The Technical University of Denmark

Bachelor Project
Spring 2018

Data analysis of the link between magnesium in drinking water and mortality

- with specific focus on cardiovascular diseases

Charlotte Friis Theisen s143922

Supervisor: Bjarne Kjær Ersbøl

External supervisors: Annette Kjær Ersbøll
Kirstine Wodschow

08.07.2018

Abstract

The association between magnesium in drinking water and the risk of cardiovascular death has been examined in many studies but never in a Danish context prior to this project. Some evidence of a protective effect of drinking water rich in magnesium is found in these studies. In this epidemiological study, register based data is used along with water samples taken during the past 37 years. The study is designed as a cohort study with a 10-year study period (2005-2014) and includes the entire Danish population aged 30 or more. A Poisson regression model for incidence rates was used to assess the association and included confounders on age, gender, cohabitation and family income as well as adjustment for calendar year. The results showed a significant protective effect of magnesium in drinking water on ischemic heart disease (IHD) and particularly acute myocardial infarct (AMI). The 20% least exposed ($\leq$ 6.65 mg/l), had an increased risk of 24% of dying from AMI compared to the 20% most exposed ($>$ 21.9 mg/l). However, no association was found between the level of magnesium in drinking water and overall cardiovascular death or death from stroke. Further extensive sensitivity analysis has to be carried out to confirm the found association.

Contents

Abstract **1**

1 Introduction **3**

2 Background **5**
2.1 Magnesium and drinking water 5
2.1.1 Magnesium in the ground 5
2.1.2 Recommended intake 5
2.1.3 Actual Intake 6
2.1.4 Magnesium deficiency - consequences 7
2.1.5 Magnesium through drinking water 7
2.2 Relevant studies and literature 7
2.3 Water Softening 11

3 Data **12**
3.1 Data collection 12
3.2 Raw data description 12
3.2.1 Data from GEUS 12
3.2.2 Data from registers 13

4 Methods **15**
4.1 Methods for the magnesium data 15
4.1.1 K Nearest Neighbours 15
4.1.2 Geographical interpolation 17
4.1.3 Linear interpolation 18
4.2 Study design 18
4.3 Statistical methods 22
4.3.1 Incidence rates 22
4.3.2 Introduction of Poisson regression of incidence rates 23
4.3.3 Multiple Poisson regression 23

5 Analysis and results **26**
5.1 Data preprocessing 26
5.2 Descriptive analysis of the Magnesium data 30
5.3 Estimation of magnesium levels 32
5.3.1 Linear Interpolation 33
5.3.2 Geographical interpolation 33
5.3.3 The KNN method 33
5.3.4 The data set of estimations 35
5.4 Descriptive analysis of the final data set 37
5.4.1 The confounding effect of age on gender 39
5.4.2 Subcategories of cardiovascular deaths 40

5.5 Statistical analysis 41
5.5.1 Sensitivity analysis 43

6 Discussion **46**
6.1 The results 46
6.1.1 Validity of results 46
6.2 Strengths of the present study 47
6.3 Limitations of the present study 48
6.4 Magnesium estimates 49
6.5 Addresses linked to WSAs 50
6.6 The perspectives of the study 50

7 Conclusion **51**

Appendices **56**

A SAS example code **57**

B Maps **58**

C Incidence rates of cohabitation per age category **60**

Chapter 1

Introduction

Almost everybody is concerned with their health to some extent. Most of us try our best to eat well, sleep well, exercise enough and in general follow the recommendations that will benefit our health. The recommendations are based on researchers finding associations between exposures and risk of all sort of diseases or death. But what about our drinking water? The water that runs in pipes of every household, the water that ends up on the table for dinner or in the tea kettle for breakfast. Does that impose any benefit to our health?

The answers to that question require substantial research and the answers are not yet cut in stone even though research has been going on for more than half a century.

The purpose of this epidemiological study is to contribute to the pool of knowledge that in the end will lead to official recommendations and perhaps even legislation regarding the water quality. This study will focus on one mineral that is part of the drinking water mineral composition. This mineral is magnesium.

Magnesium has already been studied in much research, but never has its potential health benefit been studied in a Danish context. The country of study might play a role in the findings since the level of magnesium in drinking water varies greatly across countries and even within a country.

The hypothesis that lays a ground for the study is the following:

Magnesium in drinking water has a positive effect on the risk of cardiovascular death.

The aim of the study is to find any evidence for or against the hypothesis. It will be investigated if the risk of dying from different sub-categories of cardiovascular disease is effected by the exposure of magnesium through drinking water.

The report describes the process of conducting the study and is structured as follows. The chapter following the introduction is a background chapter that describes various aspects of magnesium and drinking water. It also contains a review of recent studies examining the effect of magnesium in drinking water on cardiovascular death. The third chapter will describe the data available for the study and serves as a documentation of data. The fourth chapter contains all the methods used. This includes data science methods for estimating magnesium levels, the entire study design and a description of the statistical method used for the final analysis. The fifth chapter describes all analysis and results, starting with a preprocessing section that documents all handling of data. Furthermore, it has a descriptive analysis, a section on estimation of

magnesium levels and the results from the statistical analysis. The report will end with a discussion of methods validity of results, suggestions of further analysis and consequences of the results. Finally, a conclusion sums up the report.

Chapter 2

Background

2.1 Magnesium and drinking water

Magnesium is a chemical element with the symbol Mg and it exists naturally in its oxidation state Mg^{2+}. Magnesium occurs in combination with other materials in the ground and therefore ground water also contains small amounts of the magnesium ions. According to WHO, the mean concentration in ground water in different parts of the world is 20 ± 13 mg/l [1]. The actual concentrations in Denmark will be examined later in the report. First, a chapter on the background of magnesium in drinking water will justify why the hypothesis stated in the introduction is interesting to study.

2.1.1 Magnesium in the ground

The amount of magnesium that naturally exists in the ground water depends on the type of aquifer from which the water has been abstracted. The aquifer is the layer of the ground in which ground water is found. In an article by Kirstine Wodschow et al. [2] it is shown how the relation between the type of aquifer and the level of magnesium in drinking water samples is significant. This is shown in Denmark using the same data set as will later be used in this study. It is furthermore shown that there exists clusters in Denmark in which the concentration of magnesium is significantly higher (or lower) than in the rest of the country. The cluster with low concentrations is found in the central Jutland and the cluster with high levels are found on Sjælland and Lolland-Falster.

2.1.2 Recommended intake

In order to understand the possible impact of magnesium through drinking water it is necessary to understand how much magnesium humans are supposed to get every day. A small review of some international recommendations for magnesium intake will briefly be given. In general, recommendations are estimated using different terms to reflect the method used. Below, some of the common values are listed:

- Estimated Average Requirement (EAR): The EAR is the estimation of the nutrient value that meets the requirements in 50% of the individuals. The requirements are defined by a specified indicator.
- The Recommended Dietary Allowance (RDA): The intake level that meets the requirements in almost all individuals and it is estimated from the EAR.
- Adequate Intake (AI): An estimate of the average intake in a healthy group of the population.

The Estimated Average Requirement (EAR) is by The United States Department of Agriculture estimated to be 350 mg/day for men and 265 mg/day for women over 50. It is slightly lower for younger individuals. The Recommended Dietary Allowance (RDA) is estimated to be 400 mg/day for men and 310 mg/day for women [3].

The European Food Safety Agency estimated an Adequate Intake (AI) to be 350 mg/day for men and 300 mg/day for women [4].

The estimates are somewhat similar and it can be concluded that according to international recommendations a daily intake of more than 400 mg for men and 300 mg for women should be sufficient.

2.1.3 Actual Intake

The summarised results of studies examining the actual intake through the diet in different countries are listed in Table 2.1.

Country		**Magnesium intake (mg/day)**
USA	Men	mean: 268, total range: 50.3-1,138
	Women	Low group, mean: 255 High group, mean: 433
Canada	Men	mean±SD: 402±169
	Women	mean±SD: 307±123
France	Men	mean±SD: 377±114
	Women	mean±SD: 284±99
Spain	All	mean: 366
Sweden	All	mean: 330

Table 2.1: Table showing reported daily intakes of magnesium from food in different countries. In Sweden and Spain market basket analysis was conducted, in the other countries a cohort was followed and their diet analysed. Sources: [5, 6, 7, 8, 9, 10]

The studies from USA, Canada and France are based on the actual food consumption of a small study population. For the American study on women, the population was split into five equally sized groups based on their magnesium consumption. The mean of the groups with the lowest and highest intake are reported in the table. For the Spanish and Swedish study, market basket analysis was done and an average intake based on a regular diet was calculated. In those studies, no distinction between men and women was made.

If the values in Table 2.1 are seen in relation to the mentioned recommendations, it indicates that part of the population probably gets close to a sufficient amount of magnesium through their diet. The average in Canada is very close to the recommended intakes and in France they are only slightly lower. In Spain and Sweden the analysis shows a mean that also seems reasonable since it is the mean calculated for both men and women. In USA men seems to get too little magnesium on average and for the female population the lowest percentiles gets too little whereas the highest gets more than enough. For all the studies showing reasonable means it should be noted that the large standard deviations indicate that a large subgroup of the population gets much less than the recommendations. Unfortunately, similar estimates of

the Danish population could not be found.

2.1.4 Magnesium deficiency - consequences

The consequence of magnesium deficiency is manifold and includes hypertension, cardiac arrhythmias and ischemic heart disease. Magnesium is a vital part of the body's system as it is a cofactor for more than 300 enzymes, in particular it is involved in the metabolism for the synthesis of lipids, carbohydrates, nucleic acids and proteins. It is also vital in some organs, for example the cardiovascular system. Furthermore, it is present in the bone structure [4, 11, 12, 13].

2.1.5 Magnesium through drinking water

As mentioned in the beginning of this chapter, magnesium is a common element in ground water. For surface water the concentrations are often lower, WHO has estimated the mean concentration of surface water to be half the mean concentration of ground water [1]. In Denmark almost all tap water originates from ground water [14]. Consequently, part of most peoples daily intake of magnesium comes through drinking water. One study suggest a daily intake through drinking water to be $12 \pm 9.8\%$ of the total intake [15]. Of course this is dependent on the concentrations of magnesium in the local water and the amount of water consumed.

The bioavailability of an element refers to how well the body absorbs the element. The ion structure that the magnesium appears as in water (dissolved) is suggested to be more easily absorbed than magnesium from food [16, 17]. This could potentially make drinking water an even more important source of magnesium.

One study also shows that the magnesium in drinking water is absorbed significantly better when consumed with a meal [18].

The amount of tap water consumed is also likely to be related to the amount of bottled water consumed. Thus, if a population consumes much bottled water it might consume less tap water. In Denmark the amount of bottled water is relatively low compared to other European countries. Only the Scandinavian neighbours surpass Denmark on low consumption of bottled water. It is estimated that the consumption of bottled water over the past years has been around 20 l per person per year, which is equivalent to around 50 ml per day. However, the consumption of soft beverages in general (e.g. soft drinks, bottled water and juices) is estimated to be around 450 ml per day [19, 20].

When cooking food in water, the food loses part of its magnesium content. The loss of magnesium from food has an inverse relationship with the concentration of magnesium in the water. The higher the magnesium concentration of the water, the less magnesium is lost from the food [21].

Furthermore, it should be noted that the magnesium in the water used for brewing tea and coffee is barely affected by the boiling process [17].

2.2 Relevant studies and literature

Drinking water and its impact on public health has been studied for more than half a century in epidemiological studies [22]. Many studies from all over the world have focused on different

chemical elements of the water. Of interest for this project is of course mostly the studies examining magnesium or more generally the water hardness. Some of the more recent studies will be described in the following part and summarised in Table 2.2. It should be noted that no study related to the hypothesis from the introduction has been carried out in Denmark prior to this project.

Three Swedish case-control studies showed a significant protective effect of particularly magnesium in drinking water. The first study dates back to 1991 and estimates a correlation coefficient between the risk of different cardiovascular diseases and the amount of magnesium in drinking water. The coefficient is proven to be significantly negative for ischemic heart disease (IHD) but not for cerebrovascular disease including stroke [23]. Thus, the risk of IHD is reduced as the magnesium concentration is increased. The two following Swedish studies are part of the same study. One showing a significant effect of magnesium in men the other in women. The effect is greater in men than in women with odds ratios between highest and lowest exposure group of 0.65 and 0.70 respectively [24, 25].

In one Swedish case-control study there was, however, not found a significant relationship. In this study the actual consumption of drinking water per individual was assessed, and the highest consumption of magnesium from drinking water was registered as less than 4 mg/day, which is very low compared to the recommended intakes [26].

A Spanish study showed a significant relation between hypertension and magnesium in drinking water with an odds ratio of 3.61 between the least exposed and the most exposed. They also showed a significant relation between magnesium and cardiocvascular death (CD)[27].

A Taiwan case-control study showed a significant protective effect of high magnesium concentrations on cerebrovascular disease [28].

One cohort study from the Netherlands with a ten-year follow-up (1986-1996) showed almost no significant relationship between the magnesium in drinking water and various cardiovascular diseases. In this study they took multiple cofactors into account through a survey in the beginning of the study period. The highest exposed group is exposed to between 8.5 and 26.2 mg/l with an average around 10 mg/l [29].

A Finnish case-control study also showed a significantly higher relative risk of acute myocardial infarction (AMI) in individuals exposed to low magnesium concentrations, defining low as less than 1.2 mg/l. However, many results in the study were not significant e.g. the difference between the mean concentrations of cases versus controls. It should also be noted that the study only included 58 matched cases and controls [30].

Three ecological studies have also been assessed in conjunctions with this project. An English, a French and a Japanese study. The English and Japanese studies found no significant protective effects of magnesium [31, 32]. The French study found a slight protective effect on both IHD and cerebrovascular disease [33].

More detailed information on how the studies were carried out and their most interesting results can be found in table 2.2.

Authors	Type	Description	Measurement methods	Amount interval	Analysis method	Association found	Confounders considered
R. Rylander, H. Bonevik, E. Rubenowitz (1991)	Case-control study	Sweden, 27 municipalities. 1969-1978. Men and women	Water samples and confounder data at municipality level	0.57 mg/l - 15.0 mg/l	Special test designed for problems with environmental factors and poison distriutions.	Correlation coeff. (RR and mg/l) for Ischemic Heart Disease IHD: Men -0.618 (-0.806 - -0.319); Women -0.452 (-0.706 - -0.095) No significant correlation coefficient was found for the association between rleative risk and magnesium in drinking water for cerebrovascular diseases.	Age and gender
Rubenowitz, E. ,Axelsson, G. og Rylander, R. (1996)	Case-control study	Sweden, 17 municipalities in Skåne and Blekinge regions. 854 cases and 989 controls. Men only.	Waterworks connection to each individual through one year prior to death. Confounder data on individual level.	1.3 mg/l-20.0 mg/l	Logistic regression for unconditional maximum likelihood estimation.	Odds ratio: 0.65 (95% CI: 0.51-0.86) for areas with highest magnesium concentration comparedd to area with lowest.	Age and calcium
Rubenowitz, E., Axelsson, G. and Rylander, R. (1999)	Case-control study	Sweden, 16 municipalities in Skåne and Blekinge regions. 378 cases and 1378 controls. Women only.	Waterworks connection to each individual at last known adress. Confounder data on individual level.	1.3 mg/l - 21.5 mg/l		Odds ratio: 0.70 (95% CI: 0.50-0.99) between areas with highest magnesium concentration and areas with lowest for AMI.	Age and calcium
Rosenlund, Mats (2005)	Case-control study	Sweden (Stockhom), 497 cases and 677 controls. 1992-1994. Men and women.	Waterworks connection to individual + questionnaire on daily consumtion of tap water.	Exp. 1: <0.9mg/day, Exp. 2: 0.9-1.9mg/day, Exp. 3: 1.9-3.5 mg/day, Exp. 4: >3.5 mg/day	Logistic regression model	Small, but unsignificant, protective association for AMI.	Age, gender, hospital catchment area, smoking, socioeconomic status, hypertension, job strain, diabetes, BMI, physical inactivity
Gimeno-Ortiz, A., Jiménez Romano, R., Blanco Aretio, M. and Castillo Moreno, A. (1990)	Case-control study	Spain, 60 localities, 1254 cases and 4160 controls.	Water analyzed at localities level.			Significant relation between hypertension and Mg with OR: 3.61 (3.17-4.08). Also significant between cardiovascular death and Mg.	

Table 2.2: Continues...

Authors	Type	Description	Measurement methods	Amount interval	Analysis method	Association found	Confounders considered
Yang, Chun-Yuh (1998)	Case-control study	Taiwan, 17133 cases and 17133 controls. 1989 through 1993. Men and women.	Waterworks connection to municipalities (252) in year 1990. Confounder data on individual level.	Lowest mean: 3.8 mg/l, Highest mean: 17.3 mg/l	Conditional logistic regression	Odds ratio: 0.60 (95% CI: 0.52-0.70) between areas with highest concentration and areas with lowest for cerebrovascular diseases.	Gender, age, urbanization level and calcium
Leurs, L. et al. (2010)	Cohort study	Netherlands, 1986-1996, 120,852 men and women aged 55–69	Water hardness information on postal code level. Questionnaire at individual level.	Exp. 1: 1.7-3.8 mg/l Exp. 2: 4.2-6.0 mg/l Exp. 3: 6.0-8.0 mg/l Exp. 4: 8.0-8.2 mg/l Exp. 5: 8.5-26.2 mg/l	Cox proportional hazards model	No significant associations.	Lifestyle, diet, hypertension, gender.
Luoma, H., Aromaa, A., Helminen, S.,Murtomaa, H., Kiviluoto, L., Punsar, S., Knekt, P. (1983)	Case-control study	Finland, 58 cases, 58 hospital controls and 58 population controls. 1974-1975. Men.	Water samples from all individuals.	1 mg/l - 57.5 mg/l	Statistical method for matched case control described by Miettinen.	Few significant results, including RR between low and high exposure for AMI: 4.67 (1.30-25.32) (with population control).	
Maheswaran, R. (1999)	Ecological study	North west England, 1990-1992, population of 2,499,659 aged 45+, men and women	305 water supply zones. Confounder data on enumeration district level (13,794)	2 mg/l - 111 mg/l, mean: 19 mg/l, median: 12 mg/l	Log linear poisson regression.	No significant association between magnesium and AMI or IHD.	Age, gender, socioeconomic deprivation and lead, calcium and flouride in drinking water.
Sauvant, M. and Pepin, D. (2000)	Ecological study	France, department of Puy de dome, 1988-1992, population of 598,493, all ages, men + women.	Water analysis on cantons level (52 cantons).		Log linear regression.	Small but significant correlation coefficient between standardized mortality ratios and water hardness for IHD and cerebrovascualr diseases.	No confounders.
Miyake, Y. and Iki, M. (2003)	Ecological study	Japan. 1995. Population of 8,800,000.	Water supply on municipality level.	35.2 mg/l - 100 mg/l	Multiple logistic regression model.	Unsignificant association for cerebrovascular diseases.	Socioeconomi status (tax rates)

Table 2.2: Overview of studies connecting magnesium in drinking water to cardiovascular diseases

2.3 Water Softening

The hardness of water is calculated from the total amount of dissolved calcium and magnesium content and it is measured in degrees of hardness. The hardness concerns the total amount of the two elements and does not tell anything about the balance of them. The process of softening water involves removing these two ions from the water. They could be removed completely, but would often just be reduced.

In Denmark only a few waterworks are at the moment using water softening techniques with the municipality of Brøndby as the first and only place it has yet been introduced. However, HOFOR (the water supply company of the Copenhagen area) has already planned the introduction of water softening in many municipalities [34].

Water softening is good for many things in the household, such as a longer lifetime of many machines, less use of soap and less cleaning of calcium deposits in bathrooms. In 2011 a report from COWI A/S requested by the Ministry of Environment and Food of Denmark was developed. This report examined all the potential benefits and costs of water softening and concluded that it would be a financial benefit to reduce the water hardness centrally at the waterworks.

Many techniques for reducing the hardness of water exists and they all have different costs and consequences. However, common to all of them is that they remove a large part of the magnesium content. Because of this, the results of the present study can contribute to the discussion of potential consequences of reducing water hardness. In the COWI-report the potential risk of increasing cardiovascular deaths when removing magnesium from the water is acknowledged but not taken into account in the calculations [35].

Chapter 3

Data

In this chapter the data available for the project will be described. This includes how it was collected and which attributes it contains.

3.1 Data collection

The data used throughout this project originates from two different sources. One data set is created on basis of data extracted from the geological survey of Denmark and Greenlands (GEUS) database called Jupiter. In this database, information concerning the entire Danish water supply is stored along with detailed information on every water sample analysed (extractions contain samples back to 1980). The data from this source is publicly available and was extracted by GEUS in July 2017 by the request of Kirstine Wodschow.

The second part of the data used in this project is extracted from Danish health and health related registers at Statistics Denmark. Due to security the data is accessed through their servers.

The two data sources can be linked through two extra sources of information. This includes information on the geographical shape of all water supply areas and information about which area each waterworks supplies. This information was established during a study of the Danish drinking water by Jörg Schullehner in 2014 [36] and has been slightly modified by Kirstine Wodschow recently. The second extra source is the geographical coordinates of all Danish addresses. This data was extracted from *Styrelsen for Dataforsyning og Effektivisering - Adresse Web Services (AWS)* [37] on the 16th of May 2018 by Kirstine Wodschow.

An overview of exactly which data was accessible from the two main sources will now be given. The two extra sources of information, the geographical shapes of all WSAs and the coordinates of addresses, are not described in details.

3.2 Raw data description

3.2.1 Data from GEUS

The data from GEUS is data on water samples measuring magnesium concentrations and data on water abstraction for each waterworks.

Attributes of magnesium samples:

Sample_ID: A unique ID to identify each sample.
WSA_ID: A unique ID identifying a water supply area.
X_centroid: X-coordinate in UTM format locating the center of the WSA.
Y_centroid: Y-coordinate in UTM format locating the center of the WSA.
Waterworks_ID: An ID that uniquely identifies each waterworks.
Amount: The concentration of magnesium measured.
Date: The date at which the concentration of magnesium was measured.

Attributes of the abstraction data:

Waterworks_ID: An ID that uniquely identifies each waterworks.
Abstraction: The amount of water that was abstracted in cubic meters.
Year: The year in which the abstraction was made.

3.2.2 Data from registers

The data available for this project on the servers of Statistics Denmark comes from the Danish registers and the extraction of data was made as part of a larger project. It should be noted that only part of the attributes that exists in each register was made available for the project and only those available and relevant will be described in the following.

The Danish Civil Registration System
The Danish Civil Registration System (CRS) contains personal information about each individual in Denmark. It includes individuals immigrating to Denmark. It is structured in such a way that each year a new data set is created containing the current information of each person. This means that the information in the register is the information valid on the first of January the given year [38]. The register contains information about the address of the individual, the birthday, the family relations and the cohabitation status [39].

Attributes of the CRS data set:

PNR: Encrypted CPR-number.
Opgikom/bobikom: Encrypted address information.
Kom: Municipality code.
DateOfBirth: The date of birth.
Gender: 1 for male, 2 for female.
Age: The age at the end of previous year.
$\mathbf{Family_{id}}$: Identification of the family that the individual belongs to.
$\mathbf{Family_{type}}$: The type of family. Defined by 1 for married couples, 2 for registered partners, 3 or 4 for couples living together and 5 for individuals living alone. This is also referred to as the cohabitation status [40].
Year: Attribute created when all data sets were merged to identify in which year the information was collected.

Cause of Death
Death can have many different causes and in Denmark all deaths are registered by medical staff in *The Danish Register of Causes of Death*. All deaths are overall divided into five main categories, namely Natural, Accident, Violence, Suicide and Uncertain. Furthermore, an underlying cause and (up to several) contributory causes are specified using the International Classification

of Diseases (ICD) codes [41].

The ICD codes are a way to identify diseases and causes of death and it divides the causes into many categories with again many subcategories. Since 1994 the ICD-10 system has been in use, prior to that the ICD-8 system was in use in Denmark. The study period of the present study begins in 2005, and therefore only the ICD-10 system is described here.

All causes and diseases related to the circulatory system are registered with an 'I' and then a number between 00 and 99. Certain ranges of the numbers are then dedicated to some subcategory of diseases related to the circulatory system. This involves the following:

I05-I09:	Chronic rheumatic heart disease
I10-I15:	Hypertensive diseases
I20-I25:	Ischaemic heart diseases (IHD)
I21:	Acute myocardial infarction (AMI)
I26-I28:	Pulmonary heart diseases
I50:	Heart failure
I60-I69:	Cerebrovascular diseases
I60,I61,I63,I64:	Stroke
I70-I79:	Diseases of arteries, arterioles and capillaries

These will also be inspected further in the Analysis and Results chapter.

Attributes of the Cause of Death data set:

PNR: Encrypted CPR-number
$\mathbf{D_{date}}$: The estimated date of death.
Type: The overall type of death. Natural, accident, suicide, violence or uncertain.
Underlying cause: The main cause of death. Described by an ICD code.

The register on income
The registers on personal income and transfer payments are complex and contains information from a wide range of sources and more than 160 different variables [42]. All this information has by Statistics Denmark been combined with the family information in order to calculate a family equivalent income based on the following formula [43]:

$$Family_{income} = \frac{\text{income}_{\text{disposable}}}{0.5 + 0.5 \cdot \text{N}_{\text{Persons over 14}} + 0.3 \cdot \text{N}_{\text{Persons under 15}}} \tag{3.1}$$

In equation 3.1 the disposable income is calculated based on the income of all family members and adjusted for taxes, rents and interest expenses among many other things.

Attributes of the family income data set:

$\mathbf{Family_{id}}$: Identification of family.
$\mathbf{Family_{income}}$: The family income calculated by equation 3.1.
Year: The year of the income.

The family income is calculated by the end of the year and therefore it is match with the entries in CRS for the following year.

Chapter 4

Methods

In this chapter the approach and methods used in the project will be described. The first part is a description of the different methods used on the magnesium data in order to make it usable for further analysis. The second part will describe the overall design of the study. The third part will explain the statistical methods used in the final analysis of the link between magnesium and mortality.

4.1 Methods for the magnesium data

The magnesium data set is sparse in the sense that not all waterworks have samples measuring the magnesium concentration for every year. Therefore, an estimation of the concentration in the missing years was of absolute necessity before the data could be used. Three different methods were considered and they will be described here. Their performance will be evaluated in the next chapter.

4.1.1 K Nearest Neighbours

The K Nearest Neighbours (KNN) algorithm finds the neighbours of a data point and uses them to estimate a value for that point. Thus, in this case, it can be used to estimate missing data for the years in which no concentrations were measured. The algorithm needs the input on how distance is measured, how it is weighted and how many neighbours to take into account. In this project the distance is simply defined as the time between measurements and only measurements taken at one waterworks is used to estimate the missing years for that specific waterworks. The number of neighbours, K, and the best suited weighting scheme are estimated using an 8-fold cross validation described in the analysis section. The different distance weighting metrics evaluated in the analysis are the following:

Inverse distance weighting:
For the inverse distance weighting (IDW) the weight of each observation is related to the distance as follows:

$$w(d_i) = \frac{1}{d_i}$$

where d_i is the distance to point i and $w(d_i)$ is the weight used to calculate the estimate. The distance is measured in years.

One issue with this weighting scheme is that as the distance goes towards zero the weight goes towards infinity. However, since the distance is measured in years they are discrete and will be either 0, 1, 2... and so on. If the distance is 0 the weight is forced to be 2 (thus not using

the formula, as division by zero is impossible). This is a choice based on the idea that giving a distance of zero twice the weight of a distance of one was appropriate. A distance of zero occurs when the concentration in a year, in which a measurement was made, is estimated.

Inverse distance weighting squared:
A similar weighting scheme to the IDW, but the weight is squared. This gives a steeper downward curve to describe the weight. The curves can be seen in Figure 4.1. The formula is:

$$w(d_i) = \left(\frac{1}{d_i}\right)^2$$

where all symbols mean the same as above.

Tricube Kernel:
The Tricube kernel is used as a weighting scheme with the following formula:

$$w(u_i) = (1 - u_i^3)^3$$

where $u = \frac{d_i}{d_{max}}$ and $d_{max} = 37$ years, so that $0 \leqslant u_i \leqslant 1$. $w/u_i)$ is the weight used to calculated the estimate.

This method of weighting lets the importance of measurements be only slowly reduced as the distance increases.

Triweight Kernel:
The Triweight kernel is similar to the tricube but lets the importance of distances decrease a little faster. The formula is as follows:

$$w(u_i) = (1 - u_i^2)^3$$

where all symbols mean the same as above.

Triangle Kernel:
The Triangle kernel is very simple and impose a linear relationship between the distance and its weight:

$$w(u_i) = 1 - u_i$$

where again, all symbols mean the same as above.

To illustrate the differences between the weighting schemes, they are all shown in Figure 4.1. Here the relationship between the weight of a data point is shown as a function of its distance to the point being estimated. All distances are measured in years.

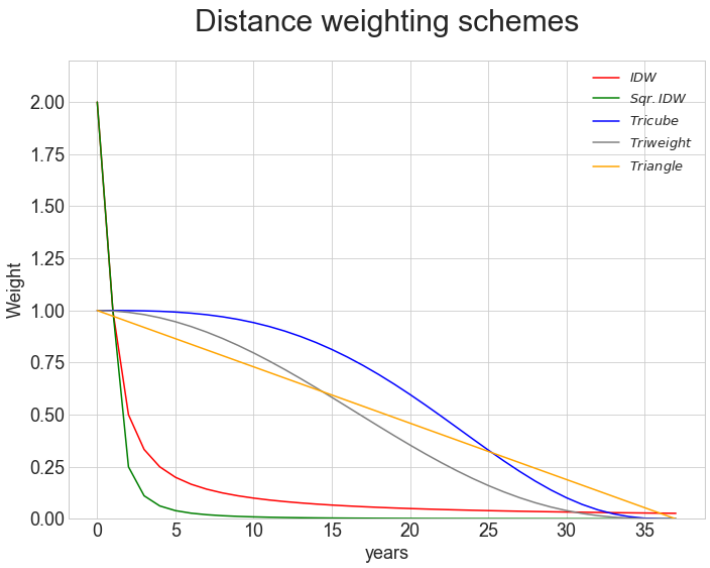

Figure 4.1: An illustration of how distances are converted to weights using the five different methods.

To illustrate how the KNN works on the data set, a plot of actual measurements and estimates is shown in Figure 4.2. Here the inverse distance weighting and 4 neighbours are used.

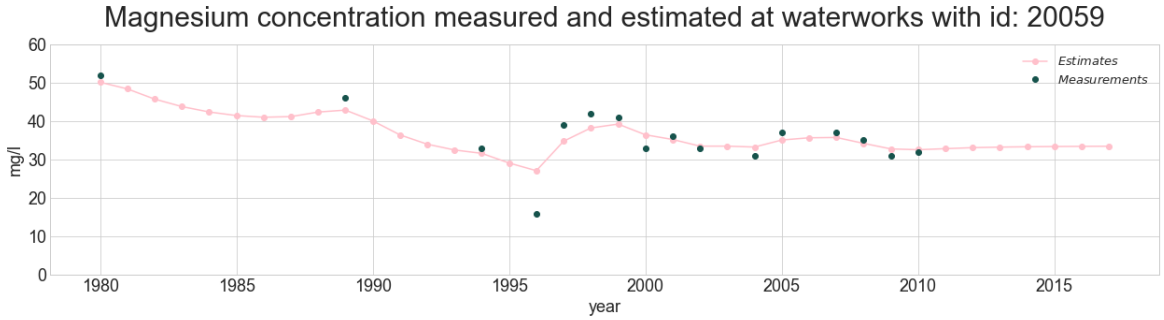

Figure 4.2: Measurements along with estimates for plant with id 20059. The estimates are based on a KNN model with K=4 and IDW.

4.1.2 Geographical interpolation

Instead of using earlier and later measurements to estimate a missing value, it is possible to use the geographically surrounding area. In this way the measurements from the surrounding areas taken in a given year are used to estimate the concentration in an area without a measurement that year. As for the KNN, it is here also necessary to determine the amount of neighbours, the distance metric and the weighting scheme. Since the distances in this case are geographical, the euclidean distance from centre to centre is calculated (in meters). For various reasons none of the above mentioned weighting schemes were fit for geographical interpolation and instead a Gaussian kernel was used with a kernel width (kw) equal to the mean of all distances:

$$w(d_i) = e^{-\frac{\sqrt{d_i}}{\sqrt{kw}}}$$

where kw is the kernel width, d_i is the distance in meters and $w(d_i)$ is the weight used in the estimation.

In order to illustrate how this method would estimate concentrations, Figure 4.3 shows estimates for the same waterworks as in Figure 4.2. Here it is evident that the surrounding areas have gen-

erally lower concentrations and thus almost all estimates are lower than the actual measurements.

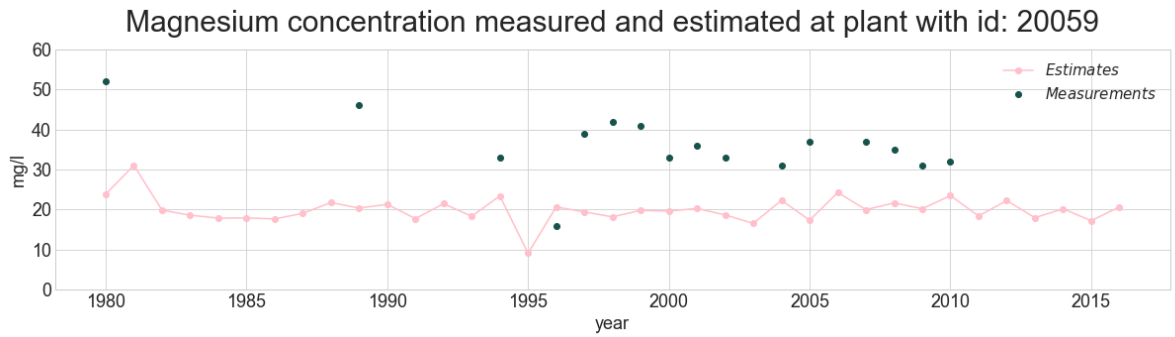

Figure 4.3: Measurements along with estimates for waterworks with id 20059. The estimates are based on a geographical interpolation with 20 neighbours and Gaussian kernel weighting.

4.1.3 Linear interpolation

In linear interpolation, the so called linear interpolant is a straight line created between two points. The linear interpolant is then used as the estimate of missing values. If the first or last data point is not in 1980 or 2017 respectively, then linear extrapolation is used. This simple technique is illustrated in Figure 4.4 again for the same waterworks.

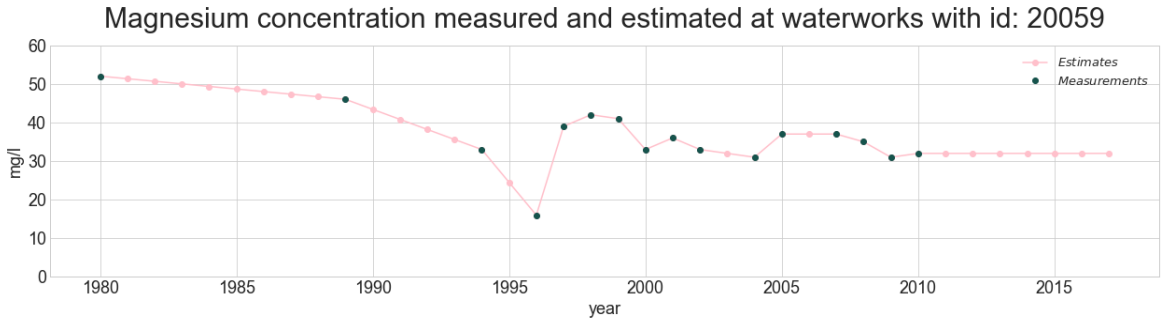

Figure 4.4: Measurements along with estimates for plant with id 20059. The estimates are based on linear interpolation and forward fill.

The evaluation of the methods is based on the negative mean absolute error, calculated as:

$$MAE_{neg} = -\frac{\sum_{i=1}^{i=n} |x_i - est_i|}{n} \tag{4.1}$$

where x_i is the actual measurement and est_i is the estimate. n is the number of actual measurements being estimated.

4.2 Study design

The study of the association between magnesium in drinking water and mortality is designed as a cohort study, also called a follow-up study. A cohort study means that a group of people are followed over a period of time. In this period it is observed whether the event of interest occurs, whether they for some reason leave the study, how their exposure changes and how their characteristics (potential confounders) change. Leaving the study is referred to as censoring. Censoring happens if for example a person dies of another cause than the one of interest or if

the person moves out of the country and can no longer be followed. If a person survives all the way through the study period the person is said to be right censored. This study is retrospective which means that the study population is followed historically and thus the study period ended before the beginning of the study. However, data are collected prospectively, as events happen. This is possible because of the well documented registers that contain information about each individual every year including information on death. The registers also make it possible to include almost the whole population in the study population. In the present study only individuals above 30 are included in the study population. Individuals younger than 30 years are not included since almost zero individuals suffers from a fatal event this young and in particular from a cardiovascular disease. Furthermore, individuals are excluded due to missing information about them. More detailed information on exclusion of individuals can be seen in the next chapter in the data preprocessing section. In this study, a so called open or dynamic cohort is used, which means that persons can enter the study after the study period has begun. This could be because they turn 30 or because they move from another country to Denmark.

The study period ranges 10 years from 2005 to 2014. This was simply the amount of data available at the time of this project. The possibility of calculating exposure in 2004 also exists, but no information concerning mortality was available for this year.

To illustrate the design, Figure 4.5 was created where examples of how different situations are handled are shown. The green bars illustrate the time in which the given person is in the study and the red dot illustrates a fatal event of interest. All the hatched areas represent time in which the individuals are not yet part of the study population, but information about them exists and is used to calculate their exposure. Person 1 illustrates a person entering the study by the beginning of the study period and surviving all the way through. This means he/she must have been over 30 in 2005. Person 2 is a person who enters the study in 2005 but dies from the event of interest during 2009. Person 3 is someone who enters the study in 2007 and survives. This person might be entering in 2007 because it was at this time he/she turned 30 and thus was allowed in the study population. The year prior to his inclusion is hatched since information from this time is used in the calculation of his exposure. Person n is a person who is only part of the study for a few years. He/she might have moved to Denmark in 2007 and left again in 2012. This means that this person is censored in 2013 and thus leaves the study without the fatal event of interest. He/she might also have died from some other cause than the one studied.

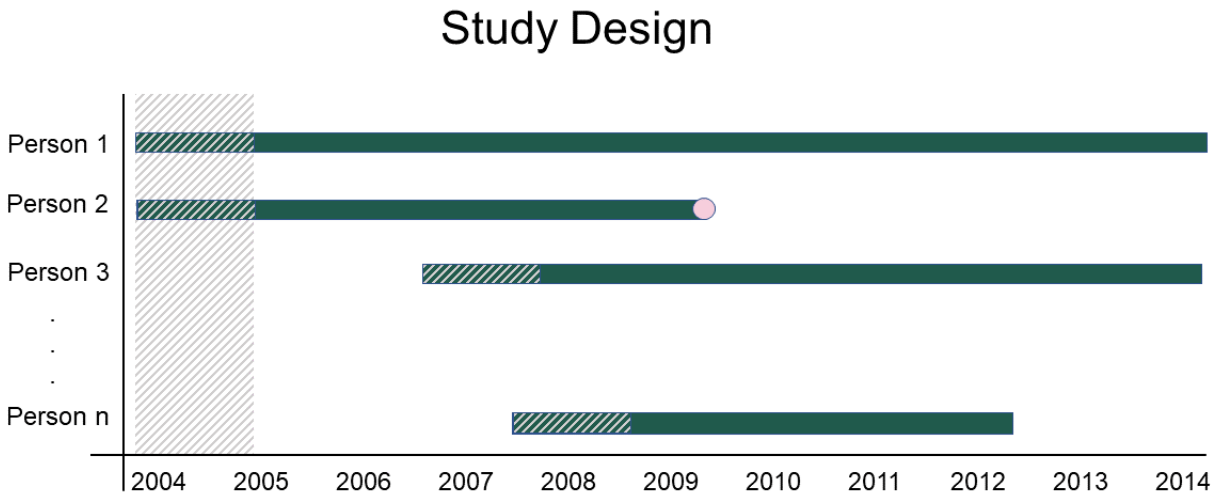

Figure 4.5: Illustration of the study design. Green bars representing time in study and red dot representing fatal event. The hatched areas represent time for calculating exposure and thus the individuals are not in the study population during this period.

For everyone in the study population the concentration of magnesium in the drinking water of the area of their residence is followed and an exposure is calculated for every year. The magnesium exposure is calculated as the average of the past two years. Thus, if a person dies mid-year 2008 then the concentrations from 2006 ($\frac{1}{2}$), 2007 (1) and 2008 ($\frac{1}{2}$) are used to calculate a weighted average (weights in parenthesis).

The event of interest is for the main analysis death from cardiovascular diseases (CD), but some sub-categories of CD will also be studied.

Furthermore, several potential confounders are followed for each individual through the study period. A confounder is a central issue for all epidemiological studies and could simply be defined as *The confusion of effects* [44]. This means that the effect of exposure is mixed with the other effects from the counfounders, thus leading to a bias if not all confounders are taken into account. The confounders chosen to be included are based on the three principles by Rothman [45]:

- A confounding factor must be an extraneous risk factor for the disease.
- A confounding factor must be associated with the exposure under the study in the source population.
- A confounding factor must not be affected by the exposure or the disease. It cannot be an intermediate step in the path between the exposure and the disease.

Moreover, the confounders considered for the study are inspired by confounders taken into consideration by similar epidemiological studies around the world. These confounders are shown in Table 2.2 in Chapter 2. All studies take age and gender into consideration and many of them also include some form of socioeconomic status. Furthermore, living alone has been shown to affect the risk of cardiovascular death [46] and is therefore also included as a confounder. As stated above, a confounder must be linked to both exposure and outcome. It is plausible that all these potential confounders are linked to both. The magnesium exposure from drinking water is dependent on the geographical location of residence and since geographical variations exist in these factors they can be linked to magnesium exposure as well as risk of CD.

This relationship between exposure, outcome and confounders is illustrated in Figure 4.6. The illustration is inspired by the causal diagrams or directed acyclic graphs also described by Rothman [45]. However, a full causal analysis was outside the scope of this project since it requires substantial expert knowledge.

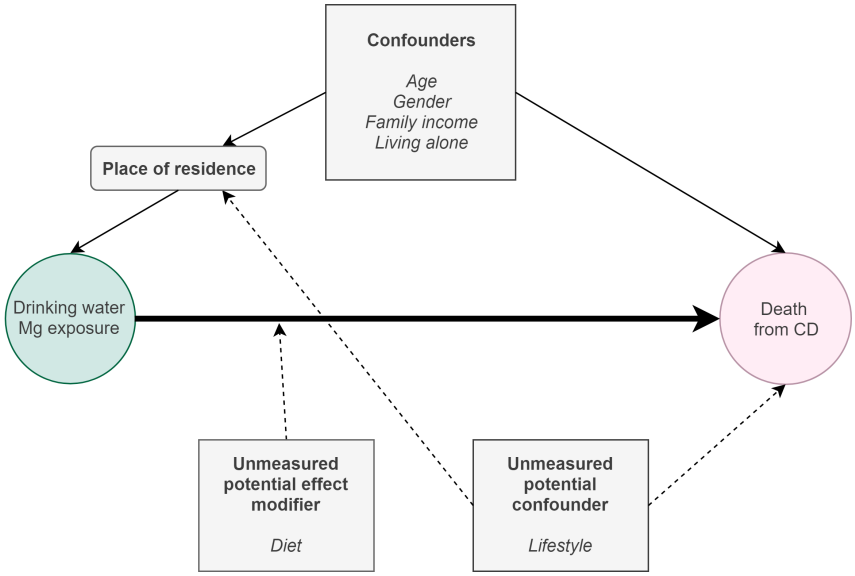

Figure 4.6: Model to illustrate that the confounders are related to both exposure and outcome. Unmeasured potential confounder and effect modifier are added with hatched line to illustrate they are just proposals.

In the figure, the bold arrow between drinking water magnesium exposure and cardiovascular death represents the link investigated in this study. The confounders, age, gender, family income and living alone, are related to cardiovascular death since they all have an effect on the risk of dying from CD. However, for them to be actual confounders they need to have an effect on the drinking water exposure. They have this indirectly through the place of residence. An unmeasured confounder is lifestyle which includes smoking and exercise habits. This confounder is linked in a similar way to exposure and outcome as the other confounders. Unfortunately, this information is not available in the study. In the figure, diet is written as a potential effect modifier because you get magnesium from your diet as well as from your drinking water. If you get plenty of magnesium through your diet then being exposed to high magnesium levels in your drinking water is not likely to have the same effect as if you have a magnesium deprived diet. However, diet is neither available in the study.

In general the motivation behind assessing effect modification is to understand whether the exposure has a different effect in groups with different characteristics, e.g. men and women. If the effect is the same across all groups then it is called homogeneous and otherwise heterogeneous. Effect modification is somewhat similar to what is denoted an interaction. Interactions are used when the aim is to investigate whether there is a joint effect of two or more characteristics on the outcome. Interactions can be used to model effects that are not constant across the categories of some other effect. For example the effect of being male versus female on the risk of dying from CD might change with the age category. This can be handled by introducing an interaction term between gender and age.

One adjustment not mentioned in Figure 4.6 is the adjustment for calendar year. This is relevant since the risk of dying from CD has been reduced during the study period and if changes in magnesium exposure also varies across the years it will be a necessary component in the model. Furthermore, this parameter will open up the possibility of estimating a trend in the relative risk of being exposed to low versus high magnesium levels. For example, magnesium in

drinking water could prove to have an increasing or decreasing importance over the study period.

The specifications of the study are summarised below:

Specifications
Type: Retrospective open cohort
Study population: Danish population aged 30 or more
Study period: 2005-2014
Exposure: 2-year magnesium average
Event: Cardiovascular death
Confounders: Age, gender, living alone, income level + adjustment for calendar year.

Certain subcategories of CD are also investigated as the event of interest. This includes acute myocardial infarction, stroke and ischemic heart disease.

Several sensitivity analysis are carried out which includes examining interactions and effect modifications. They are examined through changes in the statistical model.

However, another way of handling them is also attempted. Here an effect modifier is handled by doing an analysis only for the sub-population that is assumed to behave differently. In one sensitivity analysis, the elderly population is assumed to be more affected by their magnesium exposure and thus an analysis only including them is carried out.

4.3 Statistical methods

To analyse the association between exposure of magnesium from drinking water and mortality, Poisson regression is used. In the following, the method will be introduced along with the model used for analysis. The model is based on the main study design described above. Before introducing Poisson regression, incidence rates in general and how they are used for descriptive analysis will be described.

4.3.1 Incidence rates

In order to estimate the risk of a given event, it is common to use the incidence rate, IR. The incidence rate describes the amount of events in a specific time period in a specific group of people. This is typically the number of events per 100,000 person-years. In this study an event is death from cardiovascular disease. This can for a descriptive analysis simply be calculated as:

$$IR = \frac{d}{RT} \tag{4.2}$$

where d is the number of events and RT is the total sum of risk time in the group, often measured in person-years.

The incidence rate itself can be of great interest as it describes the risk, but the ratio between different groups are often even more interesting. This describes the relative risk between two groups - usually groups with different exposure levels. It is calculated as:

$$IRR = \frac{IR_2}{IR_1} \tag{4.3}$$

where IR_2 is the incidence rate in exposure group 2 and IR_1 is the incidence rate in the unexposed group (or the reference group, e.g. lowest exposure group).

Due to the nature of ratios, only the logarithm of IRR has an approximately normal distribution where the variance can be estimated as:

$$s^2_{\ln(IRR)} = \frac{1}{d_1} + \frac{1}{d_2} \tag{4.4}$$

where d_2 is the number of deaths (events) in exposure group 2 and d_1 is the number of deaths in the reference group.

Then a confidence interval of for example 95% can be estimated as:

$$\exp(\ln(IRR) \pm 1.96 s_{\ln(IRR)}) \tag{4.5}$$

where $s_{\ln(IRR)}$ is the standard deviation.

4.3.2 Introduction of Poisson regression of incidence rates

The incidence rates described above can only be used for a simple descriptive analysis. If the adjustment for confounders is needed, then a more advanced method is necessary. This could be the Poisson regression method. Here the incidence rates are estimated by the expected number of deaths divided by the number of person-years (risk time) in the study as follows:

$$IR_i = \frac{E[d_i|x_i]}{RT_i} \Leftrightarrow \tag{4.6}$$

$$\ln E[d_i|x_i] = \ln RT_i + \ln IR_i \Rightarrow \tag{4.7}$$

$$\ln E[d_i|x_i] = \ln RT_i + \alpha + exposure_i \tag{4.8}$$

where RT_i is the number of person-years in exposure group i and is referred to as the offset and x_i is an indicator of the exposure group. IR_i is the incidence rate of exposure group i and it is approximated by the two parameters α and $exposure_i$. α is the intercept related to the incidence rate of the reference group and $exposure_i$ are parameters related to each exposure group. For example if 5 exposure groups exists then $i = 1, 2, 3, 4, 5$ and four exposure parameters are estimated along with the intercept. This model is the simplest possible and is not yet adjusted for confounders. It would in fact yield the same results as the descriptive analysis described above [47].

4.3.3 Multiple Poisson regression

Multiple (multivariable) Poisson regression is an extension of the simple (univariable) Poisson regression introduced above. Multiple refers to the fact that the expected number of events are based on multiple parameters, thus adjusting the incidence rates for differences in many parameters. The estimated parameters are related to the exposure as in the previous section, but also to all the confounders included in the study.

All variables in the present study are categorised which makes it possible to do the analysis based on aggregated data. This means that the data set is grouped so that all persons with the same characteristics and belonging to the same exposure group are turned into one single line in the data set. This line also has information about the amount of person-years related to it

and the amount of deaths. The aggregation of data makes the computation time much faster [48].

Each such line represents a certain stratum within a certain exposure group. A stratum is defined by a specific combination of characteristics. For example one stratum could be men aged 50-55, living alone with a high income in year 2010. This subgroup of the population will exist within each exposure group. In the aggregated data set it is determined how many people from this stratum died within each exposure group. This is done for all possible combinations of confounders and calendar years. Poisson regression then estimates the number of deaths related to each line by estimating a range of parameters and using the risk time associated with each line.

For the statistical analysis one main model is used. This model is based on the study design described earlier and looks as follows:

$$\begin{aligned}\ln \mathrm{E}[d_{ijklmn}|\mathbf{X}_{ijklmn}] = \ln RT_{ijklmn} + \alpha + exposure_i + age_j + gender_k + \\ cohabitation_l + income_m + calender\_year_n\end{aligned} \tag{4.9}$$

where d_{ijklmn} is the number of deaths, RT_{ijklmn} is the number of person-years and $\mathbf{X}_{ijklmn}$ is the vector of variables within the jth age category, the kth gender, the lth cohabitation status, the mth income group and the nth calender year as well as the ith exposure group. α is the intercept related to the incidence rate of the reference group of each variable. All the other elements in equation 4.9 are parameters related to the variables. For each variable the number of parameters estimated is the amount of categories minus 1, since one arbitrarily chosen category will be the reference.

In total, this model has 32 unknown parameters that need to be estimated. The estimation is done using the software SAS9.4 and the procedure *proc genmod*. An example of how the *proc genmod* is used can be found in appendix A. This procedure uses maximum likelihood estimation to estimate all the unknown parameters. When estimating parameters 95% confidence intervals are also estimated. Furthermore, it reports the p-value of a likelihood ratio test of a model with and without each parameter [49]. The p-value is the probability of observing the likelihood ratio given the null hypothesis is true, where the null hypothesis is that the simplest model is the best fit.

As mentioned earlier, the incidence rate ratios (IRR) are often of great interest. Below is shown an example of how the IRR between exposure group 1 and 4 is calculated:

$$IRR_{1vs.4} = \frac{IR_1}{IR_4} \tag{4.10}$$

$$= \frac{\exp\left(exposure_1 + age_j + gender_k + cohabitation_l + income_m + calender\_year_n\right)}{\exp\left(exposure_4 + age_j + gender_k + cohabitation_l + income_m + calender\_year_n\right)} \tag{4.11}$$

$$= \exp\left(exposure_1 + age_j + ... + calender\_year_n - (exposure_4 + age_j + ... + calendar\_year_n)\right) \tag{4.12}$$

$$= \exp\left(exposure_1 - exposure_4\right) \tag{4.13}$$

where $exposure_1$ and $exposure_4$ are the maximum likelihood parameter estimates related to exposure groups 1 and 4 respectively.

Sensitivity analysis

In the sensitivity analysis interactions and effect modifiers will be added to the analysis. This means that an extra part will be added to equation 4.9. One interaction examined is an interaction between age category and gender, this would lead to the following part being added:

$$age\_gender_{jk} \tag{4.14}$$

which is a parameter that would be estimated for each possible combination of age categories and genders. However, one category of each variable would be chosen as the reference. If 13 age categories exists and two genders, then 12 extra parameters would be estimated. Nothing would be changed in equation 4.9, the individual effect of age and gender would still be there.

For the models including effect modifiers, a part added to equation 4.9 could look as follows:

$$exposure\_age_{ij} \tag{4.15}$$

which is also an estimated parameter associated with both exposure group and age category. Assuming five exposure groups and 13 age categories, this effect modifier would lead to an estimation of 48 extra parameters.

Chapter 5

Analysis and results

5.1 Data preprocessing

In order to work with the data and in the end make the statistical analysis of the association between magnesium in drinking water and mortality, much preprocessing had to be done to the various data sets (see Chapter 3 for details on data sets). To illustrate the process, Figure 5.1 has been created. The six squares at the top of the figure represent the raw data sets. Squares further down the tree represent processed data sets that are central to the project. The first one is the one containing estimates of magnesium concentrations. This data set demanded many considerations and analysis on its own and is therefore represented in the figure. The last data set at the bottom of the tree is the final data set that was used in the statistical analysis. The interval of years written below the title of each data set indicates the time period in which data is available for that specific data set. All steps marked as circles on the figure are actions done to the data sets and they will shortly be elaborated upon. The n on the figure refers to the number of observations in a data set, and thus it can be seen how the observations are reduced (or increased) during the preprocessing. The red box is a disquisition of observations lost due to matching of the register data to the magnesium estimates through the addresses.

First, the right side of the figure will be explained, thus beginning with the magnesium data sets. The data set containing concentrations of magnesium measured from water samples had 62,941 observations. Each sample was linked to a waterworks and each waterworks was then attempted to be linked to a water supply area. This was possible for almost all samples except for 128. These were excluded. The samples were then grouped by the year in which they were taken. This reduced the data set to 56,131 observations. For the years in which more than one sample were taken, the average of the concentrations was kept. In total, 35% of the reduction was due to samples taken on the exact same day as another sample. This data set was then used in the step called KNN, where estimates of concentrations were made. A description of this step will be given later in the chapter. The estimates were made for all waterworks with at least one sample and for all years from 1980-2017. It resulted in a total of 148,504 estimates.

The data set called *waterworks abstraction* dated far back in time and therefore not all registrations were relevant for this study. The amount of relevant registrations were further reduced by waterworks without magnesium estimates or with negative registrations. These were all excluded in the step called data cleaning. The data set was left with 103,734 relevant registrations of abstraction which should be compared to the 148,504 magnesium estimates. This means that not all waterworks had a registered abstraction for all years. It could be due to the fact that some of the waterworks have not been active during the whole period. Some waterworks were connected to more than one WSA and for those it was impossible to know how much water they delivered to each area. Therefore the abstraction was simply divided by the amount of WSAs

that the waterworks was supplying, thus assuming an equal amount was delivered to all areas. This is possibly not always the case but with no further information available, it was the most valid assumption. For more information about the connection between waterworks and WSAs see Table 5.2 and Table 5.1. This correction of abstraction levels was also contained in the data cleaning step.

The magnesium estimates for each waterworks and the abstraction data were then merged and grouped by WSA. The estimates of the waterworks were thus transformed into estimates for each area. For WSAs connected to more than one waterworks this was done by weighting the estimates for a given year by the abstractions registered for that year. For example, if an area had 90% of the water delivered from one waterworks and 10% from another, then the estimated concentrations for the two waterworks would be used to calculate a weighted estimate for the WSA, weighting the estimates 90% and 10% respectively. In case no registrations were made for one of the waterworks, it was assumed that it did not deliver any water that year. However, for years where none of the waterworks connected to a WSA had any registrations of abstraction a simple average of the relevant magnesium estimates was used. This was done with the assumption that the area must have had water delivered and that some registration error had most likely occurred. This was the case for all estimates of 2017 since the 2017 abstraction data was not available. It was also the case for many estimates before 1990. The registration process seems to have been sparse at that time. The merging of the magnesium estimates and the abstraction registrations resulted in a data set with estimates for all water supply areas based on the waterworks connected to it. In total 97,556 estimates were calculated and ready to be used in the further analysis. The amount of estimates is reduced because the original waterworks estimates were grouped by WSA.

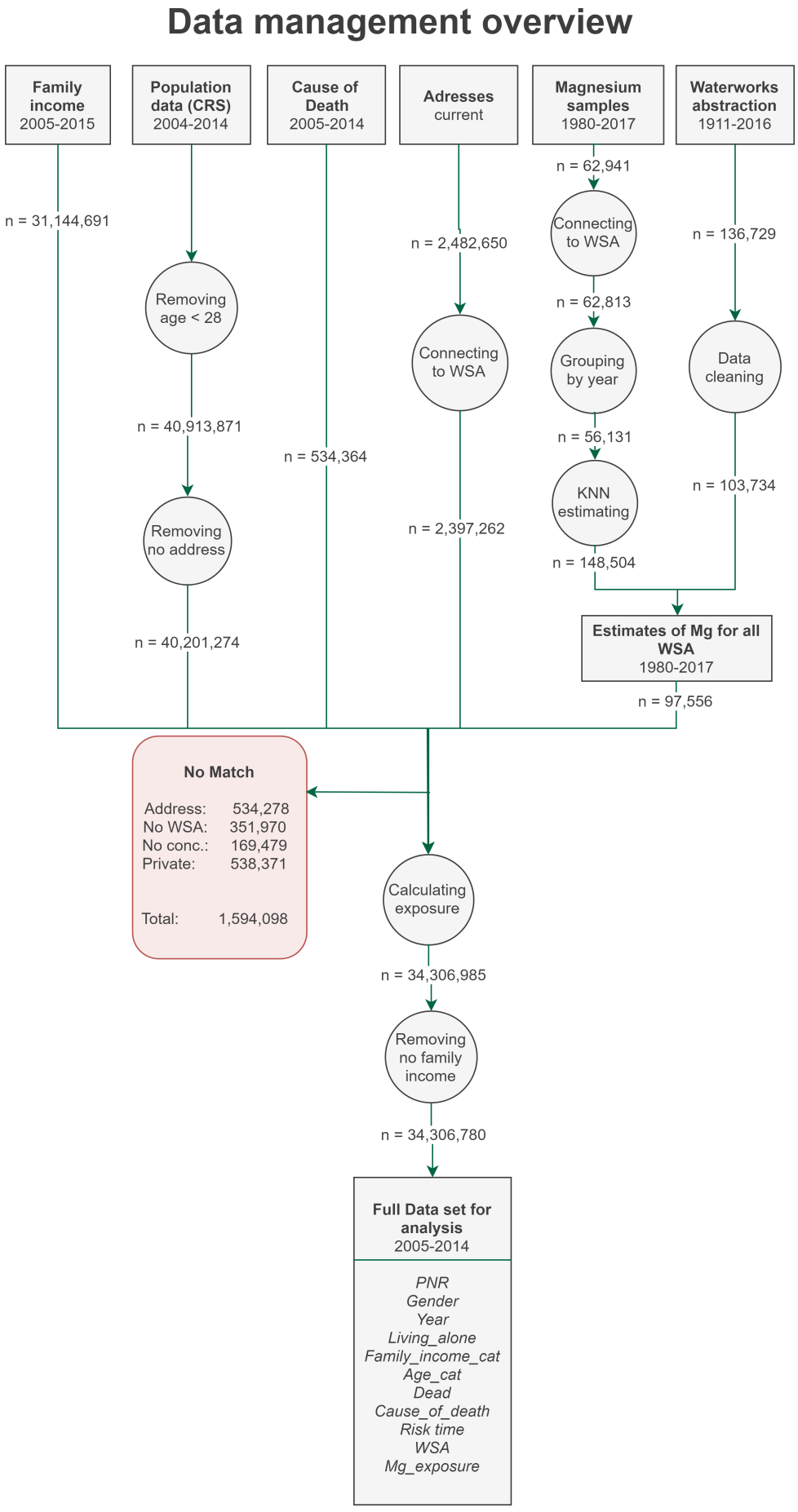

Figure 5.1: Overview over the data management process. Squares symbolise data sets and circles symbolise actions.

As can be seen in Figure 5.1, the data set with estimates of magnesium concentrations had to be linked to four other data sets. These data sets include the family income, population data (CRS) and cause of death - all originating from Danish registers. These were accompanied by a data set containing the geographical location of all current addresses. It was then determined in which WSA the coordinates of each address were located. This was done in the step called *Connected to WSA* and in total less than a 100,000 addresses could not be matched to any WSA. This was simply due to the fact that their coordinates were not inside any of the water supply areas. Kirstine Wodschow used a geographical information system (QGIS version 2.18.14) to do the geographical matching [2].

Before merging the register data with the magnesium data, all individuals aged less than 28 and individuals with no address were removed from the data set. Only very few observations did not have an address. They were excluded in such a way that all individuals who had at least one year with no address were completely removed from the data. This was done because it would not be possible to calculate their exposure correctly. The reason for keeping persons aged 28 and 29 is to be able to calculate the two-year average exposure.

When merging all these data set into one final data set many observations from the register data did not have a match in the address data set and therefore no connection to any WSA. This was the case for more than half a million of the observations and in particular this was an issue for observations before 2007. In 2007 many of the municipality codes changed due to the structural reform. As mentioned earlier, some addresses were not linked to any WSA and this resulted in around 350,000 observations not being linked to a WSA. For some observations the case was, that they were linked to a WSA with no concentration. This would be the case for people living in the few areas with no measurement and therefore no estimates. More than half a million were excluded due to the fact that they were supplied by their own well and even though the concentration might be similar to that of the surrounding area it was deemed too uncertain.

For all individuals left in the data set the exposure was calculated. This was done as an average of the past two years. A slightly simpler average was used for people dying during 2005 where only the concentrations from 2004 and 2005 were used, making it less than a two-year average. This was done so that observations from 2005 could be used and thus keeping the study period ten years long. For people having left Denmark and moved back they were only included two years after they had reentered so that an appropriate exposure could be calculated. After the calculations, observations of people aged 28 or 29, observations from 2004 and observations where calculations could not be completed (due to reentering the country) were removed from the data set. Furthermore observations where no family income was present were excluded.

This left a data set of around 34.3 million observations corresponding to a study population of 4,143,662 unique individuals. This yields an average time in the study of 8.3 years.

The only extra thing that had to be done to make the data set ready for analysis was categorising the data. The exposure was divided into five categories of equal size yielding the following exposure groups:

Group 1: Exposed to 6.65 mg/l or less.
Group 2: Exposed to more than 6.65 and up to 10.3 mg/l.
Group 3: Exposed to more than 10.3 and up to 14.6 mg/l.
Group 4: Exposed to more than 14.6 and up to 21.9 mg/l.
Group 5: Exposed to more than 21.9 mg/l with the maximum being 53.6 mg/l.

The family income was likewise divided into five categories of equal size. This was done by taking the inflation of income and the difference between retired individuals and none retired ones into account.

The age was divided into 13 categories with five-year intervals and the last category being 90+.

The attribute specifying whether the individual is living alone or not was based on the attribute $Family_{type}$ from the CRS. If this attribute was 5 they were said to be living alone and otherwise they were marked as not living alone.

5.2 Descriptive analysis of the Magnesium data

In order to work with the magnesium data set it is first important to understand what it contains and where issues might arise. In this section a descriptive analysis of the data set will be carried out. This will lead to the proposed methods for estimating missing magnesium concentrations.

As mentioned in Chapter 3, the magnesium data set contains measurements of magnesium levels and each measurement is related to a waterworks. In total, the data set contains 3,684 different waterworks. Each is related to a water supply area (WSA) and in total the data set has 2,537 unique WSAs. It appears clearly that often many waterworks must be related to the same area. However, it is also the case that one waterworks is sometimes the supplier of more than one area. To illustrate how common these two types of double supply are, the number of occurrences are listed in tables 5.1 and 5.2.

Waterworks connected	Number of WSAs
1	1901
2	331
3	158
4	64
5	33
6	16
7	13
8	10
9	3
10	1
11-22	9

Table 5.1: Table of how many WSAs have a certain number of waterworks connected.

Connected to (WSAs)	Number of waterworks
1	3619
2	25
3	14
4	9
5	4
6	4
7-19	10

Table 5.2: Table showing how many waterworks are connected to a certain number of WSAs

These types of double connection are important when the concentration in each WSA is calculated based on estimates from the waterworks as described in the previous section.

In Figure 5.2, is a box plot of all measured concentrations of magnesium in mg/l. It can be seen that the Danish drinking water on average contains around 9.8 mg/l, but it ranges from 0.005 to more than 50 mg/l with extreme values up to 90 mg/l. Half of all measurements are actually contained within a quite narrow range between 6 and 16 mg/l. In the higher concentrations many outliers are present and they are here defined as being more than 1.5 times the inter quartile range from the 3rd quartile.

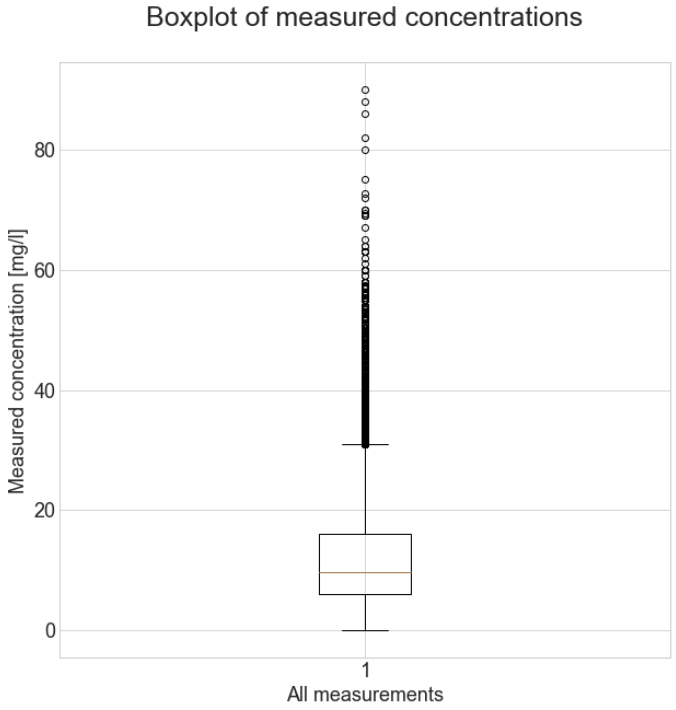

Figure 5.2: Box plot of all measured magnesium concentrations in mg/l.

In total there exists more than 60,000 measurements made between 1980 and 2017. The amount of measurements are at some waterworks only a single one whereas others have measurements from all years. All waterworks with at least one measurement are used in the further analysis. However, some waterworks have never measured the magnesium concentration and the magnesium level in the water delivered by them are thus classified as unknown. In order to examine how much water with unknown magnesium levels that was delivered to the consumers, the total abstraction of all waterworks is illustrated in Figure 5.3. In this figure the yellow bars represent abstraction from all waterworks and the green bars illustrate how much of the total abstraction we have information about. By information is meant at least one magnesium concentration measurement between 1980 and 2017. As seen in the figure quite a bit of the water is actually water with no information and thus also water that is not part of the further analysis. The black lines represent the official water abstraction numbers found at Statistics Denmark's web page [14]. They are added to confirm the correctness of the abstraction data set and as can be seen in the figure, the registrations are similar to the data. The only exceptions are in 2015 and 2016 were Statistics Denmark has slightly higher registrations, this could be due to some delay in the data available to the present study.

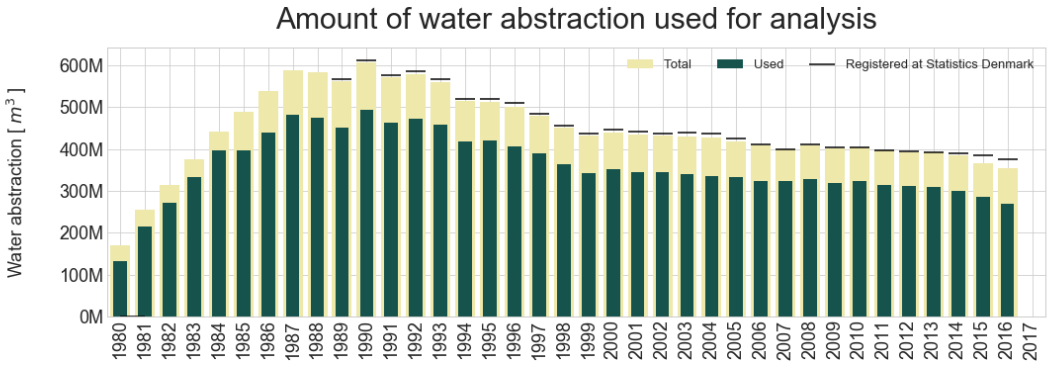

Figure 5.3: Bar plot showing the amount of water abstraction used for the analysis compared to the total amount registered each year. Black lines represent the abstraction officially registered at Statistics Denmark

Over the past 30 years the water abstraction has decreased and this downward trend is simply due to consumers using less water. The low registrations of abstraction in the 80ties are, however, due to unreliability of the registrations. At Statistics Denmark only abstractions back to 1989 are recorded [14].

5.3 Estimation of magnesium levels

In order to use the information about magnesium levels in different areas it is necessary to have an estimate of the level each year. In some areas many years might pass between measurements and estimating what the concentration was in these years can be done in many ways. No matter how the estimation is done it should be noted that long periods with no measurements does invoke some uncertainty regarding the estimated concentration used for the analysis. To assess the problems and how many areas were affected by them, Figure 5.4 was created.

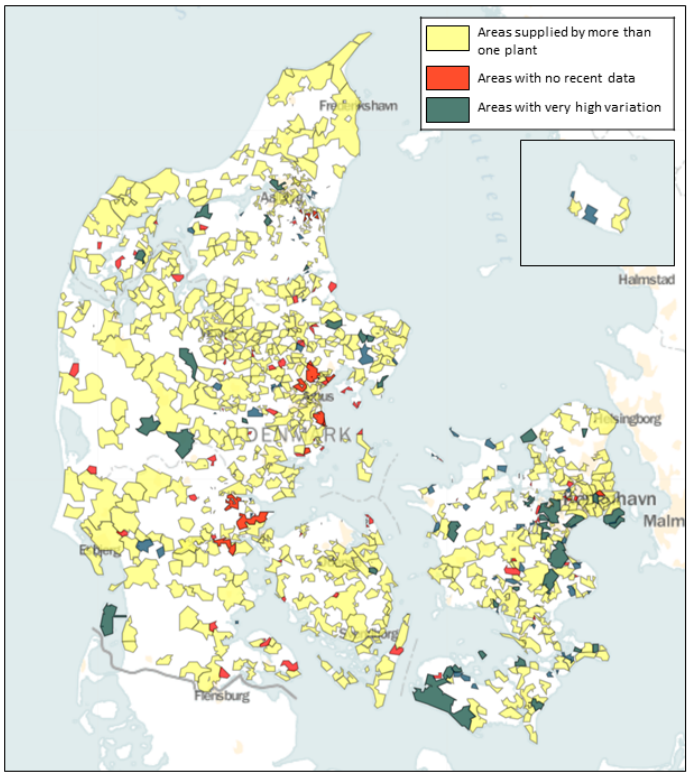

Figure 5.4: Map of problematic WSAs.

On the figure, areas that are supplied by one or more waterworks are marked in yellow, areas with no data since 2000 are marked in red and areas with very high variation are marked in green. That an area is supplied by more than one waterworks is problematic since it can be difficult to know which waterworks each household gets water from. The areas with no recent data will still be estimated but it should be noted that the estimates will have a high uncertainty. The areas with high variations defined as a standard deviation of more than 4 mg/l are also problematic since the reason for the variation is unknown. Almost all of them also have more than one waterworks connected and the variation could simply stem from that fact. However, it could also be due to a sudden change in concentration, which is difficult to handle.

All of this made it clear that concentrations had to be estimated for each waterworks and then later aggregated to WSA level. In order to figure out the best way to do the estimations, three methods were taken into account. In the following their performance will be described.

5.3.1 Linear Interpolation

First of all, the method of simple linear interpolation has the issue of using the measurement made in one year as the exact value valid for that whole year. This is of concern since a measurement made on a specific date could be unusually high or low compared to measurements made before and after. Assuming that such a measurement was the true value for the entire year seemed like too strong an assumption. Instead it was decided to use a way of estimating the true concentration based on several measurements. Therefore, the method of linear interpolation was not used for the final analysis.

5.3.2 Geographical interpolation

A second method using geographical interpolation was attempted. This method uses the neighbouring areas to determine the concentration of magnesium. Since magnesium is found in the aquifer and is related to geography, the concentrations of neighbouring areas would be expected to be similar. For testing this, leave-one-out cross validation with the euclidean distance metric, the gaussian kernel weighting and 20 neighbours was used. This method gave a negative mean absolute error of -4.719 mg/l calculated using equation 4.1. Moreover, 15,796 data points could not be estimated because they were related to areas where none of the 20 closest neighbours had any measurements. It seems as if many data points could not be estimated, but it is probably due to the fact that in the early years only very few areas had any measurements. It was also experimented with changing parameters in the geographical interpolation but none of the experiments seemed to yield results close to being as precise as the KNN method (results in next section) and therefore only this one constellation is reported.

5.3.3 The KNN method

The third method uses the K-nearest-neighbour algorithm. In order to determine how well this method performs a cross validation was carried out. The validation was made in two steps where the first step included determining the best weighting scheme and the optimal number of neighbours (K). This was done only using the waterworks with more than eight measurements so that up to 6 neighbours could be tested at all waterworks. It was an 8-fold cross validation with a test set of 10% and a training set of 90%. This means that for each waterworks 10% of the data points were removed and placed in a test set. The rest of the data was then used to estimate the data points in the test set. This was for each waterworks done 8 times, every time

placing a different 10% in the test set.

The results can be seen in Figure 5.5 where the darker green colours indicates better test scores. The test scores used for the heatmap are the mean of the negative mean absolute error.

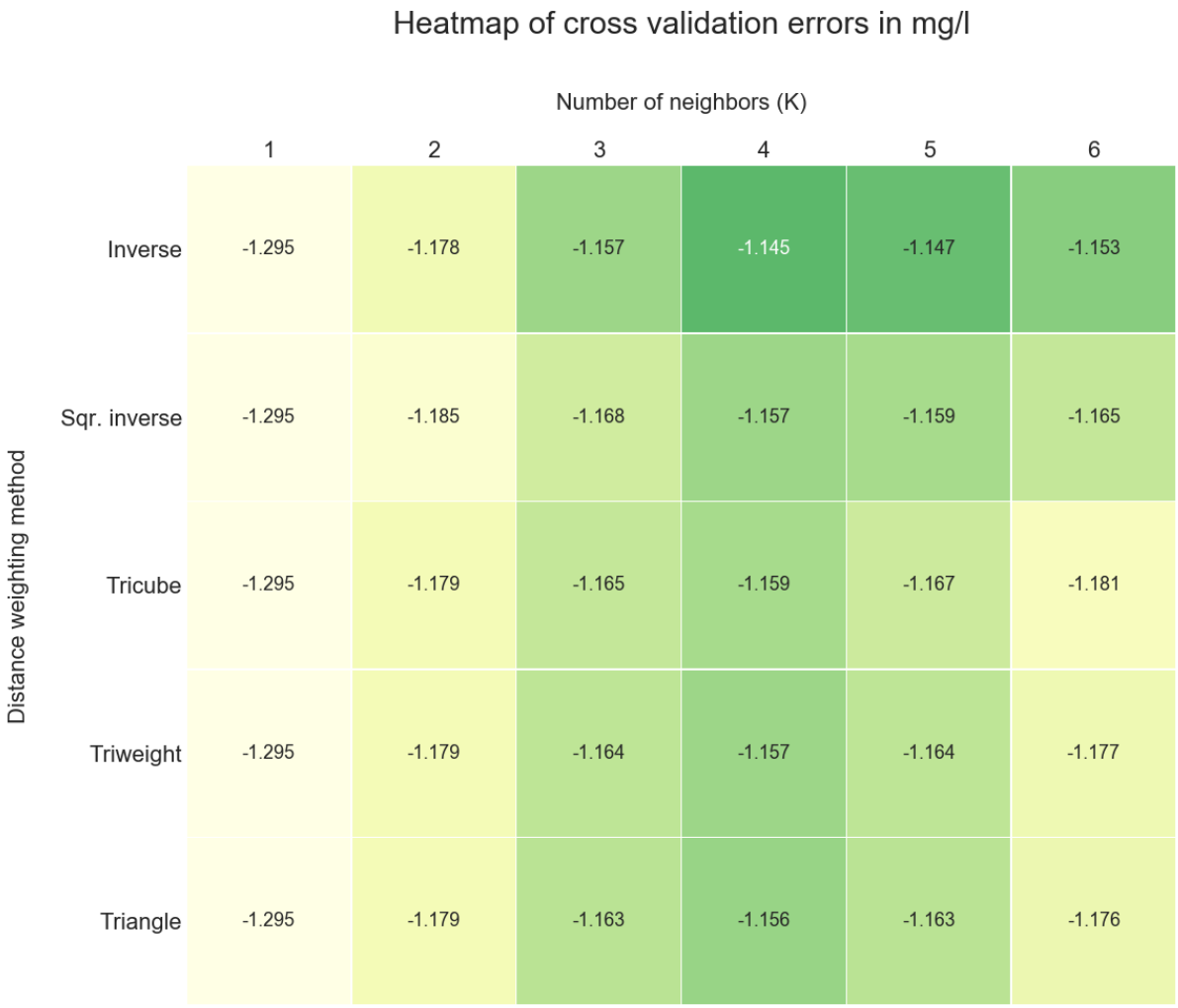

Figure 5.5: Heatmap illustrating the average negative mean absolute error for different weighting schemes used on K from 1 to 6.

From the figure it is clear that many combinations have similar errors and they might be equally good to use. However, the combination with lowest error is the inverse distance weighting and 4 neighbours. Since this combination performs slightly better than the others, it is the one used. In order to examine what happens when concentrations at all waterworks are estimated (including the ones with 8 or fewer measurements) step 2 of the cross validation was carried out. In this step a leave-one-out cross validation was made. Here each data point was estimated by the use of the others. The only waterworks excluded in this cross validation were the ones with only one measurement. For waterworks with less than five measurements only the maximum possible number of neighbours was used. Hence, if a waterworks had three measurements only two neighbours were used to estimate the third data point.

Using this approach the negative mean absolute error was -1.22 mg/l, which is slightly worse than the error from the preliminary cross validation. However, this is to be expected since also concentrations at waterworks with very few samples were estimated here.

In Figure 5.6 a histogram of the cross validation errors are shown. This shows that almost all of them are very small but a few are very large. One concentration is estimated to be more than 70 mg/l form the actual measurement.

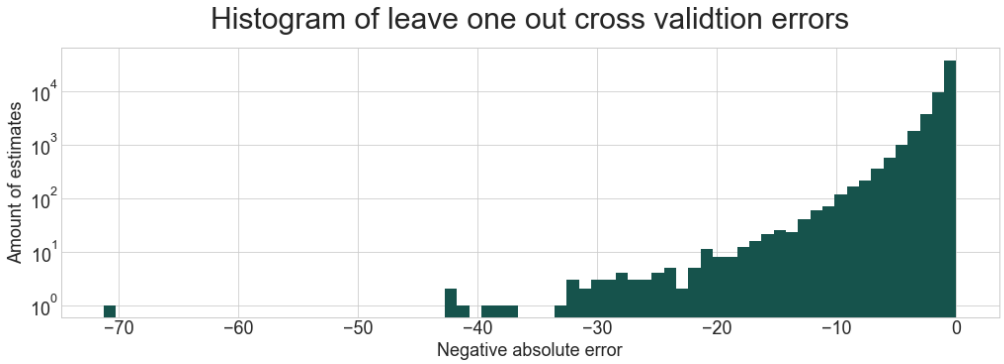

Figure 5.6: Histogram of negative absolute errors from leave-one-out cross validation using the inverse distance weighting and 4 neighbours where possible.

Note that the y-axis of the figure is on the log-scale making the large negative values visible. All large errors was double checked manually in the Jupiter database and no apparent reason for the difference could be found.

For the final data set of estimations this method was applied. It was applied on all years and thus also years with samples were estimated using the four nearest neighbours. However, the sample taken in that year was of course also used and with a distance of zero it was the closest data point and therefore also attributed the highest weight.

5.3.4 The data set of estimations

The method used for the final data set is the KNN method. The method was as described used on all waterworks with at least one measurement of magnesium. Also as described earlier all waterworks deliver water to at least one water supply area. Since it is only possible to determine in which WSA an address is located, the estimates of the waterworks needed to be transformed into estimates of concentrations of each area. This was done as described in the data preprocessing section.

This aggregation led to the final data set of estimates and in Figure 5.7 one example of how the original measurements are turned into a final weighted estimate is shown.

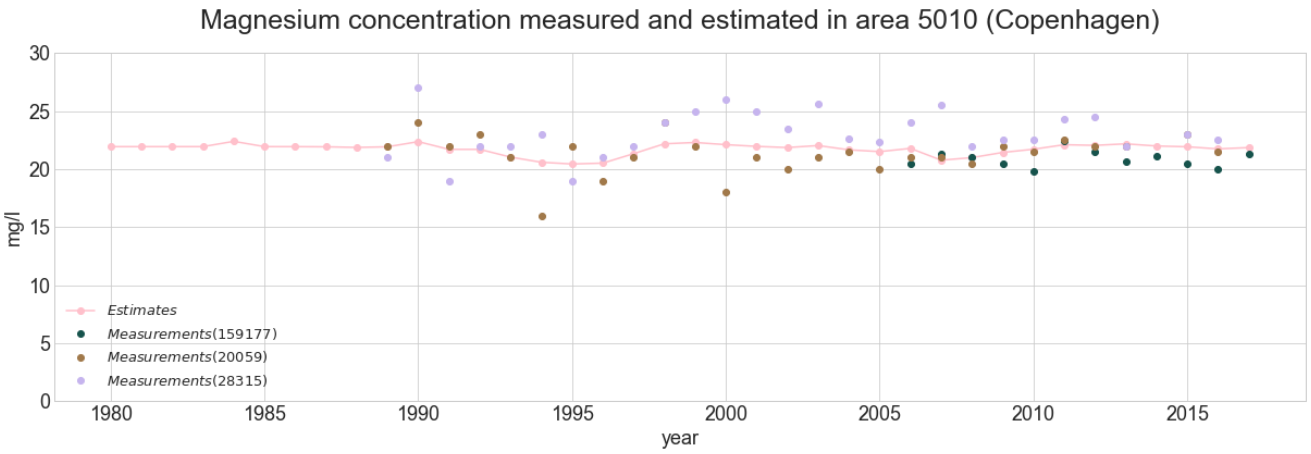

Figure 5.7: Area 5010 close to Copenhagen is connected to three waterworks. The estimations (in pink) made from all the original measurements shown as green, brown and purple dots.

In Figure 5.7 the pink line is the estimate of concentrations in this specific area close to Copenhagen. These are the ones used in the further analysis. The estimate is based on all the other dots coloured according to the waterworks at which they were taken.

In order to assess how all these estimates look geographically and time wise, several maps were created. However, their similarity is striking and only one is shown here in Figure 5.8. The rest can be found in appendix B. This map is based on estimates from 2015 and only areas marked in grey has no estimated concentration. As seen on the map it is primarily in Jutland that areas with very low concentrations exists.

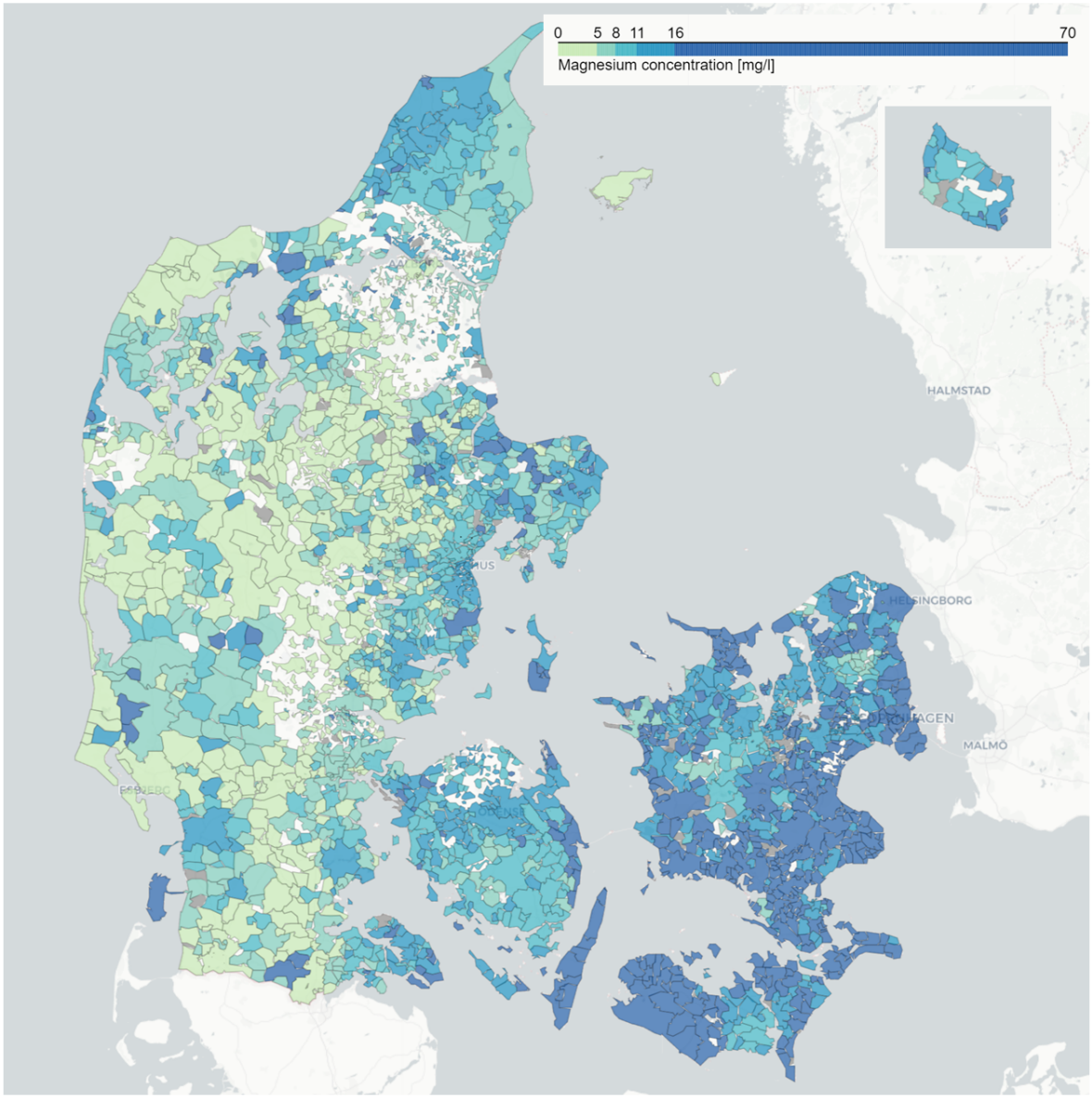

Figure 5.8: Map showing the estimated levels of magnesium in all WSAs (with data). Year 2015. Areas with no data (and hence no estimation) are marked in grey. The lighter grey is the ocean. The white is areas where no WSAs exists.

5.4 Descriptive analysis of the final data set

The final data set is created from the data set of estimations and all the register data and linked via the addresses as described in the preprocessing section. Before carrying out the statistical analysis using Poisson regression, a descriptive analysis of this final data set will be given.

In order to get a first look at the data set, all the attributes have been described in Table 5.3. This includes the amount of cardiovascular deaths within each category of each attribute as well as the amount of risk time that exists within the group. The simple unadjusted incidence rates have also been calculated. The table gives an indication of the importance of each attribute, however, one should be very careful when looking at the rates since the actual effect might not be apparent as they are all unadjusted.

Characteristics		N_{deaths}	Risk time (in million years)	Incidence rate per 100,000 person-years
Gender	Men	67,346	1.6	411.6
	Women	70,734	1.7	406.7
Age Category	30-34	190	3.7	5.2
	35-39	379	3.7	10.3
	40-44	782	3.9	20.2
	45-49	1,379	3.8	36.6
	50-54	2,402	3.5	68.9
	55-59	3,653	3.4	68.9
	60-64	6,412	3.4	190
	65-69	8,991	2.8	319
	70-74	12,712	2.1	616
	75-79	18,684	1.5	1,214
	80-84	25,960	1.1	2,360
	85-89	28,527	0.65	4,420
	90+	28,009	0.32	8,860
Family income	1	40,937	6.7	612.0
	2	35,338	6.7	525.5
	3	29,372	6.7	435.3
	4	20,092	6.7	296.3
	5	12,341	6.8	181.1
Cohabitation	Living alone	44,153	11	873.7
	Not living alone	93,927	23	191.9
Year	2005	15,559	3.3	475.5
	2006	15,241	3.3	459.5
	2007	14,925	3.3	452.5
	2008	14,398	3.3	429.8
	2009	14,205	3.3	421.2
	2010	13,842	3.3	407.6
	2011	12,968	3.4	380.0
	2012	12,926	3.4	377.0
	2013	12,256	3.4	355.6
	2014	11,760	3.4	339.5
Magnesium exposure [mg/l]	≤ 6.65	28,212	6.8	417.1
	]6.65,10.3]	27,559	6.8	407.3
	]10.3,14.6]	27,254	6.7	404.0
	]14.6,21.9]	28,355	6.7	420.1
	> 21.9	26,700	6.7	396.7

Table 5.3: Table showing a list of all characteristics being part of the analysis along with the number of cardiovascular deaths, the amount of risk time and the unadjusted incidence rate in each group.

From Table 5.3 it seems like the age category has an enormous effect along with the income and cohabitation status. From the calendar year attribute it seems like there is a downward trend towards lower incidence rates. The magnesium exposure groups seem to perhaps have a slight effect. The gender, however, seems to have almost no impact and this is a place where one should indeed be careful with the conclusions.

5.4.1 The confounding effect of age on gender

Being careful with making any conclusions based on unadjusted incidence rates is important and an example of how it can go wrong is described in this subsection.

The incidence rates per age category divided between men and women show an interesting picture as seen in Figure 5.9. Here the power of the age category is still very apparent with incidence rates close to zero for age groups below 45 and an almost exponential growth of the risk. In Table 5.3, men and women seemed to have almost identical incidence rates, but from this figure it is now visible that it is not really the case. It can be seen that in all age categories men have much higher incidence rates than women.

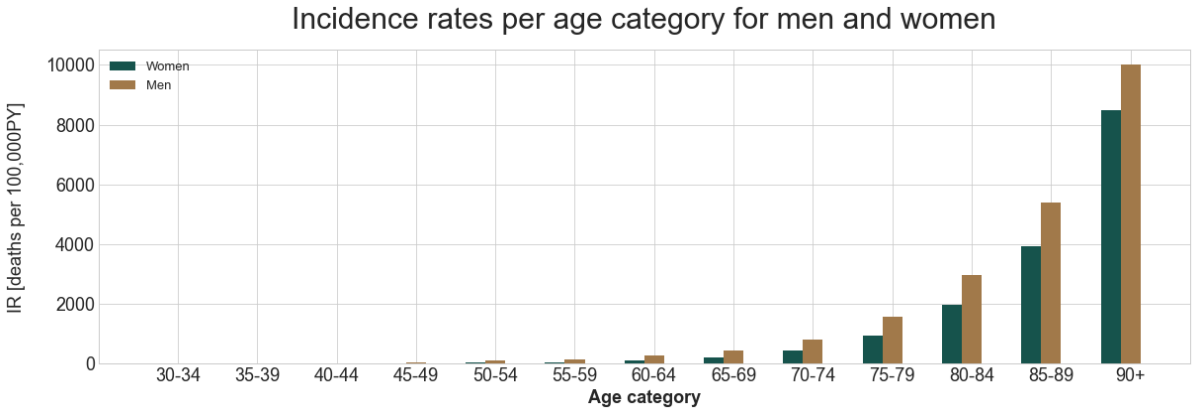

Figure 5.9: Unadjusted incidence rates of men and women separately per age category.

In order to assess the actual difference between the men and women, the incidence rate ratios (IRR) have for each age category been calculated and are shown in Figure 5.10

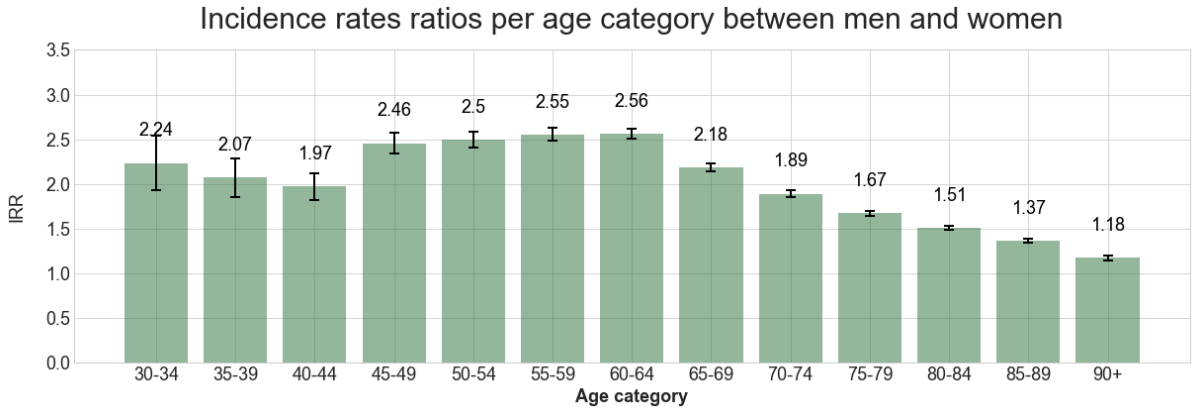

Figure 5.10: The calculated unadjusted incidence rate ratios between men and women per age category.

It is evident that in all age categories the IRR is higher than the overall incidence rate ratio between men and women of 1.01 calculated from the values in Table 5.3. In most categories men even have more than double the risk compared to women. This is an example of the classic Simpsons paradox in statistics, where grouping of data can hide certain trends. In this case it is due to the fact that women in general are older than men, thus dragging up their overall incidence rate. This shows how important it is to include all the true confounders in the analysis.

Another thing that is evident from Figure 5.10 is that the IRR is not constant across all categories which indicates that there might be an interaction between the two attributes as suggested in

the method chapter. This will be investigated in the sensitivity analysis in the following section.

5.4.2 Subcategories of cardiovascular deaths

All the aforementioned incidence rates are those of cardiovascular deaths (CD). This includes a broad variety of deaths and therefore it has been divided into subcategories based on the ICD codes as described in the Data chapter. In Figure 5.11 it is shown how many deaths occurred within each subcategory during the entire study period.

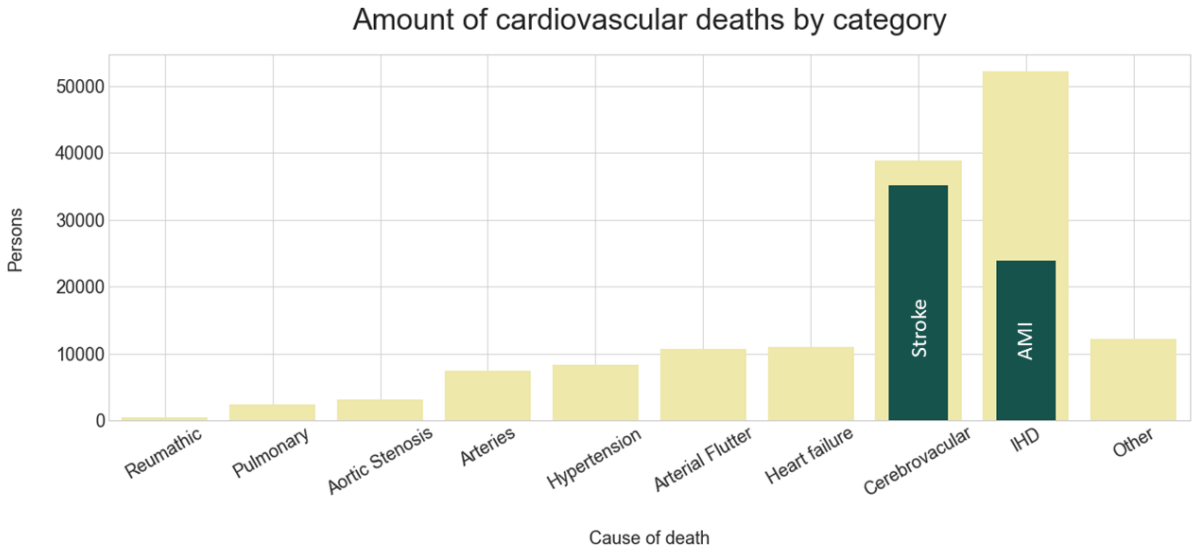

Figure 5.11: Number of deaths categorised by ICD codes. Green bars show number of cerebrovascular deaths defined as stroke and number of ischemic heart diseases defined as acute myocardial infarction.

As seen in the figure, the two largest categories are ischemic heart diesease (IHD) and cerebrovascular diseases of which almost all deaths are due to a stroke. For the IHD category almost half of the deaths are due to one disease namely acute myocardial infarction (AMI). In particular IHD and AMI have been the main subject of relevant studies examining magnesium in drinking water. Therefore it is of interest to see whether the magnesium exposure seems to have a different effect depending on the casue of death investigated.

In Figure 5.12 the unadjusted incidence rates of each exposure group are shown for three different types of outcome, namely death from all cardiovascular diseases, from IHD and from AMI specifically.

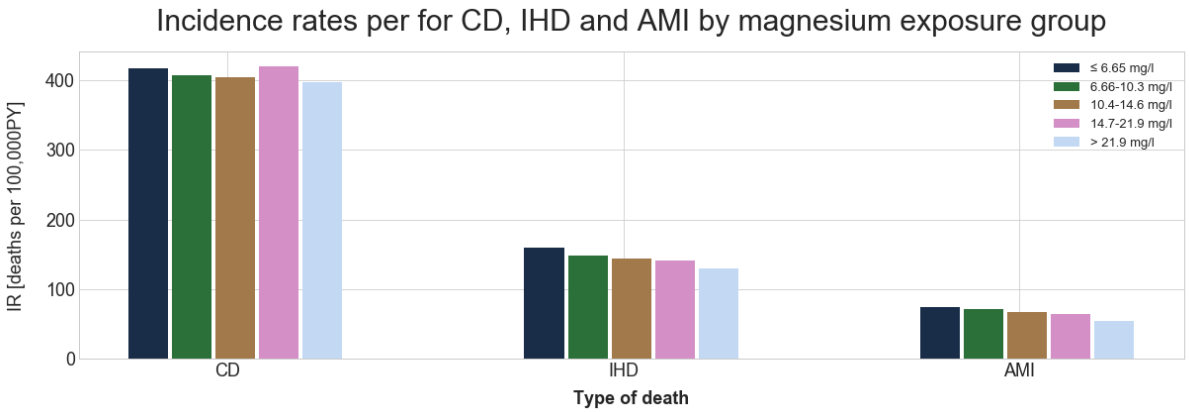

Figure 5.12: The incidence rates for CD, IHD an AMI per exposure group.

The first noticeable thing is of course that the rates are a lot lower for IHD and in particular for AMI. But it should be noted that they are still common causes of death. The absolute difference between the highest and lowest exposed group within each type of death seem similar, but what is of more interest is the relative risk. In Figure 5.13 the unadjusted incidence rate ratios between exposure groups 1 and 5 are shown for the three different outcomes.

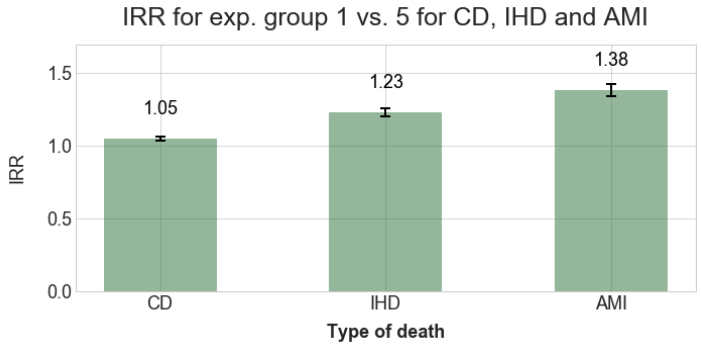

Figure 5.13: Incidence rate ratios between the lowest and highest exposed groups for CD, IHD and AMI separately.

The IRR is for all outcomes above 1 which indicates a negative effect of low magnesium exposure, however, it is clear that the IRR is particularly high for AMI. That the IRR of IHD is lower than the AMI IRR could indicate that the other half of IHD cases are not affected by magnesium as much as AMI. The confidence intervals are small for all three IRR and thus indicates a significant role of magnesium, however, it is important to emphasise that these are all unadjusted rates and could be an expression of geographical differences in the confounders. In the following section these will be taken into account through multiple Poisson regression.

5.5 Statistical analysis

The statistical analysis of the final data set is carried out using a multiple Poisson regression as defined in the chapter on methods. The model referred to as the adjusted model is the main model that all analysis are based on. Only in the sensitivity analysis will alternative models be assessed.

The examined outcome is based on the results from the descriptive analysis and the literature. This means that the overall cardiovasular deaths are examined along with the subgroups stroke

and IHD and the subgroup of IHD, AMI. The results are presented in Figure 5.14 which shows a forest plot of the incidence rate ratios with the highest exposure group as reference (i.e IRR=1) for all four outcomes separately. The incidence rates are shown as small circles and their 95% confidence intervals are shown as lines.

As can be seen in the figure, the pattern of the descriptive analysis is visible here as well. A high magnesium exposure has the highest impact on AMI with all IRR being significantly greater than 1. The IRR between exposure groups 1 and 5 is the greatest with a value of 1.24, meaning that individuals exposed to less than 6.65 mg/l have 24% greater risk of dying from AMI than individuals exposed to more than 21.9 mg/l. For IHD the pattern is similar to that of AMI but less strong. Here the IRR between group 1 and 5 is 1.12 and for the rest of groups it is 1.06 or less. For stroke no meaningful pattern can be observed and the IRR all lie close to 1 indicating no effect of a high magnesium exposure. The same is the case for overall cardiovascular deaths.

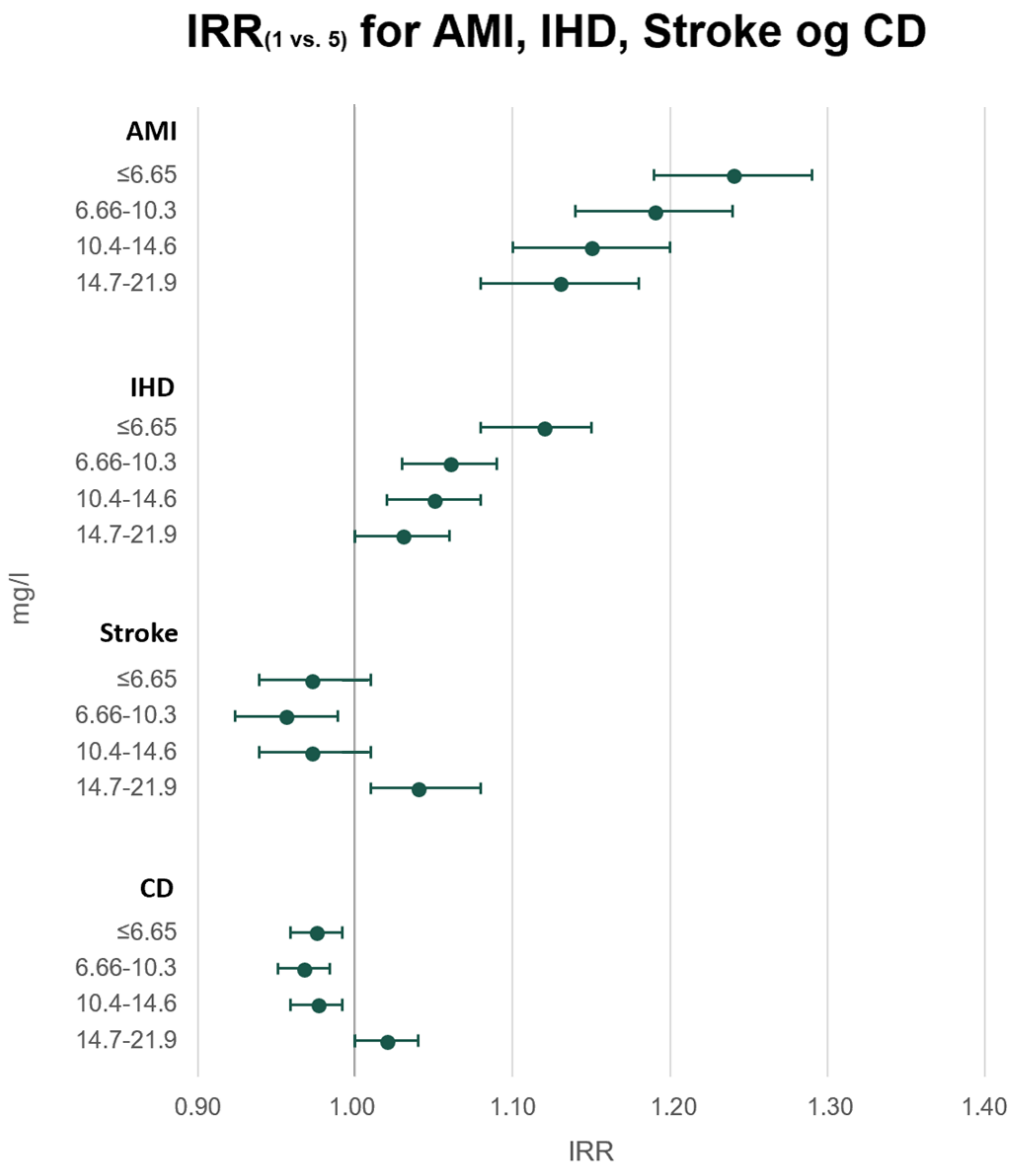

Figure 5.14: Forest plot showing the incidence rate ratios between the highest exposed group (> 21.9 mg/l) as reference and all other groups for acute myocardial infarction, ischemic heart disease, Stroke and all cardiovascular diseases respectively.

5.5.1 Sensitivity analysis

The results of the analysis depend on the model used for analysis and it is important to examine whether changes in the model specifications could potentially change the results. Therefore a sensitivity analysis is carried out in this section. If the results prove to be robust to changes in the model, it is more likely that the results are valid. However, many considerations should go into the evaluation of the validity of the results and these will be discussed in the next chapter.

Model with interactions

The specifications of the adjusted model include all the potential confounders available for this study. However, it does not include any interactions or effect modifiers. Therefore a model has been specified containing interactions between age categories and gender as well as age categories

and cohabitation. As seen in the descriptive analysis there was an indication of non constant incidence rate ratios between men and women over the age categories. A very similar pattern can be seen between age categories and cohabitation, these plots can be found in appendix C.

The results from this model are reported next to the equivalent result of the unadjusted model and the adjusted model in Table 5.4 and Table 5.5 for IHD and AMI respectively. The unadjusted model will yield the same results as in the descriptive analysis and the adjusted model is the same one as used for the results in Figure 5.14.

Incidence rate ratios for IHD death						
Exposure [mg/l]	**Unadjusted model**		**Adjusted model**		**Model with interactions**	
	IRR	**95% CI**	**IRR**	**95% CI**	**IRR**	**95% CI**
≤6.65	1.23	(1.20-1.27)	1.12	(1.08-1.15)	1.13	(1.10-1.16)
]6.65,10.3]	1.14	(1.11-1.18)	1.06	(1.03-1.06)	1.07	(1.04-1.08)
]10.3,14.6]	1.11	(1.0.8-1.14)	1.05	(1.02-1.08)	1.05	(1.02-1.08)
]14.6,21.9]	1.08	(1.05-1.12)	1.03	(1.00-1.06)	1.04	(1.01-1.07)
>21.9	1	-	1	-	1	-

Table 5.4: Table showing the incidence rate ratios for IHD between exposure group 5 as reference and all other exposure groups. The IRRs are shown for the unadjusted model, the adjusted model and the model with interactions.

Incidence rate ratios for AMI death						
Exposure [mg/l]	**Unadjusted model**		**Adjusted model**		**Model with interactions**	
	IRR	**95% CI**	**IRR**	**95% CI**	**IRR**	**95% CI**
≤6.65	1.38	(1.33-1.44)	1.24	(1.19-1.29)	1.25	(1.19-1.30)
]6.65,10.3]	1.31	(1.26-1.37)	1.19	(1.14-1.24)	1.20	(1.15-1.25)
]10.3,14.6]	1.24	(1.19-1.30)	1.15	(1.10-1.20)	1.16	(1.11-1.21)
]14.6,21.9]	1.20	(1.14-1.25)	1.13	(1.08-1.18)	1.13	(1.08-1.18)
>21.9	1.00	-	1.00	-	1.00	-

Table 5.5: Table showing the incidence rate ratios for AMI between exposure group 5 as reference and all other exposure groups. The IRRs are shown for the unadjusted model, the adjusted model and the model with interactions.

The result from the model with interactions are very similar to the ones from the adjusted model with just slightly larger IRRs. Thus, the results seem robust to the addition of these interaction terms.

Models with effect modification

Another possible misspecification of the original adjusted model is that there might exist some effect modifiers. The effect of magnesium exposure might be affected by some of the confounders and therefore it was examined whether adding this effect to the model would change the results. The effect modifiers taken into account was both gender and age categories. However, adding the gender and age separately did not prove to have a significant effect, but adding an effect modification as a combination of both of them was significant. All the incidence rates resulting from that are not reported here, but instead an analysis based only on individuals aged 70 or more shows that high levels of magnesium might actually be more important to elderly people than younger.

The analysis based on 70+ year-olds and the original adjusted model gives incidence rate ratios between the highest and lowest exposed group of 1.28 with the confidence interval ranging from 1.22 to 1.35 for AMI. This is should be compared to the 1.24 IRR from the main analysis.

Finally, an attempt in relation to effect modification was made to see if the effect of magnesium might change over time. Using the calender year to modify the effect was significant but when looking at the incidence rate ratios for each year separately, it was evident that no trend was present and that by adding this effect only some random variation was modelled. Since this is not of interest, these specific results are also not reported here.

Analysis including private wells

In the main analysis and all sensitivity analysis so far, the individuals supplied by their own well were not included. However, most of their addresses are located within water supply areas and they could thus be included by assuming their exposure was equal to the exposure of the surrounding area. As mentioned earlier, it is uncertain whether that is a valid assumption or not. In order to examine this a bit further, a sensitivity analysis including these individuals has been carried out. This meant that more than half a million more observations were added to the final data set. The results of this extended analysis were almost identical to the results of the main analysis. When rounded to three significant digits no differences in the IRRs were present. Thus, it can be concluded that assuming the exposure of the private well owners was the same as in the surrounding area did not change the results at all.

Chapter 6

Discussion

In this chapter the study and project in general will be discussed. This includes reflections and recommendations of further work that could be done to affirm the results or potentially reject them. The first part of the discussion will include a summary of results and a discussion on their validity. Then the strengths and limitations of the study and study design will be discussed. After that there is a discussion on the estimations of magnesium and the matching process of addresses to WSAs. At the end the perspectives of the study will be discussed.

6.1 The results

The results of the present study showed a significant protective effect of magnesium in drinking water on ischemic heart disease and in particular on acute myocardial infarction. The results showed that individuals from the least exposed group have 24% greater risk of dying from AMI than individuals from the most exposed group. In fact, the analysis showed a significantly greater risk for all exposure groups compared to the highest exposure group. A sensitivity analysis also indicated that older individuals might be more affected by the magnesium concentration of their drinking water. They had a 28% increased risk of AMI for the lowest versus highest exposure group.

For both cardiovascular death in general and specifically stroke no significant and meaningful relations between incidence rates and magnesium exposure levels were found.

6.1.1 Validity of results

The results of the present study indicate that the magnesium intake from drinking water constitutes such a large part of the total magnesium intake that it actually makes a difference towards the risk of dying from AMI. In order to understand if this indication is plausible, it is necessary to see the concentrations of magnesium in the Danish drinking water in relation to the recommended and actual intakes described in Chapter 2.

If it is assumed that individuals drink between 1 and 2 litres of water (including tea and coffee) every day, then individuals from the highest exposed group would consume 22-107 mg/day of magnesium and individuals from the least exposed group would only consume 0-13 mg/day of magnesium. It is clear that there is a real difference between the intakes. The recommendations are around 300 mg/day for women and 400 mg/day for men. For the highest exposed group the daily intake could thus constitute up to 33% for women and 25% for men of the recommended intake. For individuals not getting enough magnesium through their diet, it could even constitute a greater part of the actual intake. However, that would only be the case for individuals who are

exposed to very high concentrations and who drink much water. A more realistic suggestion is probably around 10-20% of the daily intake. Comparing this to individuals in the least exposed group, they would only get 0-4% of their daily recommended intake from their drinking water. If their diet is very low in magnesium, it could constitute slightly more of the actual intake.

These considerations makes it seem plausible that the magnesium intake through drinking water could actually have an effect.

Whether it can be justified clinically that the association is only present for IHD and not cerebrovascular diseases (e.g. stroke) would require expert knowledge that is outside the scope of this project.

Comparison with other studies

In Chapter 2, eleven studies on the association of magnesium in drinking water and cardiovascular death were described. They showed somewhat contradicting results. Five of them showed a significant protective effect of magnesium in drinking water on IHD or AMI. These results are in line with the results from the present study. Two Swedish studies showed an increased risk of AMI of 35% for men and 30% for women exposed to low concentrations of magnesium compared to high concentrations. This is similar to the increase in risk found in the present study. Two studies showed a significant effect of magnesium against cerebrovascular disease (including stroke), however, two other studies also examined this and did not find any connection. This is also the case for the present study. Four of the eleven studies found no associations at all. Two of these were ecological studies, where one from Japan had only very high concentrations above 35 mg/l. Compared to the present study, all the Japanese individuals would be in the highest exposure group. The third study that found no association was the study from the Netherlands. In this study an enormous amount of confounders were used in the model, including hypertension. Since hypertension is associated with magnesium deficiency it might be unsuitable to view it as a confounder. The fourth study not finding any significant associations was another Swedish study. In this study the maximum intake of magnesium through drinking water was assessed to be 4 mg/l. With such a small intake it is not surprising that no association is found. Compared to the present study, all individuals in this Swedish study would probably belong to exposure group 1.

6.2 Strengths of the present study

Some of the limitations highlighted in the studies above are not of concern in the present study. This is the case for the exposure interval. The interval is quite wide, ranging from almost zero up to more than 50 mg/l. The wide range makes it possible to find an association as opposed to studies where the range is very narrow.

A strength of the present study and a general strength of all register-based studies is that less or almost no selection and attrition bias exists [50]. Selection and attrition bias refers to the bias of who is included in the study and who leaves the study ahead of time. The reason for these biases almost not being present is that the entire Danish population is included in the study and no one leaves the study unless it is from fatal event or they leave the country. Potentially, a slight bias could be incurred from the individuals leaving the country, but it is not of concern.

That the entire Danish population is included in the study also means that each magnesium exposure group actually represent approximately 20% of the population. It is an actual 20%

who lives with the elevated risk. If a study was based on a small group with particularly low exposures, then it would be harder to generalise the results and prove that the increased risk is a problem concerning a large part of the actual population.

The register data also has a high validity and completeness. Only very little information is missing and it is never up to the individuals themselves to record anything, thus many errors and inaccuracies are avoided. If any mis-classifications do exist in the data, they will be non-differential and independent [50]. Furthermore, the validity of the AMI diagnosis in the Danish register on Causes of Death is considered high [51].

Another thing that differentiates the present study from the other studies, is that all variables vary over time. The other studies simply take the current magnesium exposure without taking any variations over time into account. Often the measurements of magnesium concentrations are taken long before or after the incidences of CD are recorded. In the present study the exposure is calculated from concurrent estimates. The present study is also the only one assessing a two-year average as the exposure. A sensitivity analysis examining alternative calculations of exposure, e.g. a five-year or one-year average, could be conducted to see if the results were robust to such a change.

6.3 Limitations of the present study

The design of this epidemiological study could have been different in many ways. Part of the design is limited by external possibilities and part of the design relies on arbitrary decisions.

The amount of data available for the study was an external limitation. Considering the time-span of the magnesium estimates, it would be possible to extent the study period both back and forth in time. This would require that the data access was granted by Statistics Denmark.

That relevant confounders, such as the lifestyle, is not included, is of course also due to limitations in the data available. This type of information would require more than the register data and therefore also another type of study that included surveys. However, this would change many elements of the study, for example the size of the study population, that would need to be much smaller. The confounder called *family income* is used in the present study as a way of taking lifestyle into account. It is not the low income itself that elevates the risk of cardiovascular death, but rather the lifestyle associated with it. This association might exist, but it will not be the case for all individuals, e.g. a high income does not guarantee a healthy lifestyle. Therefore it will not be as accurate as the true information on lifestyle. An analysis of how strong the association between income and lifestyle actually is would benefit the study.

In general, a more thorough analysis of the confounders and their association with exposure and outcome should be done as an extension of this project. It could involve an analysis and creation of a Directed Acyclic Graph (DAG) based on expert knowledge from cardiologists among others. This could also clarify the justification of including interactions or effect modifiers.

One of the general limitations of register-based epidemiological studies is the fact that data on potentially important confounders does not exist in the registers [50]. Even if a DAG showed the need to include more confounders, they would likely not be available. However, if unmeasured confounders exist, it is still possible to evaluate the results. Grøn et al. [48] suggested several methods for large-scale register-based studies. This includes a method of introducing unmeasured confounding and calculating how strong a confounder this would need to be in order to change the significance of the results. It also includes alternative ways of calculating confidence

intervals for the Poisson regression. Other studies have also suggested sensitivity analysis to estimate the effect of an unmeasured confounder [52].

Another limitation is that the huge amount of data involved in the study might make parameters statistically significant which are not clinically relevant [50]. This was seen in the sensitivity analysis, where the calendar year proved to be a significant effect modifier. However, it was not because a clinically relevant trend was present, but simply due to random variations over the study period.

6.4 Magnesium estimates

When looking at the data set of magnesium measurements, it is clear that some form of estimation of concentrations is needed. If a waterworks has almost constant magnesium levels, and only a few years exist with no measurements, then it seems obvious to assume a similar concentration in the few years without measurements. However, for some waterworks only very few measurements are taken and even 25% of the waterworks have 3 measurements or less taken since year 2000. For these, the ease of estimation is not so obvious.

Another factor that complicates the estimation is the variation in measurements within a waterworks. If a waterworks have measurements that vary greatly, how do we know the true level of magnesium? To assess this issue, the standard deviation of each waterworks was calculated and it was found that 75% of them had a standard deviation of 1.3 mg/l or less. If the standard deviation is that low, it is safe to assume that for most areas the concentrations lies within a quite narrow range. This fact makes the estimations more valid in general.

A third factor that invokes uncertainty is the double connections where areas have several waterworks connected. In these areas it is assumed that all households get water from a mix of the connected waterworks. This assumption might be true in some cases and might be completely wrong in other cases.

These three factors can make you question the accuracy of the individual exposure used in the analysis. A more complicated model perhaps using a combination of the KNN and the geographical interpolation could have been attempted, but it seemed important as well to make the estimation process so transparent that it was clear which estimates went into the final data set. Another, perhaps better, way to handle the uncertainty could have been to label each estimate with an uncertainty category based on the amount of measurements at the waterworks and the variation within measurements. Then a sensitivity analysis could be conducted only using the more certain estimates to see if it makes a difference in the results. Alternatively, a weighted analysis could be conducted, giving higher weights to more certain estimates. However, these solutions could also introduce some bias since it is mostly smaller areas that have few measurements.

With all this being said about the uncertainty of the estimates, it is also important to stress that the average cross validation error for the estimation process was only -1.22 mg/l. An error this small would probably not change the analysis since the highest exposure group is exposed to at least 15 mg/l more than the lowest exposed group. The fact that the magnesium level is dependent on the geographical area also makes the issue of double connections minor. If the waterworks are close to each other, it is also likely that the water they abstract is similar, thus making it less relevant which exact households they supply.

6.5 Addresses linked to WSAs

When the observations of the register data were linked to water supply areas, more than 1.5 million observations could not be linked. Some observations were not linked because they were supplied by a private well, others were not linked because their residence was placed in an area without any estimates, and some were not linked due to the fact that their address did not have a match in the data set of addresses.

The last group constitutes more than half a million observations and it is somewhat uncertain why their address did not exists in the data set. One explanation could be that the address data set contains the current addresses in Denmark and thus addresses that has been discontinued are not in the data set. In order to reduce the issue of failed matching, an improved matching could include discontinued addresses, thus making it possible to include these observations in the analysis.

The individuals excluded because they lived either in an area with no measurements (and thus no estimates of magnesium) or lived outside any of the defined water supply areas could also be attempted included in the analysis. This could be done using the geographically nearest estimates as estimates of their exposure.

The individuals excluded due to receiving water supply from own private well were included in a sensitivity analysis. Here it was assumed that their exposure was the same as the exposure of the surrounding area. The sensitivity analysis showed results similar to the main analysis. It is, however, still not proven that including actual samples from private wells would not change the results. This would require water samples taken from all wells analysed for magnesium content and these data are to date non-existent.

6.6 The perspectives of the study

In order to validate the results further, it is recommended to first of all make an extensive analysis of confounders as suggested in the discussion above. This requires further involvement of experts with knowledge of specifically cardiovascular diseases. If it is agreed upon that the results makes sense in a clinical context and that no further confounders should be included, then the sensitivity analysis also suggested above could be carried out to make sure the results are not changed.

Assuming that the validity of the results are upheld, then this study could contribute to the discussion about water softening as well as recommendations of diet or magnesium supplements. In particular, individuals living in areas with very low magnesium concentrations in the drinking water could be recommended a diet that is rich in magnesium. For individuals refusing to live on a magnesium rich diet, perhaps a daily supplement of magnesium could benefit them.

The discussion on water softening is more of a political discussion, since it is up to the authorities to adopt the appropriate legislation on water quality. As stated in the aforementioned report form COWI A/S there can be a financial benefit of decreasing water hardness in areas with particularly hard water. However, the results from this study indicates that reducing the water hardness and thus removing the magnesium could lead to many more cases of AMI. This is something that definitively needs to be taken into consideration before water softening becomes a standard way of treating water at waterworks.

Chapter 7

Conclusion

The present study has examined the hypothesis presented in the introduction, namely that magnesium in drinking water has a positive effect on the risk of cardiovascular death. The study has found evidence both for and against this hypothesis. In particular, an association between magnesium in drinking water and the risk of acute myocardial infarction has been found. The lowest exposed group ($\leq$ 6.65 mg/l) is found to have an increased risk of 24% of dying from AMI compared to the highest exposed group ($>$ 21.9 mg/l). The result has been affirmed in several sensitivity analysis, that also indicates an even greater protective effect of magnesium in drinking water for elderly people. The evidence against the hypothesis is that no significant association between magnesium in drinking water and the overall cardiovascular death has been found.

Before the results of the study can definitively contribute to the pool of knowledge and official recommendations, further work has to be carried out. This includes a deeper analysis of confounders and several extensive sensitivity analysis of the Poisson regression.

If the results are robust to further sensitivity analysis, then the present study could have an impact on public health. This is both regarding official recommendations on diet in specific areas and legislation on water quality.

The study has also examined different ways of handling the magnesium samples from the different waterworks. Several methods for estimating the true concentrations of the water supply areas has been attempted. Using a K-nearest neighbours approach with four neighbours and an inverse distance weighting proved to be the best suited of the assessed methods. Further methods could be suggested or the uncertainty of each estimate could be taken into account through further sensitivity analysis. It has been shown how the magnesium concentrations differ across the country, with low concentrations primarily in Jutland.

Bibliography

[1] C. N. Ong, A. C. Grandjean, and R. P. Heaney, "The mineral composition of water and its contribution to calcium and magnesium intake," in *Calcium and Magnesium in Drinking-water: Public Health Significance*, WHO, Ed., 2009, ch. 3.

[2] K. Wodschow, B. Hansen, and A. K. Ersbøll, "Stability of Major Geogenic Cations in Drinking Water — An Issue of Public Health Importance : A Danish Study , 1980 – 2017," pp. 1–16, 2017.

[3] S. A. Atkinson, R. Costello, and J. M. Donohue, "Overview of global dietary calcium and magnesium intakes and allowances," in *Calcium and Magnesium in Drinking-water: Public Health Significance*, WHO, Ed., 2009, ch. 2.

[4] E. Panel and A. Nda, "Scientific Opinion on Dietary Reference Values for magnesium," *EFSA Journal*, vol. 13, no. 7, p. 4186, 2015.

[5] L. TROPPMANN, K. GRAY-DONALD, and T. JOHNS, "Supplement use: Is there any nutritional benefit?" *Journal of the American Dietetic Association*, vol. 102, no. 6, pp. 818–825, jun 2002.

[6] Y. Song, J. E. Manson, N. R. Cook, C. M. Albert, J. E. Buring, and S. Liu, "Dietary Magnesium Intake and Risk of Cardiovascular Disease Among Women," *The American Journal of Cardiology*, vol. 96, no. 8, pp. 1135–1141, oct 2005.

[7] A. M. Jodral-Segado, M. Navarro-Alarcón, H. López-G de la Serrana, and M. C. López-Martı´nez, "Magnesium and calcium contents in foods from SE Spain: influencing factors and estimation of daily dietary intakes," *Science of The Total Environment*, vol. 312, no. 1-3, pp. 47–58, aug 2003.

[8] P. Galan, M. J. Arnaud, S. Czernichow, A.-M. Delabroise, P. Preziosi, S. Bertrais, C. Franchissaeur, M. Maurel, A. Favier, and S. Hercberg, "Contribution of Mineral Waters to Dietary Calcium and Magnesium Intake in a French Adult Population," *Journal of the American Dietetic Association*, vol. 102, no. 11, pp. 1658–1662, nov 2002.

[9] R. D. Abbott, F. Ando, K. H. Masaki, K.-H. Tung, B. L. Rodriguez, H. Petrovitch, K. Yano, and J. Curb, "Dietary magnesium intake and the future risk of coronary heart disease (The Honolulu Heart Program)," *The American Journal of Cardiology*, vol. 92, no. 6, pp. 665–669, sep 2003.

[10] W. Becker and J. Kumpulainen, "Contents of essential and toxic mineral elements in Swedish market-basket diets in 1987," *British Journal of Nutrition*, vol. 66, no. 2, pp. 151–160, 1991.

[11] W. B. Weglicki, "Magnesium efficiency: Clinical and experimental aspects," in *Calcium and Magnesium in Drinking-water: Public Health Significance*, WHO, Ed., 2009, ch. 5.

[12] R. M. Touyz and B. Sontia, "Magnesium and hypertension," in *Calcium and Magnesium in Drinking-water: Public Health Significance*, WHO, Ed., 2009, ch. 6.

[13] B. M. Altura and B. T. Altura, "Atherosclerosis and magnesium," in *Calcium and Magnesium in Drinking-water: Public Health Significance*, WHO, Ed., 2009, ch. 7.

[14] "Danmarks Statistik - vandforbrug," 2018. [Online]. Available: http://www.statistikbanken.dk/VAND2MU1

[15] J. H. Hankin, "CONTRIBUTION OF HARD WATER TO CALCIUM AND MAGNESIUM INTAKES OF ADULTS," *Journal of the American Dietetic Association*, vol. 56, no. 3, 1970.

[16] J. DURLACH, "Magnesium Level in Drinking Water and Cardiovascular Risk Factor, a Hypothesis," *Magnesium*, vol. 4, no. 1, 1985.

[17] M. R. H. Lowik, E. H. Groot, and W. T. Binnerts, "Magnesium and Public Health: The Impact of Drinking Water," *Trace Substances in Environmental Health: Proceedings of University of Missouri's Annual Conferenc*, 1982.

[18] M. Sabatier, "Meal effect on magnesium bioavailability from mineral water in healthy women," *American Journal of Clinical Nutrition*, vol. 75, no. 1, 2002.

[19] UNESDA, "Sales volume data, collected by GlobalData," 2016. [Online]. Available: https://www.unesda.eu/products-ingredients/consumption/

[20] European Federation of Bottled Water, "Key Statistics," 2016. [Online]. Available: http://www.efbw.org/index.php?id=90

[21] J. Durlach, M. Bara, and A. Guiet-Bara, "Magnesium level in drinking water: Its importance in cardiovascular risk," *Magnesium in Health and Disease*, pp. 173–182, 1989.

[22] M.-P. Sauvant and D. Pepin, "Drinking water and cardiovascular disease," *Food and Chemical Toxicology*, vol. 40, no. 10, pp. 1311–1325, oct 2002.

[23] R. Rylander, H. Bonevik, and E. Rubenowitz, "Magnesium and calcium in drinking water and cardiovascular mortality," *Scand J Work Environ Health*, vol. 17, no. 2, pp. 91–94, 1991.

[24] R. Rylander, "Magnesium in drinking water and cardio-vascular disease–an epidemiological dilemma," *Clinical Calcium*, vol. 15, no. 11, 2005.

[25] E. Rubenowitz, G. Axelsson, and R. Rylander, "Magnesium in drinking water and death from acute myocardial infarction," *Epidemiology (Cambridge, Mass.)*, vol. 143, no. 5, pp. 456–462, 1996.

[26] M. Rosenlund, "Environmental Factors in Cardiovascular Disease," Ph.D. dissertation, Karolinske Institutet, 2005.

[27] A. Gimeno Ortiz, R. Jiménez Romano, M. Blanco Aretio, and A. Castillo Moreno, "Relationship of several physico-chemical components in drinking water, hypertension and cardiovascular disease mortality," *Revista De Sanidad E Higiene Publica*, vol. 64, no. 7-8, 1990.

[28] C. Yang, "Calcium and magnesium in drinking water and risk of death from cerebrovascular disease," *Stroke*, vol. 29, no. 2, 1998.

[29] L. J. Leurs, "Research Relationship between Tap Water Hardness, Magnesium, and Calcium Concentration and Mortality due to Ischemic Heart Disease or Stroke in the Netherlands," 2010.

[30] H. Luoma, A. Aromaa, S. Helminen, H. Murtomaa, L. Kiviluoto, S. Punsar, and P. Knekt, "Risk of Myocardial Infarction in Finnish Men in Relation to Fluoride, Magnesium and Calcium Concentration in Drinking Water," *Acta Medica Scandinavica*, vol. 213, no. 3, pp. 171–176, apr 2009.

[31] R. Maheswaran, S. Morris, S. Falconer, A. Grossinho, I. Perry, J. Wakefield, and P. Elliott, "Magnesium in drinking water supplies and mortality from acute myocardial infarction in north west England," *Heart*, vol. 82, no. 4, pp. 455–460, oct 1999.

[32] Y. Miyake and M. Iki, "Ecologic Study of Water Hardness and Cerebrovascular Mortality in Japan," *Archives of Environmental Health: An International Journal*, vol. 58, no. 3, pp. 163–166, mar 2003.

[33] M. P. Sauvant and D. Pepin, "Geographic variation of the mortality from cardiovascular disease and drinking water in a french small area (Puy de Dome)," pp. 219–227, 2000.

[34] "HOFOR - blødere vand," 2018. [Online]. Available: https://www.hofor.dk/baeredygtige-byer/udviklingsprojekter/bloedere-vand/

[35] Cowi A/S, *Central blødgøring af drikkevand*, 2011, no. April.

[36] J. Schullehner and B. Hansen, "Nitrate exposure from drinking water in Denmark over the last 35 years," *Environmental Research Letters*, vol. 9, no. 9, p. 095001, sep 2014.

[37] "Styrelsen for Dataforsyning og Effektivisering - Danmarks Adressers Web API," 2018. [Online]. Available: https://dawa.aws.dk/

[38] H. Christiansen, "CPR-oplysninger," 2018. [Online]. Available: https://www.dst.dk/da/Statistik/dokumentation/Times/cpr-oplysninger

[39] C. B. Pedersen, "The Danish Civil Registration System," in *Scandinavian Journal of Public Health, volume 39, supplement 7*, L. Thygesen and A. Ersbøll, Eds., 2011, pp. 22–25.

[40] "Danmarks statistik - familietype," 2018. [Online]. Available: https://www.dst.dk/da/Statistik/dokumentation/Times/forebyggelsesregistret/familietype

[41] K. Helweg-Larsen, "The Danish Register of Causes of Death," in *Scandinavian Journal of Public Health, volume 39, supplement 7*, L. C. Thygesen and A. K. Ersbøll, Eds. SAGE, 2011.

[42] M. Baadsgaard and J. Quitzau, "Danish registers on personal income and transfer payments," in *Scandinavian Journal of Public Health, volume 39, supplement 7*, L. Thygesen and A. Ersbøll, Eds., 2011, pp. 105–105.

[43] "Danmarks Statistik - familieækvivalensindkomst," 2018. [Online]. Available: https://www.dst.dk/da/TilSalg/Forskningsservice/Dokumentation/hoejkvalitetsvariable/familieindkomst/famaekvivadisp

[44] K. J. Rothman, *Epidemiology: An Introduction*. Oxford University Press, 2012.

[45] K. J. Rothman, S. Greenland, and T. L. Lash, *Modern epidemiology*, 2008.

[46] J. A. Udell, P. G. Steg, B. M. Scirica, S. C. Smith, E. M. Ohman, K. A. Eagle, S. Goto, J. I. Cho, D. L. Bhatt, and R. A. Continu, "Living Alone and Cardiovascular Risk in Outpatients at Risk of or With Atherothrombosis," *Archives of Internal Medicine*, vol. 172, no. 14, pp. 1086–1095, 2012.

[47] W. D. Dupont, *Statistical modelling for biomedical researchers.* Cambridge Medicine, 2009.

[48] R. Grøn, T. A. Gerds, and P. K. Andersen, "Misspecified poisson regression models for large-scale registry data: Inference for 'large n and small p'," *Statistics in Medicine*, vol. 35, no. 7, pp. 1117–1129, 2016.

[49] S. Support, "Proc Genmod documentation," 2018. [Online]. Available: https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_genmod_sect010.htm

[50] L. C. Thygesen and A. K. Ersbøll, "When the entire population is the sample: Strengths and limitations in register-based epidemiology," *European Journal of Epidemiology*, vol. 29, no. 8, pp. 551–558, 2014.

[51] M. Madsen, M. Davidsen, S. Rasmussen, S. Z. Abildstrom, and M. Osler, "The validity of the diagnosis of acute myocardial infarction in routine statistics: A comparison of mortality and hospital discharge data with the Danish MONICA registry," *Journal of Clinical Epidemiology*, vol. 56, no. 2, pp. 124–130, 2003.

[52] R. H. H. Groenwold, D. B. Nelson, K. L. Nichol, A. W. Hoes, and E. Hak, "Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research," *International Journal of Epidemiology*, vol. 39, no. 1, pp. 107–117, 2010.

Appendices

Appendix A

SAS example code

```
title1 'Poisson regression, cardiovascular death';

*Poisson regression
proc sort data=final3;
  by alder_cat koen enlig aar indkomstgrp mg_exp_grp;
run;
```

```
proc univariate data=final3 noprint;
  by alder_cat koen enlig aar indkomstgrp mg_exp_grp;
  var dodc rt;
  output out=final3_agg sum=sum_dodc sum_rt;
run;
```

```
data final3_agg2;
  set final3_agg;
  log_rt = log(sum_rt);
  if sum_dodc = . then sum_dodc = 0;
run;
```

```
proc genmod data = final3_agg2;
  class alder_cat koen enlig aar indkomstgrp mg_exp_grp;
  model sum _dodc = alder_cat koen enlig aar indkomstgrp mg_exp_grp / dist=p link=log offset=log_rt type3;
  estimate 'IRR 1 vs 5 mg' mg_exp_grp 1 0 0 0 -1 / exp;
  estimate 'IRR 2 vs 5 mg' mg_exp_grp 0 1 0 0 -1 / exp;
  estimate 'IRR 3 vs 5 mg' mg_exp_grp 0 0 1 0 -1 / exp;
  estimate 'IRR 4 vs 5 mg' mg_exp_grp 0 0 0 1 -1 / exp;
run;
```

Figure A.1: Example code of how the Poisson regression using proc genmod is carried out.

Appendix B

Maps

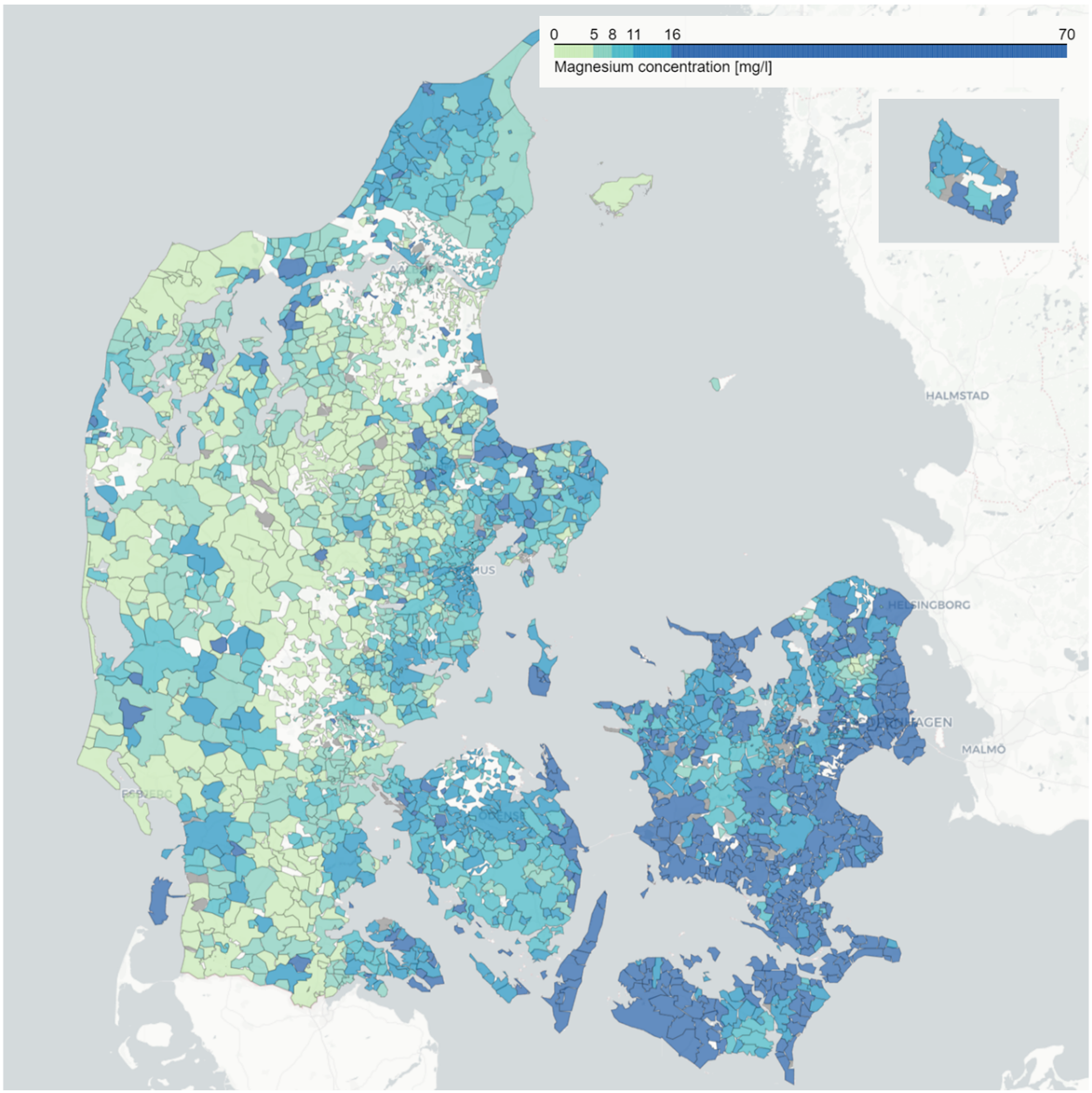

Figure B.1: Maps showing estiamted concentrations in 2005.

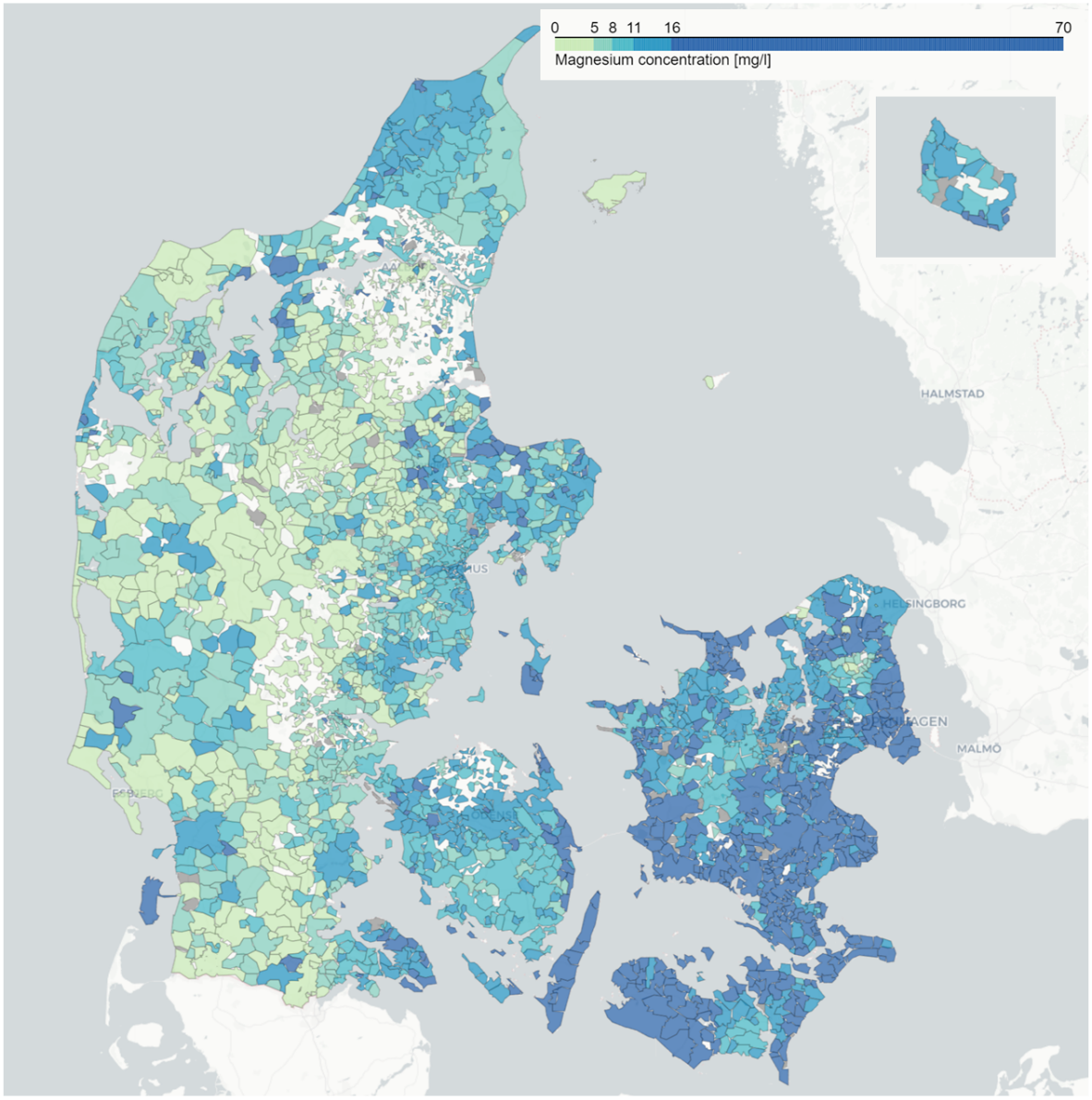

Figure B.2: Maps showing estimated concentrations in 2010.

Appendix C

Incidence rates of cohabitation per age category

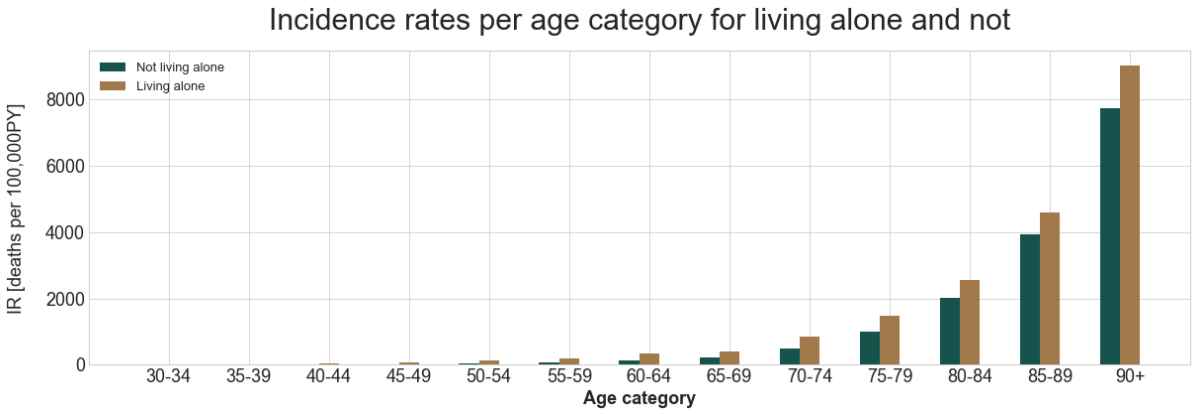

Figure C.1: IR for living alone and not living alone separately per age category

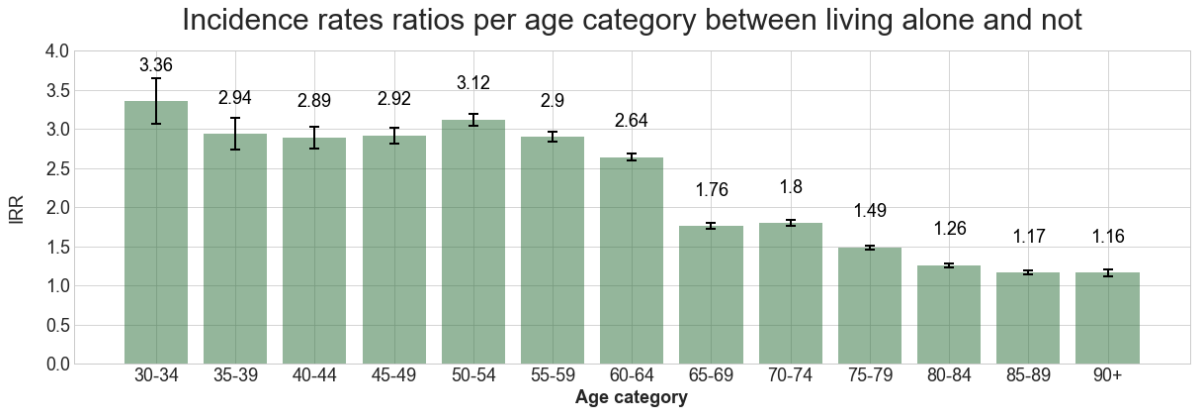

Figure C.2: IRR between living alone and not living alone per age category.