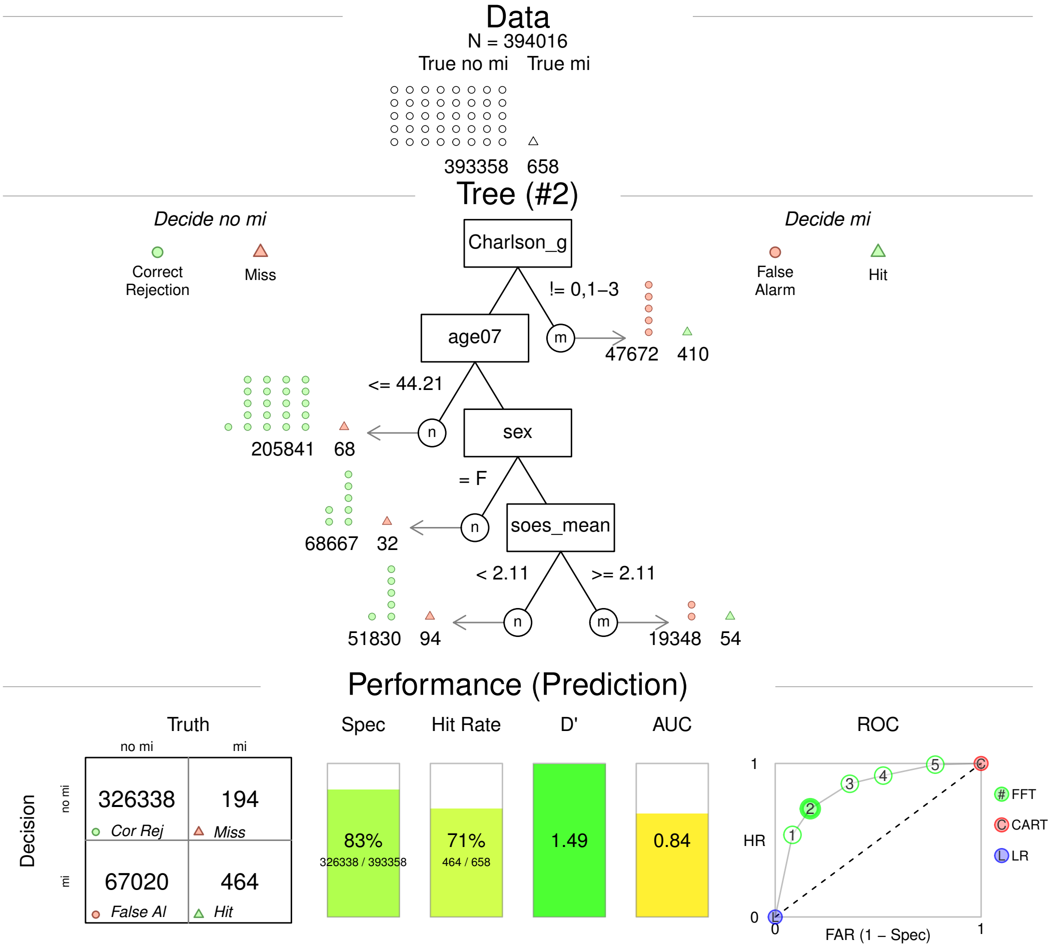


Figure 3.86: Fast and Frugal Tree: *full default* model

PPV	NPV	FPR	FNR	FDR	ACC	F1	κ
0,01	1	0,2	0,3	1	0,8	0,01	0,8

Finally, the full model including cohort assignment is shown as the fifth level in figure 3.87. The decision tree still immediately classifies all individuals in the highest (worst) comorbidity class (of three possible classes) as a positive outcome. This could be due to correlation of the variable with the outcome itself, as the need for medication, which is the source of the estimated comorbidity score, could be the outcome rather than an independent cause of diagnosed myocardial infarction. The next two choices are age and sex. As in the reference model, individuals younger than 45 years and females are classified as unaffected. These two decisions affect a large portion of the data set. The next to last decision is to classify individuals with a socioeconomic status worse than 2,1 as a positive outcome. As a result, males older than 44 years with a low comorbidity index and a low (i.e., better) SES index are additionally classified by their cohort assignment. Interestingly, individuals in the control cohort are labeled as a positive outcome and those in the intervention cohort are labeled as a negative outcome.

Consistent with the previous trees, the performance measures are listed in table 3.55. It can be observed that indicators such as sensitivity (0,76) and FNR (0,24) slightly improve, mostly at the cost of specificity (0,8) and FPR (0,2). The model's general performance measure AUC (0,84) is mostly unchanged, but Kappa (0,8) decreases and is located just between the reference model and the best performing tree of the *full default* model.

This suggests that the variable cohort is only applicable under certain circumstances. In the case where false-positive classification is not as problematic as false-negative prediction, the full model including parenthood information may be preferred over the previous decision tree with only four nodes. In this case, however, the second tree from the reference model shown in figure 3.84 and table 3.52 performs even better and is consensually simpler.

Table 3.55: FFTree full: performance measures

TP	FP	TN	FN	AUC	sens	spec
498	78.560	314.798	160	0,8	0,8	0,8

PPV	NPV	FPR	FNR	FDR	ACC	F1	κ
0,01	1	0,2	0,2	1	0,8	0,01	0,8

Several conclusions can be drawn. Overall, fast and frugal trees appear to be a comprehensive and accessible approach for the task at hand. A modest graphical representation combined with more complex performance indicators allows for a quick and informative data analysis process.

In summary, decision trees with cohort assignment do not perform better than models without this information. Even when overestimation of *positive* outcomes is accepted or intended, simpler models that reflect common knowledge still have an advantage.

Although FFTrees are known to perform well on unbalanced classes, the rarity of myocardial infarction in the study population most likely leads to serious problems. Matching of cohorts and balancing of outcome variables could yield improvements. This bias is not related to the conceptual problems introduced by the selection process that could lead to insufficient accuracy in spouse and parent detection.

Furthermore, a single performance indicator for an entire model is insufficient for FFTree and may even hide important details. In addition to interpreting the entire decision tree, the mixture of the most common measures used in this chapter allows for a quick and comprehensive evaluation of the results.

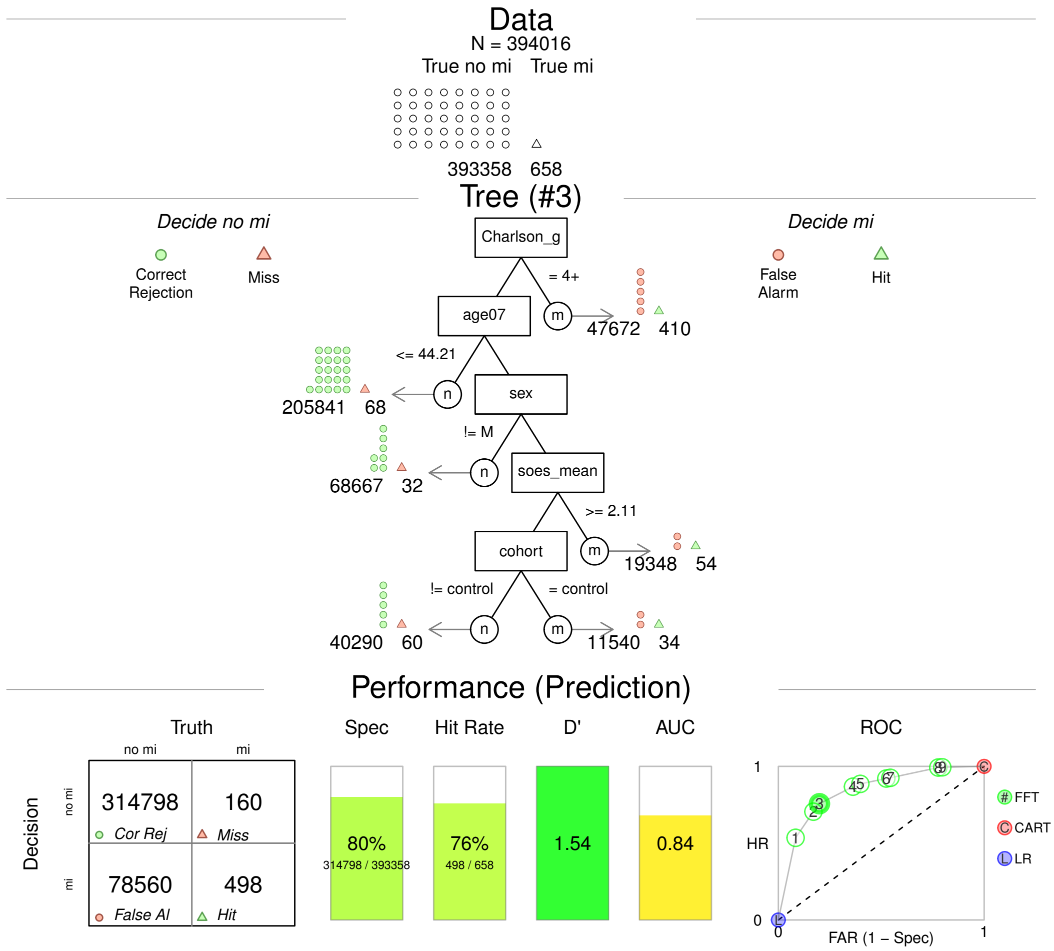


Figure 3.87: Fast and Frugal Tree: *full* model

3.5.5 Gradient Boosting Machine and class imbalance

Generalized Boosted Regression Modeling following Friedman's *Gradient Boosting Machines* (GBM) in conjunction with different approaches to balance the outcome variable are applied next.

Table 3.56 on the following page lists the AUC measures for all resulting models. Although neither the distribution of these values nor a confidence interval is included, it can clearly be stated that all GBM models show a mostly identical performance and correspond in efficiency to the mean and median logistic regression models.

All resulting ROC curves are visualized in figure 3.88 on the next page. Although the optimized viridis color palette [Garnier, 2016] is applied, the individual models are indistinguishable. There are slight variations in detail, but the overall results are virtually

Table 3.56: Area under the ROC curves for GBM models

model	AUC
down 1	0,8763
down 2	0,8729
model 1	0,8729
model 2	0,8711
SMOTE 1	0,8749
SMOTE 2	0,8737
up 1	0,8763
up 2	0,8729
weighted 1	0,8729
weighted 2	0,8737

identical.

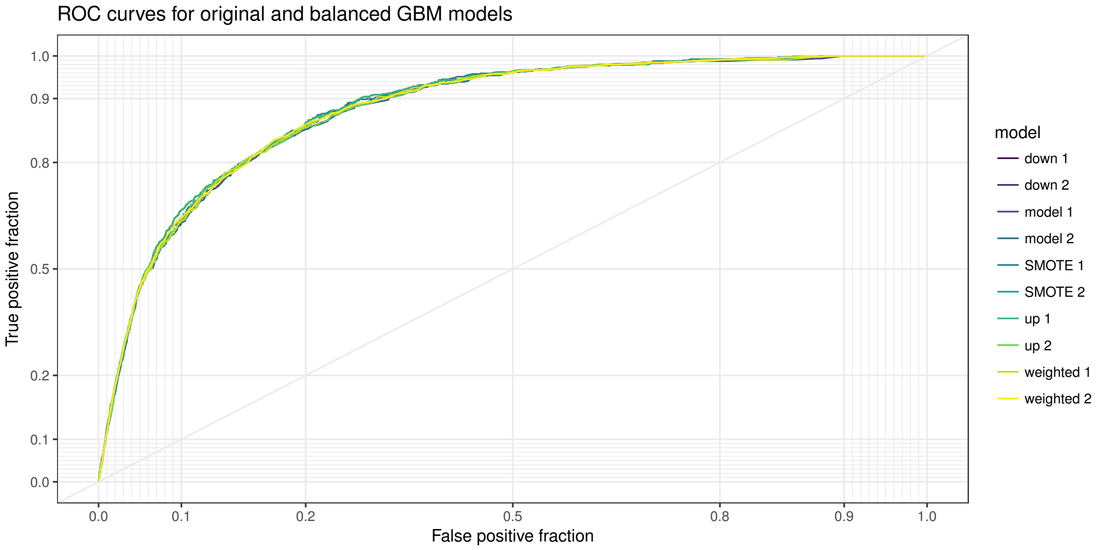


Figure 3.88: Gradient Boosting Machine: ROC curves

When evaluating the GBM models in detail, some differences can be seen in terms of conversion, the necessary and optimal number of boosting iterations, and the relative performance of each iteration. Analysis of the relative relevance of the individual variables yields a familiar picture, with the grouped Charlson score dominating. Overall, these details provide no additional or new insights.

Although no new insights or improved classifiers are obtained in this chapter, the GBM models confirm the results from the previous approaches. It can be concluded that an AUC of slightly above 0,87 is the (current) limit allowed by the available data. Furthermore, the suspicion that the logistic regression models may be biased or overfitted by the unbalanced data set can be ruled out to some extent, as an alternative method yields very similar results.

In summary, GBM is an interesting and straightforward method that can be easily adapted to the problem at hand, benefits from parallelization, and requires no prerequisites. The resulting models show identical performance to logistic regression in this project. A similar conclusion can be drawn for balancing the training data. Evaluating and implementing different methods is worthwhile, but does not significantly change the results for the data at hand.

3.5.6 Propensity score matching

The initial assumption of a positive influence of parenthood on the risk of suffering a myocardial infarction could not be confirmed. Various attempts to obtain unbiased estimators and even to balance the data set according to the rare outcome events showed similar results. This chapter presents results based on matched datasets using propensity score matching.

Tables 3.57 to 3.60 summarize the cohort sizes of the matched cohorts. It can be observed that the cohort *control* is smaller in case replacement is allowed. This could lead to less diversity in the cohort control on the one hand, and more similar matched pairs on the other. Moreover, the additional control for comorbidity leads to slightly more matches.

Table 3.57: Matched cohorts: gender, age, SES without replacement

type	Control	Treated
Matched	260.802	260.802
Unmatched	70.996	983.461

Table 3.58: Matched cohorts: gender, age, SES with replacement

type	Control	Treated
Matched	230.140	1.243.415
Unmatched	101.658	848

Table 3.59: Matched cohorts: gender, age, SES and comorbidity without replacement

type	Control	Treated
Matched	263.313	263.313
Unmatched	68.485	980.950

Table 3.60: Matched cohorts: gender, age, SES and comorbidity with replacement

type	Control	Treated
Matched	232.246	1.243.604
Unmatched	99.552	659

Frequencies per cohort for the two matched datasets are presented in the following tables. Further listings are omitted due to the highly similar results and, as a result, lack of additional information. A univariate comparison of cohorts matched on SES with replacement can be found in table 3.61 on the facing page. The matched cohort *intervention* follows roughly the same distribution in terms of age and gender as the entire unmatched population presented in table 3.42 on page 126. On the other side, cohort *control* differs more markedly in terms of size and distribution. As a possible result of the decrease in median age, the general health status concerning the comorbidity score improved. However, individuals in the control cohort still appear to have a worse comorbidity classification than those in cohort *intervention*.

The number of individuals who suffer myocardial infarction cannot be directly compared in this data set. The relative risk of suffering a myocardial infarction as a parent compared with a member of a childless couple is 0,49. [39]

In contrast, table 3.62 on page 150 compares cohorts matched on SES and comorbidity groups without replacement. It can be observed that age and sex are exactly balanced, as required by exact matching without replacement on these variables. In addition, both the SES and the number of cases in each comorbidity group are very similar. Because of the large cohorts and resulting power, the remaining SES differences still test as significant. Parents appear to have slightly higher (i.e., worse) comorbidity than individuals in the control cohort, which is in drastic contrast to previous observations. Most importantly, the number of observed myocardial infarctions is almost identical in both groups and shows no significant difference. In absolute numbers, there are more individuals experiencing myocardial infarction in the control cohort than in the cohort *intervention*.

In conclusion, there is no evidence that parents have a higher risk than couples without children. Furthermore, it can be concluded that removing the bias regarding age and sex and matching on SES and comorbidity class clearly cancels out any difference regarding the number of actual events compared with the raw statistics on the total population.

Table 3.63 on page 151 compares individuals with and without an observed myocardial infarction in the dataset matched for age, sex, and SES with replacement. Individuals with a recorded incident are significantly older and dominated by men. SES appears to be slightly unfavorable, whereas the grouped Charlson comorbidity index is significantly

[39] the relative risk is calculated as follows: $\frac{\frac{mi_{control}}{no\,mi_{control}}}{\frac{mi_{intervention}}{no\,mi_{intervention}}}$

Table 3.61: Baseline characteristics by cohort: matched by age, sex and SES with replacement. $_a$ b $_c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD. Numbers after proportions are frequencies. Tests used: [1]Wilcoxon test; [2]Pearson test .

	control	intervention	Test Statistic
	$N = 230140$	$N = 1243415$	
age07	$_{39}$ 48 $_{54}$ (47 ± 9)	$_{37}$ 42 $_{47}$ (42 ± 7)	$F_{1,1473553}$=56614, P<0.001[1]
age_stratum : [30,35)	14% (33165)	15% (192031)	χ^2_5=1e+05, P<0.001[2]
[35,40)	11% (25208)	23% (289250)	
[40,45)	12% (26574)	25% (314319)	
[45,50)	18% (41254)	19% (237219)	
[50,55)	23% (53747)	11% (135587)	
[55,60]	22% (50192)	6% (75009)	
sex : F	50% (115999)	50% (623187)	χ^2_1=6, P=0.01[2]
M	50% (114141)	50% (620228)	
partner_sum	$_{1,0}$ 1,0 $_{1,0}$ (1,1 ±0,4)	$_{1,0}$ 1,0 $_{1,0}$ (1,1 ±0,4)	$F_{1,1473553}$=199, P<0.001[1]
soes_mean	$_{1,7}$ 1,9 $_{2,2}$ (2,0 ±0,3)	$_{1,7}$ 1,9 $_{2,1}$ (1,9 ±0,3)	$F_{1,1473553}$=7579, P<0.001[1]
Charlson_group : 0	36% (82365)	39% (485825)	χ^2_2=6590, P<0.001[2]
1-3	48% (111154)	51% (631264)	
4+	16% (36621)	10% (126326)	
mi : FALSE	100% (229529)	100% (1241784)	χ^2_1=231, P<0.001[2]
TRUE	0% (611)	0% (1631)	

worse. Despite all these marked differences, there is no significant difference in cohort assignment.

Table 3.63 on page 151 compares individuals with and without an observed myocardial infarction in the dataset matched for age, sex, and SES with replacement. Individuals with a recorded incident are significantly older and dominated by men. SES appears to be slightly unfavorable, whereas the grouped Charlson comorbidity index is significantly worse. Despite all these marked differences, there is no significant difference in cohort assignment.

As expected, the removal of bias by matching the cohorts on multiple characteristics is consistent and confirms the previous results. Finally, all requirements of the study protocol are met.

Table 3.62: Baseline characteristics by cohort: matched by age, sex, SES and comorbidity without replacement. $_a$ b $_c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD. Numbers after proportions are frequencies. Tests used: ^{1}Wilcoxon test; ^{2}Pearson test .

	control	intervention	Test Statistic
	$N = 263313$	$N = 263313$	
age07	$_{41}$ 50 $_{55}$ (48 ± 9)	$_{41}$ 50 $_{55}$ (48 ± 9)	$F_{1,526624}$=0, P=1^1
age_stratum : [30,35)	13% (33589)	13% (33589)	χ^2_5=0, P=1^2
[35,40)	10% (25229)	10% (25229)	
[40,45)	10% (26612)	10% (26612)	
[45,50)	16% (42742)	16% (42742)	
[50,55)	25% (66787)	25% (66787)	
[55,60]	26% (68354)	26% (68354)	
sex : F	51% (132984)	51% (132984)	χ^2_1=0, P=1^2
M	49% (130329)	49% (130329)	
partner_sum	$_{1,0}$ 1,0 $_{1,0}$ (1,1 ±0,3)	$_{1,0}$ 1,0 $_{1,0}$ (1,1 ±0,4)	$F_{1,526624}$=2502, P<0.001^1
soes_mean	$_{1,7}$ 1,9 $_{2,2}$ (2,0 ±0,3)	$_{1,7}$ 1,9 $_{2,2}$ (2,0 ±0,3)	$F_{1,526624}$=504, P<0.001^1
Charlson_group : 0	35% (92656)	34% (89572)	χ^2_2=672, P<0.001^2
1-3	49% (128696)	47% (124666)	
4+	16% (41961)	19% (49075)	
mi : FALSE	100% (262615)	100% (262667)	χ^2_1=2, P=0.2^2
TRUE	0% (698)	0% (646)	

Table 3.63: Baseline characteristics by outcome: matched by age, sex and SES with replacement. $_a$ b $_c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $\bar{X} \pm 1$ SD. Numbers after proportions are frequencies. Tests used: [1]Wilcoxon test; [2]Pearson test .

	FALSE		TRUE		Test Statistic
	$N = 525282$		$N = 1344$		
age07	$_{41}$ 50 $_{55}$	(48 ± 9)	$_{50}$ 54 $_{57}$	(53 ± 6)	$F_{1,526624}$=521, P<0.001[1]
age_stratum : [30,35)	13%	(67168)	1%	(10)	χ^2_5=494, P<0.001[2]
[35,40)	10%	(50418)	3%	(40)	
[40,45)	10%	(53153)	5%	(71)	
[45,50)	16%	(85313)	13%	(171)	
[50,55)	25%	(133145)	32%	(429)	
[55,60]	26%	(136085)	46%	(623)	
sex : F	51%	(265722)	18%	(246)	χ^2_1=559, P<0.001[2]
M	49%	(259560)	82%	(1098)	
partner_sum	$_{1,0}$ 1,0 $_{1,0}$	(1,1 ±0,4)	$_{1,0}$ 1,0 $_{1,0}$	(1,1 ±0,3)	$F_{1,526624}$=10, P=0.002[1]
soes_mean	$_{1,7}$ 1,9 $_{2,2}$	(2,0 ±0,3)	$_{1,8}$ 2,1 $_{2,3}$	(2,1 ±0,4)	$F_{1,526624}$=105, P<0.001[1]
Charlson_group : 0	35%	(182167)	5%	(61)	χ^2_2=2364, P<0.001[2]
1-3	48%	(252976)	29%	(386)	
4+	17%	(90139)	67%	(897)	
cohort : control	50%	(262615)	52%	(698)	χ^2_1=2, P=0.2[2]
intervention	50%	(262667)	48%	(646)	

CHAPTER 4

Discussion

No evidence for a higher or modified risk of myocardial infarction in parents compared with couples without children is found in the defined cohorts. Most likely, it is not possible to find an actual difference regarding myocardial infarction in the population defined by the study protocol and the available data. Nevertheless, some new observations can be reported and conclusions drawn.

The inference of (family) relationships from administrative claims data of the GAP-DRG database appears to be possible and plausible. After passing several obstacles and technical difficulties, the carefully analyzed results appear to be suitable for specific research questions. Naturally, a margin of error is still present, and the presented method can only be applied in fitting settings. Furthermore, a deeper understanding of the origin of the information is advisable. Extracted co-insurance networks are easily adaptable to other cohort definitions. Nevertheless, the discriminatory power is expected to decrease with growing age. As a result, there might be (older) couples recognized as childless although they are *former* parents.

This new information enabled the discovery of new data quality issues and provided explanations for previously unknown structures and errors, such as the effect of Y2K miscoding. In addition, relationship networks enabled a sophisticated imputation methodology of individual SES. Although a detailed evaluation and justification of this approach is provided, further evaluation is needed to apply the basic strategy in subsequent analyses.

Another innovative development of this project is the comparison and integration of grouped and adjusted comorbidity scores based on the *ATC-ICD*. In general, the presented approach can be directly applied to subsequent research. The resulting comorbidity classification is a significant predictive factor for the outcome variable. A limitation of this interpretation is the lack of temporal relationship between the predictor and the event. Nevertheless, comorbidity, derived from filled prescriptions in combination with age and gender, could be a (more) reliable and conveniently feasible indicator of medical complications.

A key principle of this project is independence from speculation about the database and its content. Consequently, most of the constraints on the extracted dataset are not made solely on the basis of prior knowledge and conjecture, but on actual findings during data exploration. While this approach proved to be very time consuming, it supports the results presented and should make the entire study more reproducible.

The data analysis presented relies heavily on exploratory methods, graphical representations, and a mix of established methods. Various algorithms are applied to deal with highly imbalanced data. Besides the different sizes of the cohorts and their massive differences in age and sex distribution, the very rare and unbalanced outcome events are a major problem. It can be concluded that the defined age range of the cohorts and probably even of the whole database is not ideal for the hypothesis of the project, because most myocardial infarctions are recorded in older individuals, but the correct distinction between parents and childless couples on the basis of the available data is only possible in younger individuals individuals. As a result, there could be a substantial number of misclassification and unrecognized couples in older age strata, which would significantly bias the results.

A retrospective cohort design was chosen for this observational study. It is expected to be superior to its main alternative, a case-control study design, because the influencing factor is not present (i.e., an individual without a spouse) or cannot be measured with the available data for a significant portion of the total population present in the GAP-DRG database. Nevertheless, conceptual issues may remain. While the outcome event, myocardial infarction, is relatively well-defined and can be narrowed to a specific date, being in a relationship or having children is a complex and variable condition. Moreover, the intensity of this influencing factor may vary greatly from person to person, change over time, and even have a delayed and nonlinear effect, which would subsequently be correlated with age. Overall, the mixture of a inexplicit status (e.g., being a parent instead of a childless couple) without the possibility of further segmentation and a unique event could lead to an uncontrolled bias.

The identification of myocardial infarction also cannot be validated with the available data. Most real cases are likely to be treated in a hospital setting, where a clear diagnosis should be documented. However, it is unclear whether such a severe event coded as an additional diagnosis should be interpreted as an error in the hospital or have additional significance. Furthermore, because of the short observation period of 2 years, it is not possible to obtain a real incidence rate and thus treat individuals with a previous myocardial infarction differently. Although a specific diagnosis of subsequent[1] myocardial infarction is coded for some patients, it is not clear whether this occurs in every case in which it would be applicable.

[1] ICD-10 I22: *Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction*

In summary, this project presents several innovations and the derivation of unprecedented information. New insights into administrative claims data from the GAP-DRG database are obtained, but no proof or additional evidence of the initial hypothesis can be found. Nevertheless, a methodological and substantive foundation for subsequent research has been established.

Bibliography

[Arnold et al., 2010] Arnold, B. F., Khush, R. S., Ramaswamy, P., London, A. G., Rajkumar, P., Ramaprabha, P., Durairaj, N., Hubbard, A. E., Balakrishnan, K., and Colford, J. M. (2010). Causal inference methods to study nonrandomized, preexisting development interventions. *Proceedings of the National Academy of Sciences*, 107(52):22605–22610.

[Austin, 2011] Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424.

[Bauer et al., 2015] Bauer, C. R. K. D., Ganslandt, T., Baum, B., Christoph, J., Engel, I., Löbe, M., Mate, S., Stäubert, S., Drepper, J., Prokosch, H.-U., Winter, A., and Sax, U. (2015). Integrated Data Repository Toolkit (IDRT): A Suite of Programs to Facilitate Health Analytics on Heterogeneous Medical Data. *Methods of Information in Medicine*, 54(6).

[Benchimol et al., 2015] Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sørensen, H. T., von Elm, E., Langan, S. M., and RECORD Working Committee (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine*, 12(10):e1001885.

[Böhm et al., 2013] Böhm, M., Gruber, D., Koren, G., Schöny, W., and Endel, F. (2013). *"Wenn die Tür sich dreht": Personenspezifika und Inanspruchnahme ambulanter Leistungen von Psychiatriepatient/innen mit hoher Wiederaufnahmerate in ausgewählten Bundesländern in Österreich.* Pro Mente, Linz.

[Carr et al., 1987] Carr, D. B., Littlefield, R. J., Nicholson, W. L., and Littlefield, J. S. (1987). Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82:424–436.

[Charlson et al., 1987] Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*, 40(5):373–383. bibtex: charlson_index_1987.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[Chini et al., 2011] Chini, F., Pezzotti, P., Orzella, L., Borgia, P., and Guasticchi, G. (2011). Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources. *BMC Public Health*, 11(1):688. bibtex: chini_can_2011.

[Clark et al., 1995] Clark, D. O., Von Korff, M., Saunders, K., Baluch, W. M., and Simon, G. E. (1995). A chronic disease score with empirically derived weights. *Medical Care*, 33(8):783–795. bibtex: clark_chronic_1995.

[Cricelli et al., 2003] Cricelli, C., Mazzaglia, G., Samani, F., Marchi, M., Sabatini, A., Nardi, R., Ventriglia, G., and Caputi, A. P. (2003). Prevalence estimates for chronic diseases in Italy: exploring the differences between self-report and primary care databases. *Journal of Public Health Medicine*, 25(3):254–257. bibtex: cricelli_prevalence_2003.

[Deshmukh et al., 2009] Deshmukh, V. G., Meystre, S. M., and Mitchell, J. A. (2009). Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Medical Research Methodology*, 9(1):70.

[Deyo et al., 1992] Deyo, R. A., Cherkin, D. C., and Ciol, M. A. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *Journal of Clinical Epidemiology*, 45(6):613–619. bibtex: deyo_charlson_1992.

[Drucker, 1997] Drucker, H. (1997). Improving regressors using boosting techniques. In *ICML*, volume 97, pages 107–115.

[Einiö et al., 2015] Einiö, E., Nisén, J., and Martikainen, P. (2015). Is young fatherhood causally related to midlife mortality? A sibling fixed-effect study in Finland. *Journal of Epidemiology and Community Health*, pages jech–2015–205627.

[Elixhauser et al., 1998] Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical Care*, 36(1):8–27. bibtex: elixhauser_comorbidity_1998.

[Endel, 2014] Endel, F. (2014). Understanding data quality in linked administrative data. In *International Health Data Conference 2014*, Vancouver.

[Endel and Duftschmid, 2016] Endel, F. and Duftschmid, G. (2016). Secondary Use of Claims Data from the Austrian Health Insurance System with i2b2: A Pilot Study. *Studies in Health Technology and Informatics*, 223:245–252.

[Endel et al., 2012] Endel, F., Endel, G., and Pfeffer, N. (2012). PRM34 Routine Data in HTA: Record Linkage in Austrias GAP-DRG Database. *Value in Health*, 15(7):A466.

[Endel et al., 2011] Endel, F., Endel, G., Weibold, B., and Katschnig, H. (2011). Health service record linkage in a situation of multiple social health insurance institutions: the case of Austria. In *Proceedings of The methodological challenges of record linkage*, University of St. Andrews, Scotland.

[Endel and Piringer, 2015] Endel, F. and Piringer, H. (2015). Data Wrangling: Making Data Useful Again. In *MATHMOD 2015 Vienna – Abstract Volume*, volume 8, pages 111–112, Vienna University of Technology, Vienna, Austria. ARGESIM and ASIM, German Simulation Society, Div. of GI – German Society for Informatics / Informatics and Life Sciences.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

[Filzmoser et al., 2009] Filzmoser, P., Eisl, A., and Endel, F. (2009). *ATC –> ICD: Determination of the reliability for predicting the ICD code from the ATC code.*

[Freund and Schapire, 1997] Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

[Friedman et al., 2000] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407.

[Friedman, 2001] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232.

[Ganslandt et al., 2011] Ganslandt, T., Mate, S., Helbing, K., Sax, U., and Prokosch, H. U. (2011). Unlocking Data for Clinical Research – The German i2b2 Experience:. *Applied Clinical Informatics*, 2(1):116–127.

[Garnier, 2016] Garnier, S. (2016). *viridis: Default Color Maps from 'matplotlib'*. R package version 0.3.4.

[Gigerenzer and Brighton, 2009] Gigerenzer, G. and Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143.

[Gigerenzer et al., 1999] Gigerenzer, G., Czerlinski, J., and Martignon, L. (1999). How Good are Fast and Frugal Heuristics? In Shanteau, J., Mellers, B. A., and Schum, D. A., editors, *Decision Science and Technology*, pages 81–103. Springer US. DOI: 10.1007/978-1-4615-5089-1_6.

[Gigerenzer and Todd, 1999] Gigerenzer, G. and Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart*, pages 3–34. Oxford University Press.

[Hansen and Klopfer, 2006] Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627.

[Haug et al., 2011] Haug, A., Zachariassen, F., and van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, 4(2):168–193.

[Ho et al., 2011] Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42(8).

[Huber et al., 2013] Huber, C. A., Szucs, T. D., Rapold, R., and Reich, O. (2013). Identifying patients with chronic conditions using pharmacy data in Switzerland: an updated mapping approach to the classification of medications. *BMC public health*, 13(1):1030.

[Jakovljević and Ostojić, 2013] Jakovljević, M. and Ostojić, L. (2013). Comorbidity and multimorbidity in medicine today: challenges and opportunities for bringing separated branches of medicine closer to each other. *Psychiatria Danubina*, 25 Suppl 1:18–28. bibtex: jakovljevic_comorbidity_2013.

[Johnson et al., 2014] Johnson, E. K., Broder-Fingert, S., Tanpowpong, P., Bickel, J., Lightdale, J. R., and Nelson, C. P. (2014). Use of the i2b2 research query tool to conduct a matched case–control clinical research study: advantages, disadvantages and methodological considerations. *BMC medical research methodology*, 14(1):16.

[Johnston et al., 2015] Johnston, M. C., Marks, A., Crilly, M. A., Prescott, G. J., Robertson, L. M., and Black, C. (2015). Charlson index scores from administrative data and case-note review compared favourably in a renal disease cohort. *The European Journal of Public Health*, 25(3):391–396.

[Kandel et al., 2011] Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., and Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288.

[Katschnig et al., 2012] Katschnig, H., Endel, F., Endel, G., Weibold, B., Scheffel, S., and Filzmoser, P. (2012). Dementia and pathways of health services utilization in Austria: A record linkage study in a country with a fragmented provider payment system and only partially available unique patient identifiers. Perth, Australia.

[Klimont et al., 2007] Klimont, J., Kytir, J., and Leitner, B. (2007). *Österreichische Gesundheitsbefragung, 2006/2007: Hauptergebnisse Und Methodische Dokumentation*. Statistik Austria, Wien.

[Kuhn, 2008] Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26.

[Logue et al., 2016] Logue, E., Smucker, W., and Regan, C. (2016). Admission Data Predict High Hospital Readmission Risk. *The Journal of the American Board of Family Medicine*, 29(1):50–59.

[Loh, 2011] Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23.

[Maio et al., 2005] Maio, V., Yuen, E., Rabinowitz, C., Louis, D., Jimbo, M., Donatini, A., Mall, S., and Taroni, F. (2005). Using pharmacy data to identify those with chronic conditions in Emilia Romagna, Italy. *Journal of Health Services Research & Policy*, 10(4):232–238. bibtex: maio_using_2005.

[Marewski and Gigerenzer, 2012] Marewski, J. N. and Gigerenzer, G. (2012). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, 14(1):77–89.

[Martignon et al., 2008] Martignon, L., Katsikopoulos, K. V., and Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, 52(6):352–361.

[McCormick and Joseph, 2016] McCormick, P. and Joseph, T. (2016). *medicalrisk: Medical Risk and Comorbidity Tools for ICD-9-CM Data*. R package version 1.2.

[McHugh, 2012] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

[Moher et al., 2009] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7):e1000097.

[Murphy et al., 2006] Murphy, S. N., Mendis, M. E., Berkowitz, D. A., Kohane, I., and Chueh, H. C. (2006). Integration of clinical and genetic data in the i2b2 architecture. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1040. American Medical Informatics Association.

[Murphy et al., 2010] Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., and Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130.

[Natekin and Knoll, 2013] Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7.

[Nicholls et al., 2015] Nicholls, S. G., Quach, P., von Elm, E., Guttmann, A., Moher, D., Petersen, I., Sørensen, H. T., Smeeth, L., Langan, S. M., and Benchimol, E. I. (2015). The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for Arriving at Consensus and Developing Reporting Guidelines. *PloS One*, 10(5):e0125620.

[Pearson, 1900] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175.

[Porta, 2014] Porta, M. S., editor (2014). *A Dictionary of Epidemiology*. Oxford University Press, Oxford ; New York, 6th edition.

[Quan et al., 2011] Quan, H., Li, B., Couris, C. M., Fushimi, K., Graham, P., Hider, P., Januel, J.-M., and Sundararajan, V. (2011). Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *American Journal of Epidemiology*, 173(6):676–682. bibtex: quan_charlson_2011.

[Quan et al., 2005] Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., Beck, C. A., Feasby, T. E., and Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, 43(11):1130–1139.

[Robin et al., 2011] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

[Rosenbaum and Rubin, 1983] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[Rubin, 2004] Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*, 13(12):855–857.

[Safford et al., 2007] Safford, M. M., Allison, J. J., and Kiefe, C. I. (2007). Patient Complexity: More Than Comorbidity. The Vector Model of Complexity. *Journal of General Internal Medicine*, 22(Suppl 3):382–390.

[Schapire, 1999] Schapire, R. E. (1999). A Brief Introduction to Boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pages 1401–1406, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Schapire, 2003] Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer.

[Sedgwick, 2013] Sedgwick, P. (2013). Selection bias versus allocation bias. *BMJ*, 346(may24 4):f3345–f3345.

[Segagni et al., 2011] Segagni, D., Ferrazzi, F., Larizza, C., Tibollo, V., Napolitano, C., Priori, S. G., and Bellazzi, R. (2011). R Engine Cell: integrating R into the i2b2 software infrastructure. *Journal of the American Medical Informatics Association : JAMIA*, 18(3):314–317.

[Sharabiani et al., 2012] Sharabiani, M. T. A., Aylin, P., and Bottle, A. (2012). Systematic review of comorbidity indices for administrative data. *Medical Care*, 50(12):1109–1118. bibtex: sharabiani_comorbidity_2012.

[Sing et al., 2005] Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941.

[Starfield, 2006] Starfield, B. (2006). Threads and Yarns: Weaving the Tapestry of Comorbidity. *The Annals of Family Medicine*, 4(2):101–103.

[Stausberg et al., 2015] Stausberg, J., Nasseh, D., and Nonnemacher, M. (2015). Measuring Data Quality: A Review of the Literature between 2005 and 2013. *Studies in Health Technology and Informatics*, 210:712–716.

[Stuart, 2010] Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1–21.

[Sundararajan et al., 2004] Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H., and Ghali, W. A. (2004). New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *Journal of Clinical Epidemiology*, 57(12):1288–1294.

[Sundararajan et al., 2007] Sundararajan, V., Quan, H., Halfon, P., Fushimi, K., Luthi, J.-C., Burnand, B., Ghali, W. A., and International Methodology Consortium for Coded Health Information (IMECCHI) (2007). Cross-national comparative performance of three versions of the ICD-10 Charlson index. *Medical Care*, 45(12):1210–1215.

[Tang et al., 2015] Tang, W., Hu, J., Zhang, H., Wu, P., and He, H. (2015). Kappa coefficient: a popular measure of rater agreement. *Shanghai Archives of Psychiatry*, 27(1):62–67.

[Tennekes et al., 2013] Tennekes, M., de Jonge, E., Daas, P. J., and others (2013). Visualizing and inspecting large datasets with tableplots. *Journal of Data Science*, 11(1):43–58.

[Toson et al., 2016] Toson, B., Harvey, L. A., and Close, J. C. T. (2016). New ICD-10 version of the Multipurpose Australian Comorbidity Scoring System outperformed Charlson and Elixhauser comorbidities in an older population. *Journal of Clinical Epidemiology*. bibtex: toson_index_2016.

[Tukey, 1977] Tukey, J. W. (1977). *Exploratory Data Analysis*. Behavioral Science: Quantitative Methods. Addison-Wesley, Reading, Mass.

[Valderas et al., 2009] Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., and Roland, M. (2009). Defining Comorbidity: Implications for Understanding Health and Health Services. *Annals of Family Medicine*, 7(4):357–363. bibtex: valderas_defining_2009.

[van den Akker et al., 1996] van den Akker, M., Buntinx, F., and Knottnerus, J. A. (1996). Comorbidity or multimorbidity: what's in a name? A review of literature. *European Journal of General Practice*, 2(2):65–70.

[van Walraven et al., 2009] van Walraven, C., Austin, P. C., Jennings, A., Quan, H., and Forster, A. J. (2009). A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical Care*, 47(6):626–633. bibtex: van_walraven_elixhauser_2009.

[Vandenbroucke et al., 2014] Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., and STROBE Initiative (2014). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *International Journal of Surgery (London, England)*, 12(12):1500–1524.

[Venables and Ripley, 2002a] Venables, W. N. and Ripley, B. D. (2002a). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

[Venables and Ripley, 2002b] Venables, W. N. and Ripley, B. D. (2002b). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

[Vision, 1985] Vision, N. R. C. U. C. o. (1985). *Appendix B: Detection sensitivity and response bias*. National Academies Press (US).

[von Elm et al., 2007] von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., and STROBE Initiative (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS medicine*, 4(10):e296.

[Von Korff et al., 1992] Von Korff, M., Wagner, E. H., and Saunders, K. (1992). A chronic disease score from automated pharmacy data. *Journal of Clinical Epidemiology*, 45(2):197–203. bibtex: von_korff_chronic_1992.

[Wasey, 2016] Wasey, J. O. (2016). *icd: Tools for Working with ICD-9 and ICD-10 Codes, and Finding Comorbidities*. R package version 2.0.1.

[Weinlich, B. et al., 2014] Weinlich, B., Mate, S., Prokosch, H.U., Ganslandt, T., and Toddenroth, D. (2014). "R-Scriptlets" für i2b2-Endanwender.

[Wickham, 2009a] Wickham, H. (2009a). *Ggplot2: Elegant Graphics for Data Analysis*. Springer Science & Business Media, New York.

[Wickham, 2009b] Wickham, H. (2009b). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[Wickham, 2010] Wickham, H. (2010). A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics*, 19(1):3–28.

[Wickham, 2011] Wickham, H. (2011). Ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2):180–185.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.