

ET BLIK IND I FREMTIDEN

Et prædiktivt studie af sygdomsforekomst i Region Nordjylland



Cecilie Møller Høymark & Kathrine Bendixen
Sociologiuddannelsen, 10. semester, forår 2019
Vejleder: Rolf Lyneborg Lund
Antal anslag: 235.892

Summary

In this master thesis, we examine the impact of 31 social and psychological determinants due to the risk of somatic chronic condition arising within a period of three years. First, we make a prediction in which individuals are at high risk of developing at least one condition in the future. Secondly, we map these predicted risk values in order to examine geographical inequalities in future prevalence of somatic chronic conditions.

The data was provided by Region Nordjylland. The first part is public administration data on groups of chronic conditions. The data covers year 2010 to 2016. We assembled six groups of conditions in one outcome measure for somatic chronic condition, including: osteoporosis, arthritis, allergy and other inflammatory diseases, cardiovascular diseases, asthma or chronic obstructive pulmonary disease, and diabetes. The second part is data from Sundhedsprofilen, a cross sectional national health survey, conducted in 2010, 2013 and 2017. In this survey, information about the individual's health status is framed by a holistic perspective, including not only information about illness and disease, but also social support and level of trust.

The outcome-variable in this study is a binary measure of somatic chronic conditions developed from the time of measurement of the explanatory variables and three years beyond. The value 0 is assigned if the person obtains no *new* somatic chronic condition within this period. The value 1 is assigned if the person obtains at least one of the included conditions within this period, provided that he or she did not possess this condition at the time of measurement or previously. Information on future somatic chronic conditions are joined to the survey data from the public administration data. The 31 explanatory variables are classified into six themes: socioeconomic conditions, mental wellbeing, social contact and support, health related behaviour, sense of belonging in the neighborhood, and level of trust. The research methods used in this study are logistic regression and neural network predictive modelling. These supervised machine learning techniques make it possible to conduct information from data, where the output is known, and use this information in constructing a model, that predicts this outcome on new data.

The purpose of using both logistic regression and neural network is to compare the prediction accuracy of a simple model versus a more complex model. Our model evaluation is also based on the applicability of each model to provide outputs that allow for us to make conclusions on statistical inference. Neural networks are known for their high levels of prediction accuracy but are often referred to as *black boxes*, meaning it is impossible to understand the causal mechanisms leading to predictions. Logistic regressions, on the other hand, offer parameter estimates and level of significance of the correlation between variables. Our final assessment is that a combination of the two models is the better option for answering the research question in this master thesis. In this way, we use both the full potential of the high predictive value in neural network analysis, and afterwards the predicted value for each individual can be used as an outcome measure in a logistic regression analysis to determine causal relationships.

The theory used for explaining the causal connections between social or psychological determinants and somatic chronic conditions are Pierre Bourdieu's notion of the link between social position and lifestyle. This theory is applied to the field of health by Simon J. Williams, and his perspectives are also used. We also use the perspectives of Jon Elstad on the psychosocial explanation of health inequalities. In explaining the geographical inequalities of health, we use Mike Savage, Gaynor Bagnall og Brian Longhurst's application of Bourdieu's concept of field.

The variables with greatest impact on the probability of future presence of at least one new somatic chronic condition were found to be: age, education, gender, smoking habits and the feeling of anxiety or depression. These findings are based on the results of logistic regression, as we found the neural network analysis was probably troubled by overfitting. We also found that geographical inequalities in the risk of future illness exist. This combination makes it possible not only to establish which social and psychological determinants that have the most dominating effect on the risk of getting ill, but also where this high risk group is geographically located. Both are useful informations in planning public health care supplies, in order to minimize the health care expenditures and social inequalities in health.

Forord

En stor tak skal lyde til Region Nordjylland for at stille data til rådighed. Yderligere vil vi gerne takke vores engagerede og hjælpsomme vejleder, Rolf Lyneborg Lund, for altid at have sin dør åben og prioritere os i sin ellers travle kalender. Og til sidst et tak til Kristian Gade Kjermann og Jakob Svidt for metodisk og teknisk assistance.

Indholdsfortegnelse

1. Specialets ramme	6
2. Problemfelt	7
2.1. Indledning	7
2.2. Ulighed i sundhed	9
2.2.1. Forklaringsmodeller til ulighed i sundhed	9
2.2.2. Forskning om determinanternes betydning for sundhed	11
2.3. Kroniske sygdomme	13
2.3.1. Forskning om determinanter for somatiske kroniske sygdomme	14
2.3.2. Somatiske kroniske sygdomme i Danmark	15
2.4. Områdeulighed	16
3. Problemformulering	18
3.1. En problemformulering med et prædiktivt formål	18
3.2. Den videnskabelige placering	20
4. Specialets opbygning	24
5. Beskrivelse af data	26
5.1. Beskrivelse af de anvendte datasæt	26
5.1.1. Sundhedsprofilerne fra 2010, 2013 og 2017	26
5.1.2. Registerdata fra 2010-2016	27
5.1.3. Sammenkobling af survey- og registerdata	28
5.2. Beskrivelse af de anvendte variable	33
5.2.1. Afhængige variable: Nyopstået somatisk kronisk sygdom	34
5.2.2. Uafhængige variable	38
6. Metodeafsnit	42
6.1. Forsknings- og undersøgelsesdesign	43
6.2. Machine learning	45
6.2.1. Hvad er machine learning?	46
6.2.2. Prædiktive modeller	47
6.2.3. Supervised machine learning	47
6.2.4. Forskellige typer af algoritmer	50
6.3. Specialets undersøgelsesmetoder	51
6.3.1. Logistisk regressionsanalyse	51
6.3.2. Neural netværksanalyse	55
6.4. Diskussion af de to undersøgelsesmetoder	60

6.5. Forudsigelse af somatiske kroniske sygdomme	62
6.6. Kortlægning af somatiske kroniske sygdomme	63
6.6.1. Inddeling af områder	63
6.6.2. Fremgangsmåde for kortlægning	64
6.7. Kvalitetskriterier	65
6.7.1. Ekstern validitet	65
6.7.2. Intern validitet	66
6.7.3. Reliabilitet	68
7. Teori om determinanter	69
7.1. Socioøkonomiske faktorer og sundhedsrelateret adfærd	69
7.2. Mental trivsel og tillid	73
8. Analyse af determinanternes betydning	77
8.1. Analyse med logistisk regression	77
8.1.1. Konstruktion af model	78
8.1.2. Test af modellens præcision med datasættet fra 2013	81
8.1.3. Præsentation af den endelige model	81
8.1.4. Fortolkning ud fra odds ratio	84
8.2. Analyse med neuralt netværk	88
8.2.1. Konstruktion af modellen	88
8.2.2. Evaluering af modellen	90
8.2.3. Forudsigelse af somatiske kroniske sygdomme 2018-2020	91
8.2.4. Analyse af de prædikterede værdier med logistisk regression	91
8.2.5. Fortolkning ud fra odds ratio	94
8.3. Diskussion af analyseresultater	98
9. Teori om geografisk segregering	102
10. Analyse af geografisk fordeling af sygdom	105
11. Diskuterende konklusion	110
11.1. Sociale og psykologiske determinanternes betydning	111
11.1.1. Metodernes evne til besvarelse af problemformuleringens første del	115
11.2. Geografisk fordeling af nyopståede sygdomme	117
11.3. Afslutning	118
12. Litteraturliste	120
13. Oversigt over bilag	132

1. Specialets ramme

Dette speciale er udført i samarbejde med Region Nordjylland. Der er ønske om at kunne forudsige sygdomme og derudfra danne geografiske kort, hvor det kan ses, i hvilke områder der er særligt stor risiko for udvikling af disse. Når det vides, hvilke grupper der har særlig risiko for at blive syge og samtidig, hvilke områder der er særligt udsatte, er det i højere grad muligt at planlægge eksempelvis forebyggelse samt geografisk placering af forskellige sundhedstilbud.

2. Problemfelt

2.1. Indledning

Den overordnede sundhedstilstand blandt samfundets borgere har altid været et veldebatteret politisk emne, omend det har omhandlet forskellige fokuspunkter og sundhedsdiskurser i tidens løb. Den britiske sociolog Nikolas Rose sammenligner nutidens sundhedsdiskurser med racehygiejnen under anden verdenskrig. I dag handler den politiske interesse for sundhed dog ikke længere om befolkningens potentiale som en organisk helhed i forhold til at kunne kæmpe mod andre nationer, men i stedet handler den om økonomi og moral (Rose 2001:1ff). I en dansk kontekst kan det økonomiske aspekt betragtes således, at nogle mennesker er dyrere for samfundet end andre ved for eksempel oftere at gøre brug af offentlige sundhedsydelser. Det moralske aspekt handler om ønsket om at formindske social ulighed i sundhed (Rose 2001:5). Det er et velkendt problem, at en relativt lille gruppe af et samfunds befolkning udgør en forholdsvis stor andel af de samlede sundhedsudgifter (Shenas et al. 2014:9; Maidman & Wang 2018:1104; Lu et al. 2015:318). Også i en dansk kontekst er der en andel borgere, der bruger sundhedsvæsenets mangeartede tilbud mere end andre (Sundhedsdatastyrelsen 2018). Med udgangspunkt i det økonomiske aspekt kan det, at nogle borgere relativt set bruger mange sundhedsydelser, betragtes som et problem på samfundsniveau. Dette er ud fra en betragtning om, at det i stedet anses som fordelagtigt, at der er så få offentlige omkostninger som muligt. Ud fra det moralske perspektiv kan det anses som uretfærdigt, at nogle individer får flere sundhedsydelser for deres skattekrone, end andre gør. På den anden side kan det også moralsk anses som uretfærdigt, at der er systematiske forskelle, hvor social position hænger sammen med fordelingen af sundhedsrisici og sygdomme (Flachs et al. 2015). I forlængelse af det moralske aspekt kan det nævnes, at der i moderne velfærdssamfund er en fremherskende medicinsk- og politisk ideologi, som giver det individuelle ansvar forrang. Den offentlige debat er i høj grad præget af et individualistisk fokus (Jensen 2005:75; Cockerham 2005:51). Et eksempel herpå kan siges at være KRAM-faktorerne, der indebærer et fokus på **k**ost, **r**ygning, **a**lkohol og **m**otion (Christensen et al. 2009). Dette kan sættes i forbindelse med begrebet "blaming the victim", hvor skylden for sygdommen pålægges det syge individ (Jensen 2005:74f). Når vi fremadrettet i specialet bruger betegnelser som "god" eller "dårlig"

sundhedsrelateret adfærd, skal det forstås i relation til den sundhedsdiskurs, der præger det danske samfund, eksempelvis hvad angår Sundhedsstyrelsens anbefalinger i forhold til KRAM-faktorerne (Sundhedsstyrelsen 2018b; Sundhedsstyrelsen 2018c; Sundhedsstyrelsen & Fødevarestyrelsen 2018; Sundhedsstyrelsen 2019). Vi betragter, at dette er forbundet med et ulighedsperspektiv, eftersom vi ser, at det ikke er alle personer, der formår at leve op til samfundets normer og forventninger til passende sundhedsrelateret adfærd, og derfor finder vi det særligt vigtigt i dette speciale at fokusere på strukturerne, der påvirker individets "valg" og handlemuligheder.

Vi finder, at både den økonomiske og den moralske problematik afspejles i den internationale forskning på området, hvor der samtidig er fokus på at lave statistiske modeller til at forudsige, hvilke befolkningsgrupper der bliver syge i fremtiden (Bla. Perkins et al. 2004; Fleishman & Cohen 2010; Orueta et al. 2013; Charlson et al. 2014; Shenan et al. 2014; Lu et al. 2015; Tamang et al. 2017; Maidman & Wang 2018). Det er ud fra et ønske om at effektivisere forebyggelsesindsatser i sundhedssystemerne i lande verden over – både for at spare flest mulige omkostninger men også for at forbedre levevilkårene for det enkelte menneske (ibid.). I den internationale forskning bruges betegnelsen "high-cost users" eller "high-cost patients" om den forholdsvis lille gruppe af mennesker, der står for en relativt stor andel af et lands samlede sundhedsudgifter (ibid.). High-cost gruppen defineres oftest som de fem eller ti procent dyreste personer i den samlede population (Fleishman & Cohen 2010:534; Orueta et al. 2013:4; Charlson et al. 2014:4; Shenan et al. 2014:10; Lu et al. 2015:318; Tamang et al. 2017:2). Ifølge Sundhedsdatastyrelsen skyldes 71 procent af det samlede sundhedsbudget på 84 milliarder kroner i 2017 de ti procent dyreste patienter. De ti procent dyreste patienter har i gennemsnit modtaget behandling for 112.400 kroner om året per patient, hvilket er 22 gange mere end de gennemsnitlige udgifter for de resterende 90 procent af befolkningen (Sundhedsdatastyrelsen 2018:2). For at forklare hvorfor der er denne ulighed i sundhed, vil det i det efterfølgende afsnit inddrages, hvad der har betydning for, at der opstår ulighed i sundhed ud fra forskellige forklaringsmodeller og empiriske observationer.

2.2. Ulighed i sundhed

I første del af dette afsnit vil der tages udgangspunkt i forskellige typer af forklaringer på, hvorfor der findes ulighed i sundhed inden for det sygdoms- og sundhedssociologiske domæne. Disse forklaringer trækker både på sociale, psykologiske og fysiologiske aspekter, og de rummer således et bredt udvalg af faktorer. Dette er blandt andet socioøkonomisk status, livsstil, boligforhold, arbejdsforhold, social samhørighed, selvværd og social kapital, og modellerne belyser dermed vigtigheden af at nuancere det tidligere nævnte individualistiske fokus. I dette speciale vil forklaringsmodellerne anskues som supplerende frem for konkurrerende, og dermed kan de på forskellig vis være med til at forklare, hvorfor der er ulighed i sundhed. Den anden del af afsnittet vil med udgangspunkt i eksisterende forskning have fokus på konkrete determinanternes betydning for sundhed.

2.2.1. Forklaringsmodeller til ulighed i sundhed

Inden for de forskellige forklaringer på hvorfor der er ulighed i sundhed, er der særligt fundet to skel, som vi vil læne os op ad i en kort gennemgang af de forskellige forklaringsmodeller. For det første er der forskellige perspektiver på, hvad der har en indvirkning på hvad. Eksempelvis kan det nævnes, at det ud fra den *kulturelle- og adfærdsrelaterede forklaringsmodel* betragtes, at den sociale position er af betydning for individets livsstil og adfærd, og dette er afgørende for individets sundhedstilstand. Uligheden opstår altså på baggrund af menneskers forskellige adfærdsmønstre og opfattelser af, hvad der betegnes som henholdsvis sundt eller usundt. I modsætning hertil antages det i *selektionsmodellen*, at det er individets sundhedsstatus, der er af betydning for den sociale position, og uligheden består således i, at individer med en bedre sundhed oftere oplever opadgående social mobilitet (Nettleton 2013:161f). Hvordan vi forholder os til retningen af denne sammenhæng vil indgå i afsnit "6.7.2. Intern validitet".

For det andet findes en forskellighed i forhold til, hvorvidt fænomenet betragtes nomotetisk eller idiografisk, hvilket refererer til forskelle i antagelsen om, hvorvidt der er interesse i subjektive faktorer, eller faktorer, hvor mennesker påvirkes ens (Møhl & La Cour 2008:135f). En forklaring med et nomotetisk perspektiv er den *materialistiske forklaringsmodel*, hvor det anskues, at eksterne faktorer såsom fattigdom og forurening

rammer individer, der udsættes herfor, ens. Her kan ulighed eksempelvis forklares med, at mennesker bor i forskellige geografiske områder, hvor de er udsat for forskellige risikofaktorer (Nettleton 2013:162). På den anden side findes den *psykosociale forklaringsmodel* med et idiografisk udgangspunkt, hvor det betragtes, individer udsættes for eksterne stressorer, der kan beskrives således:

"(...) any kind of situation which is perceived as challenging or demanding: danger, a university test, the duties and tasks involved in one's job, unemployment, the loss of a relative." (Elstad 2000:72)

Situationens indvirkning på individet betragtes som socialt konstrueret og dermed forskellig for forskellige individer (Elstad 2000:73). Disse stressorer udgør en ulempe i forhold til individets sundhedsstatus, der enten kan give direkte udslag i individets sundhed som biologiske reaktioner såsom forhøjet blodtryk eller dårligt immunforsvar men også indirekte gennem livsstil (Nettleton 2013:164). Den psykosociale forklaringsmodel er opstået som et resultat af, at den materialistiske forklaring ikke længere har samme relevans i velstillede samfund, hvor borgere generelt har gode boligforhold, gode madvarer og jobs med relativt lav fysisk belastning. Derfor må uligheden forklares med andre determinanter. Det skal forstås på den måde, at når et vist niveau af levestandard er opnået, erstatter psykosociale faktorer de biokemiske og fysiske faktorer som hovedårsag til forskelle i sundhed (Elstad 2000:70).

Livsforløbet som forklaringsmodel kombinerer elementer fra de øvrige modeller. Her antages det, at der igennem et menneskeliv ophober sig en akkumulation af positive og negative eksponeringer forstået som summen af samtlige betydelige faktorer, et individ bliver eksponeret for i løbet af sit liv (Nettleton 2013:165). Slutteligt findes også *artefakt forklaringen*, hvor ulighed i sundhed forklares ud fra, at både klasse og sundhed anskues som artefakter af de sociale processer, der har fundet sted i forbindelse med udarbejdelse af målene, og derfor er også sammenhængen mellem de to artefakter (Nettleton 2013:161).

Det efterfølgende afsnit vil indebære forskellige sociale og psykologiske determinanternes betydning for sundhed med udgangspunkt i eksisterende forskning, og disse vil betragtes i forhold til de forskellige forklaringsmodeller.

2.2.2. Forskning om determinanternes betydning for sundhed

I 2005 etablerede the World Health Organization (WHO) "the Commission on Social Determinants of Health", hvis primære formål er at reducere den globale ulighed i sundhed ved at fokusere på de sociale determinanter for sundhed. Disse determinanter er ifølge WHO defineret ved: "*the conditions in which people are born, grow, work, live, and age, and the set of forces and systems shaping the conditions of daily life*" (WHO 2019). Michael Marmot, medlem af kommissionen, har skrevet en artikel om kommissionens arbejde i forbindelse med dens oprettelse (Marmot 2005). Heri beskriver han, hvordan der på globalt plan både findes en stor ulighed i sundhed mellem lande men også internt i lande – eksempelvis er der i USA 20 års forskel på forventet levealder blandt den højeste og laveste socialklasse. I artiklen lægges der vægt på skellet mellem infektionssygdomme og ikke-smitsomme sygdomme, hvor det betragtes, at ikke-smitsomme sygdomme (livsstilssygdomme) kræver en nuanceret forståelse af de sociale determinanter, eftersom de ikke udelukkende kan forklares med den materialistiske forklaring, som eksempelvis hygiejneforhold, sult eller mangel på adgang til medicinsk behandling (Marmot 2005:1102).

Analyser af sundhed på globalt plan har fastslået vigtigheden af risikofaktorer relateret til individets sundhedsrelaterede adfærd som eksempelvis KRAM-faktorerne. Det er alment anerkendt, at disse har indflydelse på individets overordnede sundhed (Marmot 2005:1102). Ud fra den tidligere inddragelse af de forskellige forklaringer på ulighed i sundhed kan det siges, at der her særligt er fokus på den kulturelle- og adfærdsrelaterede forklaringsmodel. I artiklen af Marmot (2005) henvises dernæst til en række studier, der overordnet har påvist, at dårlig sundhed er socialt graderet. Således er individets overordnede sundhedsstatus dårligere, jo lavere social position han eller hun har (Marmot 1999; Mackenbach & Howden-Chapman 2003; Galobardes et al. 2006; Braveman et al. 2010). Derudover beskrives det, hvordan både uddannelse, arbejde og indkomst har en effekt på sundhed. Jo højere uddannelse, individet har, des større viden

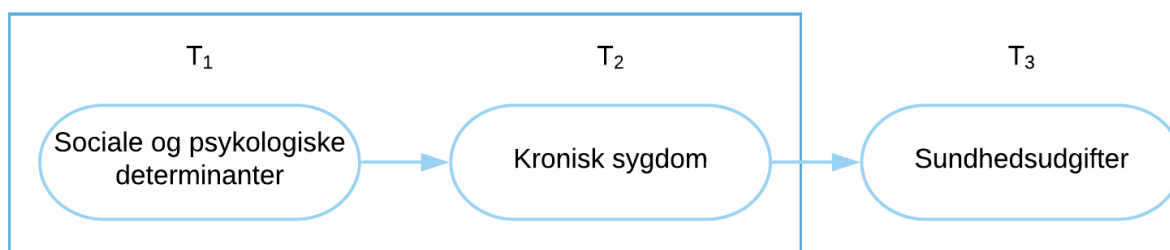
har det om sundhed, og deraf følger bedre forudsætninger for at træffe gode sundhedsrelaterede beslutninger (Barbeau et al. 2004; Dewalt et al. 2004; Sanders et al. 2009). Uddannelse påvirker desuden sundhed gennem sociale og psykologiske faktorer, idet en høj uddannelse medfører blandt andet en højere grad af social støtte, som har en positiv effekt på individets sundhed (Leganger & Kraft 2003; Mickelson & Kubzansky 2003; Uchino 2006). Yderligere former uddannelse mulighederne for individets arbejdsliv, hvor det betragtes, at de fysiske arbejdsbetingelser samt de psykosociale aspekter ved arbejdet såsom social støtte eller mængden af overarbejdstid ligeledes har en effekt på sundhed (Marmot et al. 1997; Stansfeld et al. 1997; Stansfeld et al. 1998a; Stansfeld et al. 1998b; Stansfeld et al. 1999; de Jonge et al. 2000). Yderligere viser andre studier, at økonomisk velstand og indkomst har en signifikant positiv effekt på individets sundhed efter kontrol for uddannelsesniveau (Daly et al. 2002; Braveman et al. 2005; Herd et al. 2007; Avendano & Glymour 2008; Larson & Halfon 2009; Kawachi et al. 2010).

Udover ovenstående sociale og psykologiske determinanternes betydning for sundhed fremkommer det ligeledes i eksisterende forskning, at etnicitet har betydning, hvilket primært skyldes racisme. Racisme skal i denne forstand ikke forstås som intentionelle, diskriminerende handlinger men som nogle dybereliggende sociale strukturer, der systematisk undertrykker individers muligheder på baggrund af deres etniske tilhørsforhold. Personer af anden etnisk herkomst går eksempelvis oftere på dårlige skoler, og får dermed et dårligere arbejde og lavere indkomst. Effekten mellem etnicitet og sundhed er altså medieret af de øvrige sociale determinanter som uddannelse, beskæftigelse og indkomst (Braveman et al. 2011:387f). Derudover kan sundheden også påvirkes gennem stress forårsaget af, at en person er udsat for racistiske oplevelser af forskellig art, hvilket eksempelvis kan være bemærkninger eller handlinger fra andre (ibid.; Williams & Mohammed 2009). Netop stress har en afgørende betydning for ulighed i sundhed, eftersom det er relateret til den kausale forbindelse mellem sundhed og samtlige af de øvrige determinanter, der er nævnt, forstået på den måde, at individer i lavere sociale positioner generelt oftere er udsat for stressreaktioner. Empirisk er det bevist, at stressreaktioner kan skabe fysiske reaktioner i både hjernen, immunforsvaret, hjertekarsystemet og de vitale organer (Steptoe & Marmot 2002; Evans & Kim 2007; McEwen & Gianaros 2010). Her inddrages altså den psykosociale forklaringsmodel.

2.3. Kroniske sygdomme

Blandt mange tidligere studier er der enighed om, at forekomsten af multiple kroniske sygdomme er en af de bedste indikatorer til at måle, hvorvidt et individ enten allerede befinder sig i denne omkostningsfulde gruppe, eller ender i den på et senere tidspunkt (Fleishman & Cohen 2010; Orueta et al. 2013; Lu et al. 2015; Yurkovich et al. 2015; McPhail 2016). I dette speciale har vi ladet os inspirere af den tidligere forskningsprædiktive metoder, og derudfra vil vi undersøge, hvem der i fremtiden får en kronisk sygdom. Dermed vil der i dette speciale være fokus på kroniske sygdomme som den afhængige variabel, og sundhedsmkostninger vil således ikke blive studeret direkte. I stedet undersøges kronisk sygdom som en *indikator* for, hvem der i fremtiden relativt set bliver dyre for det danske sundhedsvæsen, og der vil være fokus på de sociale og psykologiske determinanter, der har betydning for, at en person udvikler en sådan sygdom. Derudfra kan der betragtes en kausalkæde som nedenstående, hvor den blå firkant viser det led, der vil undersøges i dette speciale.

Figur 1: Kausalkæde



I økonomisk veludviklede lande er kroniske sygdomme steget markant de seneste år (Shenas et al. 2014:9; McPhail 2016; Kyed et al. 2018). I dag udgør kroniske sygdomme langt størstedelen af det samlede sygdomsbillede i disse lande, hvilket først og fremmest skyldes en udvikling i den medicinske teknologi, der markant har sænket forekomsten af akutte og livstruende sygdomme som eksempelvis infektioner (McPhail 2016; Kyed et al. 2018:160). At folk i dag bliver ældre og ældre er desuden med til at øge stigningen i antallet af folk med kroniske sygdomme, eftersom sandsynligheden for at have én eller flere kroniske sygdomme er større, jo ældre en person er (Huntley et al. 2012:13; Orueta et al. 2013:1; McPhail 2016:144; Pathirana 2018:186). Dertil kommer, at den forbedrede medicinske teknologi også har bidraget til udviklingen af bedre behandlingsmuligheder

til kronisk syge patienter, som gør, at også denne gruppe lever længere (Pathirana et al. 2018:186). I Danmark anslås det, at cirka en tredjedel af befolkningen lever med én eller flere kroniske sygdomme (Sundhedsstyrelsen 2017).

I dette speciale undersøges kronisk sygdom med udgangspunkt i de sygdomsgrupperinger, der findes i Region Nordjyllands patientadministrative system. Her indgår patienter, der har modtaget behandling for en kronisk sygdom i det pågældende år. Der indgår både somatiske og psykiatriske kroniske sygdomme, hvor vi har valgt at afgrænse os til somatiske kroniske sygdomme. Dette er valgt, fordi der er et ønske om at se på betydningen af forskellige former for psykisk trivsel for kroniske sygdomme, hvormed de forklarende faktorer i høj grad ville være overensstemmende med de undersøgte sygdomme. Derudover fremgår det også af rapporten for den nordjyske sundhedsprofil fra år 2017, at personer med psykiatriske og somatiske kroniske sygdomme adskiller sig væsentligt fra hinanden, hvad angår socioøkonomiske faktorer som eksempelvis køn og alder, men også i forhold til både social- og mental trivsel (Kyed et al. 2018:163).

2.3.1. Forskning om determinanter for somatiske kroniske sygdomme

I dette afsnit vil indgå, hvilke sociale og psykologiske determinanter på individniveau der har betydning for, at et individ udvikler en somatisk kronisk sygdom. Overordnet kan det siges, at risikoen for forekomsten af somatiske kroniske sygdomme i høj grad afhænger af individets socioøkonomiske forhold såsom køn, uddannelse, og indkomst og alder (Huntley et al. 2012:13; Schafer et al. 2012; Orueta et al. 2013:1; Flachs et al. 2015; McPhail 2016:144; Pathirana 2018:186; Kyed et al. 2018). Ligesom det er tilfældet for sundhed generelt set, har KRAM-faktorerne også en effekt på udviklingen af somatiske kroniske sygdomme (Skivington et al. 2015; Carter et al. 2015; Cockerham et al. 2017). Rygning er den faktor, der har størst betydning, og der er påvist en effekt af rygning i forhold til mere end 21 forskellige somatiske kroniske sygdomme (Cockerham et al. 2017). Derudover har også mental trivsel hos individet en betydning, idet det er påvist, at forekomsten af angst og depression medfører en øget risiko for forekomsten af en række somatiske kroniske sygdomme (Lando et al. 2006; Scott et al. 2007). I forbindelse med angst og depression kan det nævnes, at stress ligeledes har betydning, da personer

med et højt stressniveau har større risiko for at udvikle somatiske kroniske sygdomme (Stansfeld et al. 2002; Chandola et al. 2006; Fortin et al. 2006; Nielsen et al. 2008; Prior et al. 2016). Slutteligt har individets sociale netværk betydning for dets sundhedsrelaterede adfærd, og har derfor også en indirekte effekt på forekomsten af somatiske kroniske sygdomme (Almedom 2005; Vassilev et al. 2010; Pentecost & Taket 2011). Der er imidlertid også en ulighed knyttet til potentialet af individets sociale netværk, eftersom nogle netværk er mere gunstige end andre, hvilket ofte også vil hænge sammen med individets sociale position (Umberson & Montez 2010; Cockerham et al. 2017).

2.3.2. Somatiske kroniske sygdomme i Danmark

Af de inkluderede sygdomme i specialets mål for somatisk kronisk sygdom er gigtsygdomme den mest udbredte i Danmark. 820.000 personer lever hver dag med slidgigt, som er den mest hyppige form for gigt. Personer med slidgigt har årligt 2,7 millioner flere besøg hos alment praktiserende læge end personer uden samme diagnose, og sygdommen koster hvert år 2.440 millioner kroner i behandling og pleje (Flachs et al. 2015:355). Fælles for alle somatiske kroniske sygdommene er, at forekomsten på landsplan er større blandt personer, der har en kort eller ingen uddannelse sammenlignet med personer, der har en mellemlang eller lang uddannelse. Både diabetes og hjertekarsygdomme forekommer hyppigst blandt mænd, mens gigtsygdomme og knogleskørhed optræder hyppigere blandt kvinder. For samtlige sygdomme gælder det, at forekomsten stiger markant, jo ældre individet bliver (Flachs et al. 2015). Dette er dog med undtagelse af astma, der forekommer hyppigst hos børn, omend denne sygdom også optræder hyppigt hos voksne (Flachs et al. 2015:299).

I Region Nordjylland er muskelskeletsygdomme, som er en samlet betegnelse for diverse gigtsygdomme, knogleskørhed og rygsygdomme, den mest udbredte (Kyed et al. 2018:160). Fælles for både denne, diabetes, KOL og hjertekarsygdomme er desuden, at der også i Nordjylland forekommer en ulighed i forekomsten af disse, baseret på uddannelsesniveau, idet der er en hyppigere forekomst blandt personer med erhvervsfaglig- eller grundskoleuddannelse (Kyed et al. 2018:163). Blandt nordjyske borgere har mænd en højere risiko for både hjertekarsygdomme, KOL og diabetes, mens

kvinder har højere risiko for muskelskeletsygdomme (ibid.). Derudover ses det, at forekomsten af somatiske kroniske sygdomme er langt større blandt den ældre del af den nordjyske befolkning, end det er tilfældet for de unge (Kyed et al. 2018:161). Der er altså både nationalt og regionalt en social ulighed i forekomsten af disse sygdomme, både hvad angår køn, alder og uddannelsesniveau.

2.4. Områdeulighed

Udover at der i eksisterende forskning findes evidens for, at der findes ulighed angående sundhed på individniveau, findes det ligeledes, at der er ulighed i sundhed på områdeniveau (Pickett & Pearl 2001). Med udgangspunkt i eksisterende forskning antager vi, at der kan findes to forklaringer herpå. For det første kan det skyldes, at områdeeffekter har en betydning for individers sundhed og derigennem somatiske kroniske sygdomme. Områdekarakteristikas betydning for sundhed opsummeres som nedenstående i reviewartiklen af Kate E. Pickett og Michelle Pearl (2001):

“The neighbourhoods in which people live may influence health, operating through such mechanisms as: the availability and accessibility of health services; infrastructure deprivation (lack of parks, stores selling healthy foods at affordable prices, etc); the prevalence of prevailing attitudes towards health and health related behaviours; and stress and a lack of social support.” (Pickett & Pearl 2001:111)

Dette kaldes områdeeffekter (Pickett & Pearl 2001:111). I denne artikel og samtidig også reviewartiklen af J. Michael Oakes et al. (2015) fremgår det, at karakteristika på områdeniveau har indflydelse på en lang række sundhedsrelaterede outcomes. Eksempelvis har områdefattigdom vist sig at have en effekt på somatiske kroniske eller langvarige sygdomme (Diez-Roux et al. 1997; Davey Smith et al. 1998; Robert 1998; Reijneveld 1998).

For det andet er områder ikke blot tilfældige konstellationer, eftersom det ikke er tilfældigt hvor mennesker bosætter sig (Sampson 2008; Sampson & Sharkey 2008; South & Crowder 1997; Lund 2019). Angående bosætningsmønstre har det tidligere SFI fremlagt, at det danske boligmarked typisk er opdelt således, at individer med samme

uddannelses- og indkomstmæssige baggrunde ofte bosætter sig i samme områder (Christensen et al. 2015:18). Dermed kan det siges, at når individer bosætter sig med nogle, der ligner dem selv, har den tidligere nævnte sociale ulighed på individniveau betydning for, at der fremkommer geografisk ulighed på områdeniveau. Vi betragter de to forskellige forklaringer på geografisk ulighed i sundhed som supplerende.

3. Problemformulering

Ud fra ovenstående gennemgang af teoretiske perspektiver og eksisterende forskning om sociale og psykologiske determinanter og somatiske kroniske sygdomme, understreges relevansen af at undersøge somatiske kroniske sygdomme i forhold til at belyse både økonomisk og social ulighed i sundhed. Dette gøres som nævnt ud fra et fokus på at kunne forudsige forekomsten af somatisk kronisk sygdom, og derudfra har vi opstillet følgende problemformulering:

Hvilke sociale og psykologiske determinanter på individniveau har størst betydning for sandsynligheden for at udvikle en somatisk kronisk sygdom i fremtiden, og hvordan vil forekomsten af disse fordele sig geografisk i Region Nordjylland?

De sociale og psykologiske determinanter består i denne undersøgelse af 31 forklarende variable, der kan inddeles i seks overordnede temaer: 1) socioøkonomiske faktorer, 2) mental trivsel, 3) socialt samvær og støtte, 4) sundhedsrelateret adfærd, 5) social samhørighed, tillid og tryghed i lokalområdet samt 6) generel tillid til medmennesker og samfundsmæssige instanser. Områdeulighed vil studeres ved hjælp af en kortlægning af forekomsten af både eksisterende og fremtidig somatisk kronisk sygdom i Region Nordjylland. En nærmere redegørelse for de anvendte variable, undersøgelsesdesign samt anvendte metoder til at besvare specialets problemformulering vil fremgå af de kommende afsnit: "5. Beskrivelse af variable" samt "6. Metodeafsnit".

3.1. En problemformulering med et prædiktivt formål

Når det i eksisterende forskning forsøges at forudsige, hvem der bliver særligt dyre for sundhedsvæsenet, anvendes ofte machine learning i form af eksempelvis neurale netværk og random forrest modeller (Shenas et al. 2014; Bertsimas et al. 2008; Moturu et al. 2013; Maidman & Wang 2018). Selvom det ikke er sundhedsudgifter men somatiske kroniske sygdomme, der er et ønske om at forudsige i dette speciale, vil denne metodologiske tilgang inddrages grundet fokus på forudsigelse.

Selvom machine learning er et forholdsvis nyt begreb inden for sociologien, kan det bruges til at løse mange af de problemer, som sociologien altid har beskæftiget sig med (Molina & Garip 2019:12). Faktisk findes et overlap mellem machine learning og de statistiske metoder, der traditionelt anvendes inden for det kvantitative sociologiske felt, og det er en begrebsverden, der er fundet svær konceptuelt helt at adskille fra hinanden. Nogle refererer til de klassiske statistiske metoder som regressions- eller clusteranalyse som eksempler på machine learning teknikker (eks. Domingos 2012; Murphy 2013), mens der modsat ofte også refereres til statistik og machine learning som to forskellige ting (eks. Nielsen 2010; Molina & Garip 2019).

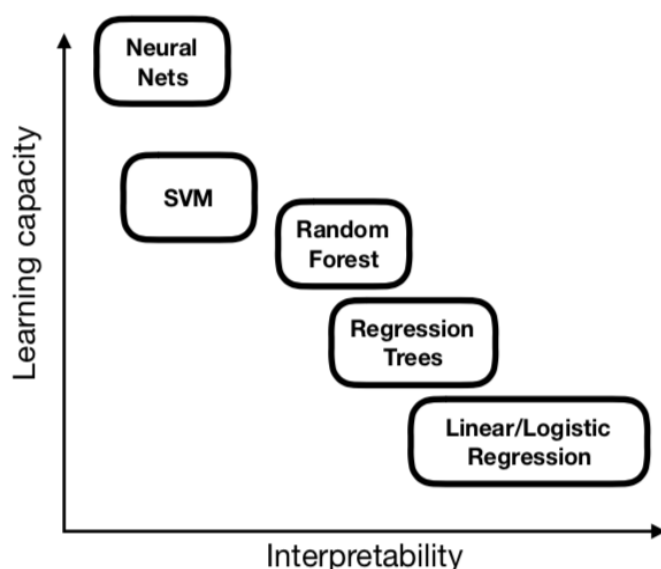
For at komme nærmere en afklaring omkring, hvordan machine learning og de klassiske statistiske metoder relaterer sig til hinanden, og hvordan machine learning finder sin anvendelse i dette speciale, findes det relevant at inddrage distinktionen mellem *forudsigelse* og *inferens* (Donoho 2017:751). De traditionelle statistiske metoder, som vi anvender inden for det sociologiske felt, er koncentreret omkring den statistiske inferens, altså undersøgelser af hvordan et kendt outcome (Y) er relateret til en række inputs (X). Denne type af modeller er ofte simple at fortolke, men de er udfordret så snart, det ønskes at generalisere til et andet sample. Machine learning refererer derimod til en prædiktiv model, hvor formålet er forudsigelse af et ukendt (fremtidigt) outcome (Y) ud fra inputs (X), og her udgør evnen til at kunne generalisere til andre (fremtidige) populationer altså essensen (Molina & Garip 2019:2). Mange af de matematiske eller statistiske teknikker der traditionelt anvendes med fokus på inferens indenfor den kvantitative sociologi, kan også anvendes som prædiktive modeller. Eksempelvis har flere af de tidligere nævnte studier anvendt logistiske regressionsmodeller til at forudsige, hvem der bliver high-cost patient (Charlson et al. 2014; Orueta et al. 2013; Lu et al. 2015; Tamang et al. 2017).

Den primære årsag til, at machine learning (endnu) ikke er så udbredt inden for socialvidenskabelige studier, er, at undersøgelsen af sociale faktorer netop har kredset omkring et ønske om at forklare de kausale forbindelser (Seligman et al. 2017:95). Det traditionelle fokus på statistisk inferens bliver i disse år nytænkt efter fremkomsten af nye måder at arbejde med store datamængder på. Brugen af kunstig intelligens og machine learning må siges at vinde frem i disse år, hvor også samfundsvidenskaberne så

småt begynder at få øjnene op for de nye muligheder (eks. Hain & Jurowetski 2018; Molina & Garip 2019).

I dette speciale vil anvendes et neuralt netværk og en logistisk regression. Metoderne er udvalgt med udgangspunkt i nedenstående figur for på den måde at sammenligne udfaldet ved anvendelse af en simpel, let fortolkelig metode med udfaldet ved anvendelse en kompleks metode med stor læringskapacitet.

Figur 2: Læringskapacitet og fortolkningspotentiale i forskellige machine learning teknikker



Kilde: Hain & Jurowetski (2018:3)

3.2. Den videnskabelige placering

I dette afsnit vil vi redegøre for vores videnskabelige ståsted.

Den biomedicinske model, der er dominerende inden for medicinsk forskning i den vestlige verden, er kendetegnet ved, at forklaringer på sygdom udelukkende ansues ud fra biologiske forandringer, hvilket hænger sammen med en dualistisk opfattelse af forholdet mellem krop og sind. De sociale og psykologiske mekanismer er altså ikke af interesse for studier funderet i det biomedicinske paradigme (Nettleton 2013:2). Det

betragtes, at årsagsbegrebet i den biomedicinske model lægger sig op ad den positivistiske videnskabsteoretiske retning, hvor det ikke forklares, hvorfor et fænomen efterfølger et andet (Wad 2015:379). I stedet er der interesse for at fastslå, hvorvidt én hændelse medfører en anden (ibid.), hvilket eksempelvis kan være, hvorvidt rygning medfører lungecancer. Inden for det sundheds- og sygdomssociologiske domæne tages der imidlertid afstand til både det biomedicinske paradigme og det positivistiske årsagsbegreb, hvilket vi ligeledes gør i dette speciale. I stedet antager vi en mere holistisk tilgang til sundhed, hvor krop og sind betragtes som gensidigt påvirkende, og sociale og psykologiske mekanismer anses for at have betydning for individets sundhed. Dette findes der som nævnt belæg for i eksisterende forskning, hvor det betragtes, at eksempelvis oplevelser kan give stressreaktioner, der kan medføre fysiske forandringer (jf. problemfeltet).

Dertil er vi inspirerede af årsagsbegrebet indenfor kritisk realisme, hvor der arbejdes med et *generativt kausalitetsbegreb* med henblik på at kunne forklare de sociale mekanismer, der udgør årsagen til, at én begivenhed fører til den næste (Wad 2015:379). I dette kausalitetsbegreb forkastes idéen om empiriske og faktuelle regelmæssigheder, da det aldrig er sikkert, at en begivenhed fører til den næste, blot fordi det er sket tidligere. Det skyldes en antagelse om, at objekter (for eksempel individer) besidder kræfter til at handle, men at disse kræfter afhænger af den givne sociale kontekst. En sådan kausalitetsforståelse fordrer dermed et behov for teorier, der kan tilbyde en forklaring på kausale sammenhænge, der studeres (ibid.). I dette speciale spiller sundheds- og sygdomssociologiske teorier derfor en væsentlig rolle i forhold til forklaringen af de empiriske fund.

Inden for sundheds- og sygdomssociologien er kritikken af den biomedicinske model ofte funderet i et socialkonstruktivistisk syn på sygdom og sundhed, som vi også antager i dette speciale. Det kommer til udtryk ved, at forståelsen af sygdomme og kropslige dysfunktioner anskues som noget, der er socialt defineret frem for, at de anskues som objektive, videnskabelige sandheder (Bury 2013:113). Det betyder ikke, at det inden for socialkonstruktivismen ikke anerkendes, at der sker biologiske forandringer i kroppen, men at indfangelsen og erkendelsen af disse sker gennem sociale konstruktioner og dermed er influeret af den sociale kontekst (Nettleton 2013:13ff). Det vil dermed også

sige, at hvad der eksempelvis betegnes som henholdsvis “sundt” eller “usundt” i dette speciale ikke betragtes som objektive sandheder men derimod som begreber, der er konstruerede gennem sociale processer, og de er således et produkt af den sociale kontekst, de indgår i.

Med udgangspunkt i kritisk realisme anskuer vi en ontologisk realisme, hvilket indebærer, at virkeligheden ikke er afhængig af vores bevidsthed men eksisterer uafhængigt heraf (Wad 2015:370). Det anskues, at der er tre virkelighedsdomæner. Dette er det empiriske, subjektive domæne, det aktuelle domæne og det reale domæne. Førstnævnte domæne består af erfaringer og observationer, og det kan siges, at det er dette domæne, den biomedicinske model beskæftiger sig med. Det aktuelle domæne udgøres af både erfarede og ikke-erfarede begivenheder og fænomener (Wad 2015:372). Det reale domæne indebærer det samme som de andre to domæner, og det indebærer yderligere dybereliggende kræfter og mekanismer, der er afgørende for de faktisk eksisterende forhold (ibid.). Med denne virkelighedsforståelse betyder det, at når vi undersøger, hvilke sociale og psykologiske determinanter der har størst betydning for sandsynligheden for at udvikle en somatisk kronisk sygdom i fremtiden, og hvordan forekomsten af disse vil fordele sig geografisk, har vi øje for både erfaringer, faktisk eksisterende forhold og samtidig de dybereliggende kræfter og mekanismer. Det subjektive, empiriske domæne kan ses i de erfaringer og observationer, der findes i både survey- og registerdataene anvendt i dette speciale. Det aktuelle domæne kan i denne sammenhæng henvise til de faktuelle biologiske forandringer, hvilket vil sige, at det udover observationer og erfaringer også kan indebære biologiske forandringer, der forstås som somatisk kronisk sygdom, selvom det hverken fremgår af registret eller i folks egen bevidsthed. Det reale domæner omfatter ligeledes dette men yderligere også eksempelvis sociale strukturer, der har betydning for de biologiske forandringer. Det er i undersøgelsen af det reale domæne, at de teoretiske perspektiver spiller en vigtig rolle.

Når vi i specialet undersøger sociale og psykologiske determinanters betydning for sandsynligheden for at udvikle en somatisk kronisk sygdom, findes det relevant at forklare dette speciales tilgang til aktør-struktur-problematikken. Her tager vi udgangspunkt i den kritiske realist Roy Bhakars perspektiv. Ud fra hans perspektiv er social struktur og handling ontologisk distinkte niveauer, der gensidigt påvirker

hinanden. Dog eksisterer samfundet forud for det individuelle menneske, og derfor tillægges samfundet den vigtigste rolle (Wad 2015:376f). Det betyder, at vi anskuer, at individet til en vis grad er underlagt strukturer, hvilket sker gennem socialisering, mens individet har en indvirkning på strukturerne gennem reproduktion og transformation (Wad 2015:376).

4. Specialets opbygning

I dette afsnit præsenteres specialets opbygning for at give et overblik over de enkelte delelementer, der indgår.

Afsnit "5. Beskrivelse af data" indeholder en beskrivelse af de anvendte datasæt, hvilke er henholdsvis spørgeskemadata fra Sundhedsprofilerne samt registerdataene fra Region Nordjyllands patientadministrative system. Derudover findes en redegørelse for sammenkoblingen af de to datasæt og en redegørelse for de variable, der anvendes i dette speciale. Efterfølgende vil der i afsnit "6. Metodeafsnit" redegøres for specialets forsknings- og undersøgelsesdesign samt de metoder, der anvendes til analyse, hvilket er henholdsvis logistisk regression og neural netværk. I forlængelse heraf bliver fordele og ulemper ved anvendelsen af disse to analysemetoder desuden diskuteret. Metodeafsnittet afsluttes med en redegørelse for den fremgangsmåde, der benyttes i specialets sidste del, hvor geografisk ulighed studeres ved hjælp af kortlægning af sygdom.

Derefter vil være afsnit "7. Teori om determinanter", som indeholder de teoretiske perspektiver, der anvendes til besvarelse af første del af problemformuleringen. I næste afsnit, "8. Analyse af determinanternes betydning", vil der trækkes på perspektiver fra det foregående afsnit. Udover teoretiske fortolkninger af analyseresultater indeholder afsnittet også en beskrivelse af fremgangsmåden for anvendelsen af henholdsvis logistisk regression og neural netværk.

De teoretiske perspektiver til anden del af specialets problemformulering redegøres for i afsnit "9. Teori om geografisk segregering", og samtidig vil disse anskuelser anvendes til at belyse, hvorfor det er relevant at inddele Region Nordjylland i mindre områder og undersøge geografisk ulighed i sundhed. Dernæst følger afsnit "10. Analyse af geografisk fordeling af sygdom", der indeholder deskriptive kortlægninger af den geografiske ulighed i forekomsten af somatiske kroniske sygdomme, som efterfølgende suppleres med en teoretisk fortolkning af den geografiske ulighed, der findes i dette speciale.

I afsnit “11. Diskuterende konklusion” vil problemformuleringen besvares, og afsnittet vil afsluttes med idéer til, hvordan fremtidige undersøgelser inden for området med fordel kan opbygges.

5. Beskrivelse af data

I dette afsnit beskrives det anvendte data. Dette vil først ske med en redegørelse for de anvendte datasæt, og hvordan de er sammenkoblet. Derefter vil de inddragede afhængige og uafhængige variable blive gennemgået, og det vil ligeledes indgå, hvordan disse er bearbejdet.

5.1. Beskrivelse af de anvendte datasæt

I dette speciale anvender vi data fra spørgeskemaundersøgelsen "Sundhedsprofilen", som består af fire tværsnitsundersøgelser af repræsentative tilfældige udvalgte samples i Region Nordjylland i årene 2007, 2010, 2013 og 2017. I dette speciale har vi adgang til årene 2010, 2013 og 2017. Desuden anvender vi registerdata fra Region Nordjyllands patientadministrative system, der indeholder samtlige registreringer af en række grupperinger af kroniske sygdomme fra 2010-2016, samt registerdata over indkomst, køn, alder og etnicitet i årene 2009-2016. Disse data er koblet til spørgeskemadataene. Variablene i datasættene fra Sundhedsprofilerne vil sammen med indkomstdata anvendes som uafhængige variable, mens registerdata over somatiske kroniske sygdomme vil anvendes som de afhængige variable. I de tilgængelige variable indgår også bopælskoordinater for den enkelte respondent, nærmere betegnet koordinater for længde- og breddegrader. Disse koordinater vil anvendes i besvarelsen af problemformuleringens sidste halvdel, hvor det undersøges, hvordan forekomsten af somatiske kroniske sygdomme vil fordele sig geografisk i Region Nordjylland.

5.1.1. Sundhedsprofilerne fra 2010, 2013 og 2017

Sundhedsprofilerne i Region Nordjylland er en del af en landsdækkende undersøgelse gennemført i samarbejde med Sundhedsstyrelsen, Statens Institut for Folkesundhed samt landets øvrige regioner. Det overordnede formål med undersøgelsen beskrives som værende:

"(...) at kortlægge trivsel, sundhed og sygdom hos borgerne i Region Nordjylland, der kan fungere som et planlægningsredskab, når der skal prioriteres og tilrettelægges forskellige

sundheds- og forebyggelsestilbud i kommuner og region. Sundhedsprofilerne danner også grundlag, når de kommunale sundhedspolitikker udformes.” (Region Nordjylland 2019)

Undersøgelsernes resultater bruges således til at planlægge sundheds- og forebyggelsestilbud i regionen ud fra de tendenser og behov, som undersøgelsen viser, og de udgør således en vigtig del af grundlaget for udviklingen af de regionale sundhedspolitikker.

I 2017 indeholdte spørgeskemaet 77 spørgsmål, hvoraf de fleste var gengangere fra de to forrige år. Således er det muligt at sammenligne resultater på tværs af årene. Spørgeskemaerne indeholder en række spørgsmål om borgerens sundhed i form af blandt andet psykisk og fysisk trivsel, sygdom, samt eftervirkninger heraf. Derudover indeholder de informationer om personens sundhedsrelaterede adfærd (KRAM-faktorer), samt en række spørgsmål der blandt andet vedrører det sociale liv, den generelle følelse af tillid, samt opfattelse af lokalsamfundets rammer og muligheder.

Svarprocenter

I 2010 blev spørgeskemaet sendt ud til 35.700 tilfældigt udvalgte nordjyske borgere over 15 år, og heraf svarede 23.392 personer. Det gav en svarprocent på 65,5 procent (Region Nordjylland 2011:6). I 2013 blev spørgeskemaet ligeledes sendt ud til 35.700 nordjyder, hvoraf 20.220 denne gang besvarede, og svarprocenten var således på 56,6 procent (Region Nordjylland 2014:18). I 2017 blev 37.600 borgere inviteret til at besvare spørgeskemaet, hvoraf der blev opnået 22.583 besvarelser. Det resulterede i en svarprocent på 60,1 procent (Region Nordjylland 2018:18).

5.1.2. Registerdata fra 2010-2016

Det anvendte registerdata fra Region Nordjyllands patientadministrative system indeholder information om ti forskellige sygdomsgrupper for kroniske sygdomme. Dataene består af samtlige registreringer i årene 2010-2016 ved enten indlæggelse eller ambulant besøg samt indløste recepter til receptpligtig medicin for disse sygdomme. Diagnoserne i det patientadministrative system er baseret på 10. udgave af International Classification of Diseases (ICD-10). De ti sygdomsgrupper er følgende: Adipositas (svær

overvægt), alkoholmisbrugsrelaterede sygdomme, allergiske og inflammatoriske sygdomme, demens, gig, hjertekarsygdomme, KOL eller astma, osteoporose (knogleskørhed), diabetes samt psykiatriske lidelser (Bilag 1). Grundet problemformuleringens efterspørgsel af somatiske kroniske sygdomme, indgår ikke de psykiatriske sygdomme i det samlede mål. Hvilke konkrete somatiske kroniske sygdomsgrupper, der inddrages, vil redegøres for i afsnit "5.2.1. Afhængige variable: Nyopstået somatisk kronisk sygdom".

Inden for den samlede årrække fra 2010-2016 har der i alt været 2.548.584 registreringer, som fordeler sig jævnt over årene. Det skal bemærkes, at der er tale om antal registreringer og ikke antal personer, hvilket vil sige, at den samme person i samme år kan være registreret med flere af de pågældende diagnoser. Kvinder optræder hyppigere i registrene end mænd, da de udgør 54,4 procent af alle registreringer. Derudover ses en overrepræsentation af ældre mennesker, idet hele 32,8 procent af samtlige registreringer er foretaget på personer over 70 år.

5.1.3. Sammenkobling af survey- og registerdata

I dette afsnit beskrives fremgangsmåden, der er anvendt til at koble spørgeskema og registerdata. Eftersom registerdataene med somatisk kronisk sygdom kun indeholder registreringer til og med 2016, og de i dette speciale anvendes til at opnå viden om sygdomme hos undersøgelsens personer frem i tiden, har det udelukkende været meningsgivende at koble spørgeskemadataene fra henholdsvis 2010 og 2013 med registerdata.

Fremgangsmåde

På grund af datasikkerhedslovgivningen var det ikke muligt at få cpr-numre sammen med registerdatasættet, og dette har medført, at vi har måttet koble registerdataene til spørgeskemadataene ud fra bopælskoordinater, x og y, samt køn. Det er ikke lige så optimalt som at koble på cpr-numre, der er unikke for hver person, eftersom der godt kan bo flere personer af samme køn på samme koordinater. Det er eksempelvis ofte tilfældet ved etagebyggeri. Samtidig skal det nævnes, at det ved denne kobling antages, at personer ikke flytter ofte, da der ved flytning er risiko for, at personer af samme køn flytter ind på

samme koordinator. Dermed kan det fremstå som om, personen får en somatisk kronisk sygdom, selvom det i stedet er tilfældet, at det er en ny person med en eksisterende sygdom. Ud fra egne beregninger fra tal fra Danmarks Statistisk flytter omkring 15 procent af Danmarks indbyggere hvert år (Danmarks Statistik 2019a; Danmarks Statistik 2019b). Dette betragtes som relativt få, og derfor antages det, at selvom det kan medføre forkerte estimeringer, vil fejlen ikke være af markant betydning.

Første step i sammenkoblingen har været at slette dupletter i både registerdatasættet og i de enkelte spørgeskemadatasæt – det vil sige, at bopælskoordinater, hvorpå der indgik flere personer af samme køn, er blevet udeladt i den videre klargøring af data. På denne måde blev vi i stand til at tildele de resterende personer i registerdatasættet et unikt id-nummer, der efterfølgende kunne kobles til cpr-nummeret i spørgeskemaet, såfremt bopælskoordinater og køn var de samme for de to datasæt.

For spørgeskemadataene betød sletningen af dupletter et tab på 760 besvarelser i 2010 (fra 23.392 til 22.632) og 441 besvarelser i 2013 (fra 20.220 til 19.779).

I registerdatene blev fremgangsmåden kompliceret af, at den enkelte observation her bestod i forekomsten af en registrering af en diagnose indenfor en specifik sygdomsgruppe hos en person i et specifikt årstal. En enkelt person kan dermed udgøre flere observationer, hvis han eller hun har flere sameksisterende diagnoser i samme år eller optræder med en eller flere diagnose(r) i flere forskellige år. I sammenkoblingsprocessen var det først nødvendigt at konstruere én streng for hver person i hvert år, hvor samtlige diagnoser, denne person måtte have, blev placeret under samme observation. Sagt på en anden måde har vi ændret datastrukturen fra et langt til et bredt layout, hvilket er en almindelig metode til håndtering af data, der består af flere gentagne målinger (Acock 2014:249ff). Det blev gjort ud fra en antagelse om, at forskellige diagnoser tildelt i samme år til en person af samme bopælskoordinater, køn og alder, med størst sandsynlighed svarer til tilstedeværelsen af flere forskellige diagnoser hos den samme person. Selvom alder ikke anvendes til at koble registerdata med spørgeskemadata, er aldersvariablen alligevel benyttet i dette led, hvor det bedømmes, hvorvidt der er tale om den samme person på et specifikt sæt af bopælskoordinater i forskellige år. Aldersvariablen er anvendt som en ekstra sikkerhed

for sammenkoblingsmetoden, idet to observationer fra henholdsvis det ene og andet år kan udelukkes at være den samme person, såfremt alderskategorien ikke er enten konstant eller stigende med ét trin. Tabel 1 nedenfor illustrerer et eksempel på datastrukturen i registerdatasættet, som den så ud til udgangspunktet. I det følgende vil vi bruge flere eksempler til at tydeliggøre de enkelte led i datamanagementprocessen, der er anvendt for at muliggøre en sammenkobling til de enkelte surveydatasæt.

Tabel 1: Eksempel på datastruktur i registerdatasættet 2010-2016.

Observation	x	y	Køn	Alder	Årstal	Kronisk sygdom
1	x1	y1	kvinde	60-69	2010	gigt
2	x1	y1	kvinde	70+	2011	gigt
3	x1	y1	kvinde	60-69	2010	diabetes
4	x1	y1	kvinde	70+	2011	diabetes
5	x1	y1	kvinde	70+	2012	diabetes
6	x2	y2	mand	40-49	2011	gigt
7	x2	y2	mand	40-49	2012	gigt
8	x2	y2	mand	50-59	2013	gigt
9	x2	y2	mand	40-49	2012	diabetes
10	x2	y2	mand	50-59	2013	diabetes

I tabel 1 ovenfor antages personen i observationerne 1-5 således at være den samme, eftersom bopælskoordinater og køn er de samme, og alderskategorien ligeledes er den samme eller plausibelt stigende for hvert årstal. Ligeledes antages det, at personen i observationerne 6-10 er den samme. Af tabel 2 nedenfor fremgår et eksempel til at illustrere det første led i omstruktureringen af data. Bemærk at observationerne 1-3 i denne tabel stadig er den samme kvinde, som i tabel 1 udgør de fem observationer 1-5, og observationerne 4-6 er den samme mand, som i tabel 1 udgør 6-10.

Tabel 2: Eksempel på datastruktur i registerdatasættet 2010-2016, fortsat.

Observation	x	y	Køn	Alder	Årstal	Gigt	Diabetes
1	x1	y1	kvinde	60-69	2010	1	1
2	x1	y1	kvinde	70+	2011	1	1
3	x1	y1	kvinde	70+	2012	.	1
4	x2	y2	mand	40-49	2011	1	.
5	x2	y2	mand	40-49	2012	1	1
6	x2	y2	mand	50-59	2013	1	1

Efter dette led er mange personer stadig gengangere i flere observationer, hvor der dog maksimalt er én observation for hver person for hvert år, som eksemplet ovenfor illustrerer. Efter dette led blev det muligt at slette personer med samme diagnose, der i samme år havde samme køn og bopælskoordinater, hvilket medførte et tab af 238.172 personer i registerdatasættet. Dertil blev 1.412 personer slettet på baggrund af manglende registeroplysninger om køn. Efter det samlede tab på 239.584 endte vi dermed med 1.318.270 personer i årrækken 2010-2016. Selvom tabet kan fremstå højt, har denne frasortering været nødvendig for at kunne besvare specialets problemformulering så validt som muligt. Desuden vil vi pointere, at flere af de frasorterede personer er gengangere, der optræder i registrene flere gange i løbet af årrækken, så der er altså ikke tale om et tab af 239.584 unikke individer.

I det sidste led komprimerede vi dataene yderligere, hvormed personerne hver især udgjorde en enkelt observation, og herefter tildelte vi hver observation/person et unikt id-nummer. I tabel 3 nedenfor fortsættes eksemplet med de to personer, der nu hver udgør en observation med et unikt id-nummer:

Tabel 3: Eksempel på den endelige datastruktur i registerdatasættet 2010-2016.

ID	x	y	Køn	Gigt	Gigt	Gigt	Gigt	Diabetes	Diabetes	Diabetes	Diabetes
1	x1	y1	kvinde	1	1	0	0	1	1	1	0
2	x2	y2	mand	0	1	1	1	0	0	1	1

Efter den endelige tildeling af unikke id-numre fremgår det, at samtlige registreringer i løbet af årrækken er foretaget blandt 327.784 unikke individer. Da den samme kombination af x, y og køn efter denne komprimering af data udelukkende optrådte én gang, og det samme gjorde sig gældende i spørgeskemadataene efter sletning af doubletter, var det nu muligt at koble de to datasæt.

Det er en væsentlig detalje, at vi først udelukkende har koblet registreringer i registrene fra 2010 med spørgeskemadataene fra samme år, og fremgangsmåden var ligeledes den samme for året 2013. Det er gjort ud fra en antagelse om, at sandsynligheden for at det er den samme person på samme adresse er relativt stor, når det er inden for samme år, men at denne sandsynlighed falder jo flere år, der er imellem. Det betyder, at det kun er de personer, der har været registreret i disse to år, det har været muligt at koble. Derefter er registerdataene fra samtlige efterfølgende år koblet til spørgeskemadataene ved hjælp af cpr-nummer og det unikke id-nummer, således at spørgeskemaet fra 2010 indeholder registerdata fra 2010 til 2013, og ligeledes indeholder spørgeskemaet fra 2013 registerdata fra 2013 til 2016. Fra 2010 kunne kobles 5.004 unikke id-numre til spørgeskemaets cpr-numre, mens der i 2013 kunne kobles 7.300. Det lyder måske ikke af meget, når man tænker på, at der i surveydatasættene fra 2010 og 2013 indgår henholdsvis 22.632 og 19.779 unikke cpr-numre efter sletning af gengangere. Her er det dog vigtigt at huske på, at det kun er de personer, der har været registreret i det patientadministrative system med en kronisk sygdom i de pågældende år, der figurerer i registerdatasættet, og flere af disse registreringer forekommer hos de samme (multisyge) personer. Den primære årsag til, at de resterende personer i spørgeskemaet ikke kan kobles til registerdatasættet er altså, at de ikke har været i behandling i systemet i enten 2010 eller 2013. Disse personer bliver i det kombinerede datasæt altså kodet til ikke at have de pågældende sygdomme i disse år. Sammenkoblingens største fejlmargen er dog, at vi ikke kan være sikre på, at årsagen til, at personen ikke optræder i registrene, ikke i stedet skyldes andre forhold. Andre muligheder kan være, at de har optrådt som nogle af de (få) doubletter, vi har været nødt til at slette, eller at de ikke længere er bosat i Region Nordjylland, hvilket eksempelvis kan være på grund af flytning eller død. I det følgende afsnit vil blive redegjort for de teknikker, vi har anvendt til at mindske denne fejlmargen mest muligt, men også til at fastslå i hvor høj grad den udgør en egentlig udfordring for undersøgelsen.

Håndtering af udfordringer i forbindelse med sammenkoblingen af data

Vi har forsøgt at tage højde for de ovenstående mulige fejkodninger af personer som værende uden diagnose(r), selvom der er en reel sandsynlighed for, at de har en eller flere diagnoser i det pågældende år. For det første har vi undladt at erstatte manglende værdier med værdien 0 for "Ingen diagnose" i tilfælde, hvor vi har vurderet, at der er større sandsynlighed for, at personen eksempelvis er afdød ved døden eller er flyttet fra Region Nordjylland. Det har vi gjort ved at kode manglende værdier som datastop i de tilfælde, hvor registeroplysninger har optrådt i tidlige år, men hvor der mangler oplysninger fra et givent år og de resterende år frem i tiden.

Derudover har vi foretaget en test for at fastslå, hvor korrekte vores rekodninger af data er. I testen har vi sammenlignet sygdomme i registrene med folks selvrapporterede diagnoser fra spørgeskemaerne. Begge målinger er fra samme år således, at testene er foretaget på registerdata og spørgeskemadata i henholdsvis 2010 og 2013. Resultatet af undersøgelserne viser et overordnet godt match mellem registrerede sygdomme og selvoplyste sygdomme, men der er visse afvigelser. Det er især gældende i forhold til gigtsygdomme, hvor cirka 20 procent har selvrapporteret at have en gigtsygdom i både 2010 og 2013, selvom registerdataene fra samme år tyder på det modsatte. Her kan man imidlertid forestille sig, at mange af sygdommene ikke er behandlingskrævende i sundhedssystemet i alle år, når diagnosen er stillet, og patienten eksempelvis er selvbehandlende. Vi kunne have valgt at erstatte observationer med folks selvoplyste diagnose i tilfælde, hvor vi har været nødsaget til at antage "ingen diagnose" på grund af manglende værdier i registerdataene. Det har vi imidlertid undladt, eftersom vi mener, at det er mere retvisende at forholde sig til, hvorvidt en person i det pågældende år har en behandlingskrævende diagnose, der figurerer i systemet. Det skyldes, at undersøgelsen er funderet i ønsket om at forudsige, hvem der bliver særligt omkostningsfulde for Region Nordjylland. Resultater fra testen fremgår af bilag 5.

5.2. Beskrivelse af de anvendte variable

Det følgende afsnit består af henholdsvis en præsentation af de afhængige variable samt de uafhængige variable, der er anvendt i dette speciale.

5.2.1. Afhængige variable: Nyopstået somatisk kronisk sygdom

Som vi allerede indledningsvist i dette speciale var inde på, har vi med udgangspunkt i eksisterende forskning valgt at undersøge forekomsten af somatiske kroniske sygdomme som en indikator for, hvornår en person kan tænkes at være hyppig bruger af sundhedsvæsenet (jf. indledningen). Der indgår to afhængige variable i dette speciale: En for undersøgelsesåret 2010, der indeholder somatiske kroniske sygdomme i årene 2011-2013 og en for undersøgelsesåret 2013, der indeholder somatiske kroniske sygdomme i årene 2014-2016. De afhængige variable består hver især af et samlet mål for forekomsten af en eller flere af en række udvalgte somatiske kroniske sygdomme. I konstruktionen af dette mål har vi foretaget nogle afgrænsninger, som vi her vil gøre rede for.

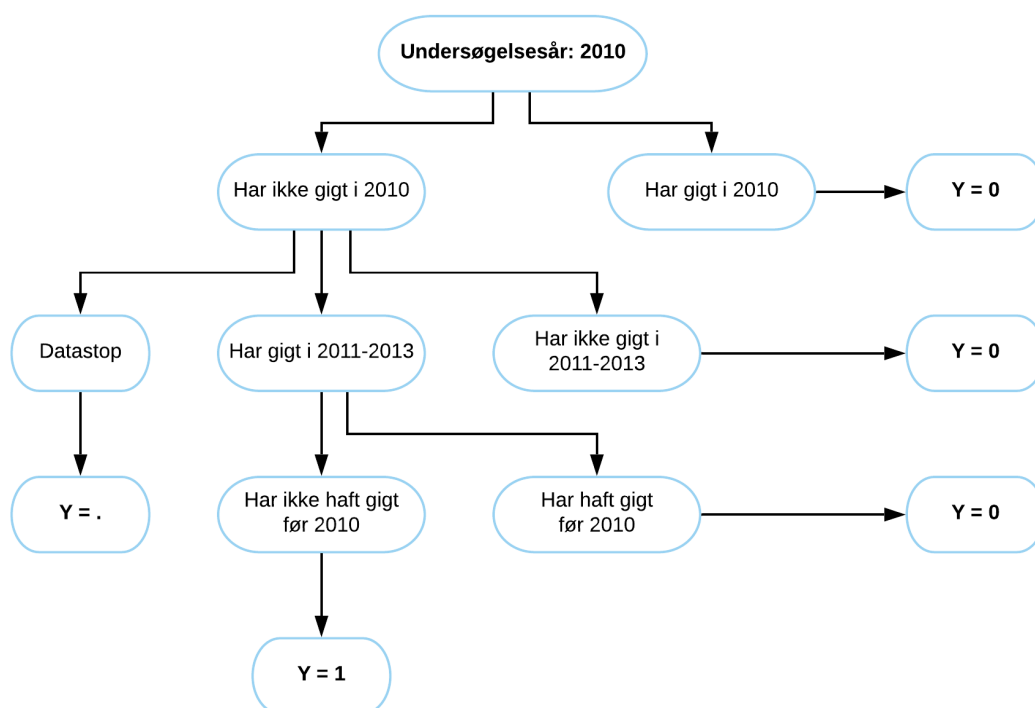
Afgrænsning: Somatiske kroniske sygdomme

Af de ti kroniske sygdomsgrupper fra Region Nordjyllands patientadministrative system indgår følgende sygdomme i de afhængige mål for forekomsten af nyopstået somatisk kronisk sygdom: Astma og KOL, allergiske og inflammatoriske sygdomme, diabetes, gigt, knogleskørhed samt hjertekarsygdomme. Samtlige af disse sygdomme optræder i Sundhedsstyrelsens rapport "Sygdomsbyrden i Danmark" fra 2015, hvor de er blevet udvalgt på baggrund af både deres afgørende betydning i de danske sundhedsbudgetter, men også fordi de vurderes at have forebyggelsespotentiale (Flachs et al. 2015:48). Vi har lagt dem sammen til et sammensat mål, hvor variabelen er kodet som en dummy, og værdien 1 opnås, hvis personen får en hvilken som helst somatisk kronisk sygdom. Selvom alle kroniske sygdomme er aldersbetingede, har alder imidlertid en særligt stor betydning for demens (Nationalt Videnscenter for Demens 2019). Aldersaspektet er ikke af særlig interesse i dette speciale, og demens er af samme årsag ekskluderet fra det samlede afhængige mål for kronisk sygdom. Derudover har vi, som nævnt i indledningen, afgrænset os til somatiske sygdomme, og sygdomsgruppen af psykiatriske sygdomme er derfor udeladt. Ligeledes kan alkoholmisbrug anses som værende en psykiatrisk lidelse (Psykiatrifonden 2019), og selvom der er i det patientadministrative system er tale om en registrering af følgesygdomme efter alkoholmisbrug, har vi valgt at ekskludere denne sygdom. Slutteligt skyldes ekskluderingen af adipositas (ekstrem fedme), at denne sygdom optræder hos en meget lav andel af populationen (0,01 procent), og samtidig indgår en variabel for BMI som en del af de forklarende variable.

Konstruktion af de afhængige variable

I konstruktionen af den afhængige variabel for undersøgelsesåret 2010, har vi givet værdien 1, hvis en person i årene 2011-2013 har fået en ny somatisk kronisk sygdom, og værdien 0 hvis ikke. At få en ny sygdom indebærer, at en eller flere af de specifikke udvalgte sygdomme optræder hos en person i årene 2011-2013, og personen ikke havde sygdommen i forvejen. Eksempelvis har personer, der ikke er registreret med allergiske eller inflammatoriske sygdomme i 2010, men som figurerer i registret i enten 2011, 2012 eller 2013 med denne sygdom, fået tildelt værdien 1 for nyopstået sygdom. Hvis en person har en af de somatiske kroniske sygdomme i forvejen, hvilket eksempelvis kan være gigt, og får en af de øvrige sygdomme, vil personen ligeledes få værdien 1 på den afhængige variabel. Derudover har personer, der i spørgeskemaet har svaret "Har haft tidligere" til en specifik sygdom i stedet fået tildelt værdien 0, da de ikke betragtes som havende fået en ny sygdom. Som tidligere nævnt skal det desuden bemærkes, at personer der optræder i registerdatasættet i nogle år, men hvor registreringerne stopper i et givent år, er blevet frasorteret med begrundelsen "datastop". Nedenfor ses et flowchart over kodningen af den afhængige variabel for undersøgelsesåret 2010, hvor forekomsten af gigt er brugt som eksempel.

Figur 3: Flowchart over forekomsten af nyopstået gigtsygdom fra 2010 og tre år frem.



Samme fremgangsmåde er fulgt i konstruktionen af den afhængige variabel for spørgeskemaundersøgelsen fra 2013, hvor nyopstået somatisk kronisk sygdom angiver, at individet har udviklet en eller flere af de udvalgte sygdomme i løbet af årrækken 2014-2016. Når vi fremadrettet henviser til data fra årene 2010 og 2013, inkluderer det således også de afhængige mål for somatisk kronisk sygdom i henholdsvis årene 2011-2013 og 2014-2016. Ligeledes henviser data fra år 2017 til fremtidig sygdomsforekomst i årene 2018-2020.

De studier, vi har ladet os inspirere af i dette speciale i forhold til den prædiktive metode, der anvendes, har blot studeret et outcome ét år ud i fremtiden (Perkins et al. 2004; Fleishman & Cohen 2010; Orueta et al. 2013; Charlson et al. 2014; Shen et al. 2014; Lu et al. 2015; Tamang et al. 2017; Maidman & Wang 2018). Vi betragter det imidlertid som en fordel at kunne undersøge sygdomsforekomsten flere år frem, idet det må antages, at jo flere år undersøgelsesdesignet i dette speciale strækker sig over, des flere nye tilfælde af somatiske kroniske sygdomme vil vi kunne indfange. Da det er en forudsætning for at kunne foretage prædiktioner ved hjælp af supervised machine learning, at de anvendte datasæt er ens for alle tre år (Brownlee 2017a; Raschka & Mirjalili 2017:3f), har vi dog en maksimal grænse i forhold til at kunne studere mulig sygdomsforekomst tre år ud i fremtiden. Det skyldes, at der for datasættet i 2013 maksimalt kan kobles registerdata frem til 2016. En uddybning af supervised machine learning vil forekomme af afsnit "6.2. Machine learning". Selvom det ikke har været muligt at studere det ønskede udfald længere frem end tre år fra måletidspunktet for de forklarende variable, hvor flere år havde været mere ønskværdigt, så betragtes det alligevel som anvendeligt. Det skyldes en antagelse om en vis kontinuitet i individets adfærd og øvrige determinanter. Denne antagelse har vi på baggrund af Bourdieus teoretiske begrebsapparat, hvor det betragtes, at det enkelte individ indgår i et differentieret og differentierende system af sociale positioner, og dette system er med til at definere både individets sociale position og dets *habitus*, som har en indvirkning på individets livsstil og handlinger (Bourdieu 1984:166; Bourdieu 1997:24; Bourdieu 2000:130). Disse teoretiske betragtninger vil vi komme mere ind på i afsnit 7, der udgør dette speciales teoriafsnit angående de sociale og psykologiske determinanter. Pointen er, at det på baggrund af Bourdieus teoretiske perspektiv antages, at når der på et givent tidspunkt eksempelvis haves en bestemt

socioøkonomisk status eller bestemt sundhedsrelateret adfærd, har dette med stor sandsynlighed også været gældende tidligere.

I dette speciale er de afhængige variable som nævnt konstrueret således, at individer optræder som havende en nyopstået somatisk kronisk sygdom, hvis den specifikke sygdom optræder til tidspunkt 2 og samtidig ikke optræder til tidspunkt 1 eller tidligere. Haves sygdommen i forvejen, optræder individet som "0 – Ingen nyopstået sygdom" på lige fod med personer, der hverken har sygdommen til tidspunkt 1 eller 2. På den måde bliver personer med en nyopstået sygdom stillet over for både personer, der ikke får sygdommen (og aldrig har haft den) men også med personer, som allerede har sygdommen i forvejen. Det anses som problematisk, idet sandsynligheden for at tilhøre gruppen af personer med værdien "1 – Nyopstået sygdom" bliver mindre jo flere af de udvalgte sygdomme, personen har i forvejen. Ud fra denne betragtning havde det været en fordel at frasortere personer, der var syge til tidspunkt 1, og kun sammenligne den fremtidige eventuelle sygdomsforekomst for personer, der var raske i udgangspunktet. En måde at løse dette bias vil være en ændring af undersøgelsens design, hvor blot én udvalgt sygdom studeres eller eventuelt et design, hvor de enkelte sygdomme studeres separat. En anden fordel ved at undersøge de respektive sygdomme enkeltvis er, at det kan tænkes, at de sociale og psykologiske determinanter indvirker forskelligt på forskellige sygdomme (se fx Lange 2015; Almdal 2017). De påviste effekter af determinanterne i forhold til det samlede mål, der fremkommer i specialets analyse, kan således tænkes at være influeret af forskellige kausale årsagsforhold, som udfordrer undersøgelses interne validitet. På den baggrund kan separate undersøgelser tænkes at lette arbejdet i forhold til at sikre de rette forebyggende indsatser for den enkelte sygdom. Bevæggrunden for at lave et samlet mål for somatiske kroniske sygdomme var imidlertid, at vi ikke var interesserede i én specifik sygdom men i stedet havde et ønske om at kunne sige noget generelt om den overordnede "sundhedsbyrde". Det betyder også, at selvom en person for eksempel har en hjertekarsygdom, vil vi stadig kunne indfange nyopståede tilfælde af andre sygdomme. Det vurderes, at et sammensat mål har været nødvendigt for at kunne sige noget om flere sygdomme i dette speciale, eftersom det ikke har været muligt at lave et undersøgelsesdesign bestående af separate undersøgelser af samtlige sygdomme, fordi et sådant design havde været markant mere omfattende.

Deskriptive statistikker over de afhængige variable

Af tabel 1 og 2 nedenfor fremgår hvor mange personer der får en ny somatisk kronisk sygdom i de efterfølgende tre år fra de to undersøgelsestidspunkter. Som det fremgår af tabel 1, udgør gruppen af personer, der får en ny sygdom i årene 2011-2013, 7 procent af det samlede sample på 22.335 personer.

Tabel 4: Frekvenstabel over den afhængige variabel (Y) i 2010.

	Frekvens	Procent
Nyopstået sygdom (Y = 1)	1.566	7,0
Ingen nyopstået sygdom (Y = 0)	20.769	93,0
Total	22.335 *	100,0

* Slettet pga. datastop: 297 personer

I tabel 4 ses frekvensen over den afhængige variabel i 2013, hvor gruppen af personer, der får en ny sygdom i årene 2014-2016, udgør en lidt højere andel af det samlede sample på 19.014 personer. Her har lidt over 11 procent af samplet en nyopstået somatisk kronisk sygdom.

Tabel 5: Frekvenstabel over den afhængige variabel (Y) i 2013.

	Frekvens	Procent
Nyopstået sygdom (Y = 1)	2.157	11,3
Ingen nyopstået sygdom (Y = 0)	16.857	88,7
Total	19.014 *	100

* Slettet pga. datastop: 765 personer

5.2.2. Uafhængige variable

De uafhængige variable i dette speciale er først og fremmest udvalgt på baggrund af et ønske om at konstruere en model, der er så nuanceret som muligt. Derfor har vi inddraget flest muligt variable fra både Sundhedsprofilerne og registerdata. Eftersom vi ikke på forhånd har udvalgt nogle specifikke variable på baggrund af antagelser om kausale

sammenhænge, må udvælgelsen af de forklarende variable i dette speciale siges at være induktivt funderet. I sin artikel beskriver Domingos (2012) den induktive fremgangsmåde som en løftestang til viden:

“[Induction] turns a small amount of input knowledge into a large amount of output knowledge. Induction is a vastly more powerful lever than deduction, requiring much less input knowledge to produce useful results, but it still needs more than zero input knowledge to work. And, as with any lever, the more we put in, the more we can get out.” (Domingos 2012:81).

Ud fra dette citat kan den induktive fremgangsmåde og inddragelsen af så mange variable som muligt, betragtes som en fordel. Kuhn og Johnson (2013) er imidlertid mere skeptiske over for den rent induktive fremgangsmåde. Ifølge dem bør det ikke være helt tilfældigt, hvilke variable der indgår i en prædiktionsmodel, da en prædiktionsmodel altid vil finde forudsigelser. I stedet bør man som forsker anvende sin ekspertviden til at indhente relevante data i forhold til det undersøgelsesspørgsmål, man søger at besvare: *“In the end, [predictive modeling] is not a substitute for intuition, but rather a complement”* (Kuhn & Johnson 2013:5). Sundhedsprofilerne indeholder imidlertid ingen variable, der ud fra vores sociologiske forståelse af sygdomsårsager er meningsløse at inddrage. Således er alle de anvendte variable nogle, der kan genfindes i den eksisterende forskning samt i de sundheds- og sygdomssociologiske forklaringsmodeller til ulighed i sundhed, som vi kort har beskrevet indledningsvist i specialet.

Fordi formålet med denne undersøgelse er at kunne forudsige et udfald over tid, har vi dog udelukkende anvendt variable, der er til stede i spørgeskemaundersøgelserne i alle tre år. I det følgende vil forekomme en præsentation af specialets i alt 31 uafhængige variable, under de seks førnævnte temaer.

For det først indgår i alt fem **socioøkonomiske variable**, henholdsvis køn, alder, indkomst, etnicitet og uddannelse. De er alle fra registerdata, med undtagelse af uddannelsesvariablen, der er fra spørgeskemadataene.

Næste tema handler om **trivsel** og indeholder i alt seks variable. Tre variable for hvor ofte respondenter inden for de seneste fire uger har henholdsvis følt sig rolig og afslappet, været fuld af energi samt følt sig trist til mode. Derudover indgår en variabel for, hvorvidt respondenter er ængstelig eller deprimeret, samt en variabel for overordnet livstilfredshed. Temaets sidste variabel er "Sheldon Cohens 10-Item Perceived Stress Scale" (PSS), som er en skala til måling af oplevet stress hos voksne. Skalaen er valideret i Danmark (Eskildsen et al. 2015).

I temaet **socialt samvær** indgår fem variable. Der indgår tre variable, der handler om, hvor ofte respondenter har kontakt til henholdsvis familie han eller hun ikke bor sammen med, venner og naboer. Derudover er der et spørgsmål, der handler om, hvor ofte respondenter er alene, selvom individet har lyst til selskab, samt hvorvidt respondenter har nogle at tale med, hvis han eller hun har behov for hjælp eller støtte.

I temaet **sundhedsrelateret adfærd** indgår ni variable, der relaterer sig til KRAM-faktorerne, altså **kost**, **rygning**, **alkohol** og **motion**. Det indebærer et spørgsmål om hvorvidt og hvor ofte respondenter ryger. Yderligere indgår to spørgsmål om henholdsvis, hvor ofte respondenter drikker alkohol i det hele taget, samt hvor ofte der drikkes mere end fem genstande ved samme lejlighed. To spørgsmål om brugen af fedtstof i kosten, hvilket er henholdsvis, hvor ofte respondenter spiser forskellige typer af fedtstof på brød, samt hvor ofte der anvendes fedtstof i madlavningen. Her har respondenter haft mulighed for at svare på, hvor hyppigt der bruges fedtstof inden for forskellige typer af madlavning eller på forskellige brødtyper. Disse har vi efterfølgende lagt sammen til henholdsvis et sammensat mål for den samlede anvendte mængde fedt på brød, samt et sammensat mål for den samlede anvendte mængde fedt i madlavningen. Derudover er der et spørgsmål, der omhandler, hvor ofte der indtages grøntsager, samt hvor ofte der indtages frugt. I forhold til motion indgår en variabel for respondenterens fysiske aktivitetsniveau i fritiden. Endelig indeholder temaet desuden en variabel for BMI.

Det næstsidste tema handler om respondenterens **lokalområde**. Her indgår tre spørgsmål omhandlende i hvor høj grad respondenter stoler på personer i sit lokalområde, hvor tæt knyttet han eller hun føler sig til lokalområdet, samt hvor tryk han eller hun føler sig ved at gå alene i området efter mørkets frembrud.

Det sidste tema handler om **tillid** og indeholder ligeledes tre variable. De to første angiver, i hvor høj grad respondenter mener, at man henholdsvis kan stole på de fleste mennesker, eller modsat i hvor høj grad man bør være forsigtig, når man har med andre mennesker at gøre. Det sidste spørgsmål har vi konstrueret som et samlet mål for graden af tillid til forskellige samfundsmæssige instanser: Sundhedsvæsen, politi, sociale myndigheder, domstole og Folketinget.

Imputation

I nærværende speciale er dele af analysen udført i programmet Python, og heri er det ikke muligt at udføre analyser på variable, hvor der findes manglende værdier. For at løse denne problematik er der anvendt multipel imputation. Med denne metode anvendes andre variable til at udregne de manglende værdier (Acock 2014:395ff). Det kan altså siges, at der med udgangspunkt i mønstre i respondenternes besvarelser udregnes kvalificerede bud på, hvad de manglende værdier kan være. Der findes flere forskellige metoder inden for multipel imputation. Når de manglende værdier i et datasæt ikke er tilfældige, men følger et bestemt mønster, kan man anvende Stata kommandoen *impute monotone*, der udfylder værdier på flere variable, der har et såkaldt *monotone-missing pattern* (StataCorp 2017:192). Monotone missing pattern forekommer, når der optræder manglende værdier i den samme observation for flere variable (StataCorp 2017:6f). I vores tilfælde betyder det, at nogle respondenter har undladt at besvare en række af Sundhedsprofilens spørgsmål. Det er et alment kendt fænomen inden for surveymetodik, at der er nogle personer end andre, der i højere grad undlader at besvare dele af spørgeskemaet, hvilket også betegnes *missing data bias* (de Vaus 2014:358). De variable, der er anvendt til at imputere manglende værdier i dette speciale, er henholdsvis køn, alder, indkomst, uddannelse og etnicitet. Eftersom samtlige af disse variable, med undtagelse af uddannelse, er fra registerdatasættet, indeholder de ingen manglende værdier, hvilket betragtes som en fordel, når de anvendes til dette formål. Uddannelsesvariablen fra 2010 indeholdte 707 manglende værdier, mens den fra 2013 indeholdt 1258, og den fra 2017 indeholdt 2750. Imputationsmetoden er blevet anvendt på uddannelsesvariablen som det første, inden denne indgik i imputation på de øvrige variable. Det er en almindelig procedure, at variable med imputerede værdier indgår i imputationen af andre variable (StataCorp 2017:192). Vi kunne have undladet at bruge uddannelsesvariablen som imputationsvariabel, men denne betragtes, sammen med de

øvrige socioøkonomiske baggrundsvariable, som særligt vigtige i forhold til svarmønstret i vores spørgeskemadata. Som vi har redegjort for i specialets problemfelt, viser flere undersøgelser nemlig, at uddannelse har en signifikant effekt på forskellige mål for sundhed og sundhedsrelateret adfærd (jf. problemfeltet), og derfor må det også forventes, at uddannelsesniveau har en betydning for respondenternes svar i Sundhedsprofilundersøgelserne.

I spørgsmålet "Alt taget i betragtning - hvor tilfreds eller utilfreds er du for tiden med livet?", hvor svarkategorierne er en skala 1-10 samt kategorien "Ved ikke", har vi tildelt midterkategorien (5), i tilfælde hvor respondenter har svaret "Ved ikke", i stedet for at angive dem som manglende svar. Hermed antages "Ved ikke" som udtryk for, at respondenter er middel tilfreds. Det kan diskuteres, hvorvidt denne antagelse er plausibel, men valget er taget, da vi ellers skulle imputere på disse besvarelser, hvilket betragtes som mindre retvisende.

Af bilag 2, 3 og 4 fremgår deskriptive statistikker af de anvendte forklarende variable. Der er både inddraget statistikker på variablene før og efter imputationen er foretaget, så det er muligt at se, hvor mange manglende værdier de oprindelige variable har haft, hvilke konsekvenser imputationen har haft for de procentvise fordelinger af de kategoriske variable, samt gennemsnittet af de intervallskalerede variable.

Standardisering

Fordi input-variablene i en neural netværksanalyse skal være standardiserede (Hain & Jurowetzki 2018:23), er dette udført for samtlige uafhængige variable, der er anvendt i dette speciale, således at værdierne placerer sig indenfor et interval mellem 0 og 1. Derudover er det også vurderet, at en sådan standardisering medfører en mere intuitiv fortolkning af odds ratio og parameterestimer i en logistisk regressionsmodel.

6. Metodeafsnit

Følgende afsnit vil først indebære det forskningsdesign, der anvendes i dette speciale, og yderligere vil specialets undersøgelsesdesign beskrives. Efterfølgende vil vi redegøre for machine learning. Herunder vil indgå, hvad machine learning er samt en uddybende forklaring af prædiktive modeller og supervised learning, som er den form for machine learning, der anvendes i dette speciale, og afslutningsvist vil indgå en redegørelse for forskellige typer af algoritmer. Derefter vil specialets undersøgelsesmetoder til besvarelse af første del af problemformuleringen beskrives, hvilket som nævnt er logistisk regression og neuralt netværk. Efterfølgende vil forekomme en diskussion af de anvendte undersøgelsesmetoder, hvor de to metoders potentialer vil diskuteres. Herefter vil komme en redegørelse for den fremgangsmåde, der er brugt til besvarelsen af problemformuleringens anden del, herunder en redegørelse for den områdeinddeling af Region Nordjylland, der anvendes i specialet. Metodeafsnittet vil afsluttes med en inddragelse af kvalitetskriterierne intern validitet, ekstern validitet og reliabilitet.

6.1. Forsknings- og undersøgelsesdesign

I en undersøgelse kan der skelnes mellem et forskningsdesign og en arbejdsplan (de Vaus 2001:8f). Dette afsnit vil indeholde en redegørelse for begge dele, hvor arbejdsplanen dog her vil omtales som undersøgelsesdesign. De to adskiller sig ved, at forskningsdesignets formål i højere grad er overordnet, og dets funktion indebærer at kunne besvare problemformuleringen så entydigt som muligt ud fra den opnåede evidens. Derimod er undersøgelsesdesignets funktion mere lavpraktisk på den måde, at det er, hvad der konkret skal foretages for at gennemføre projektet, og det er altså ud fra forskningsdesignet, at undersøgelsesdesignet udarbejdes (de Vaus 2001:9).

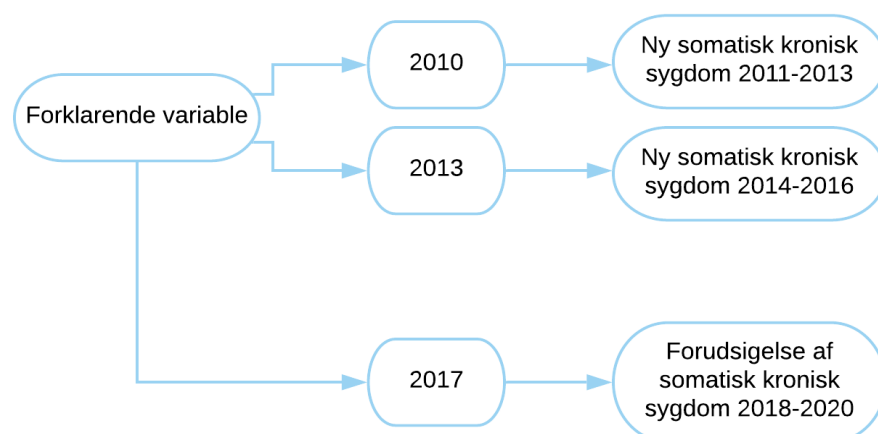
For bedst muligt at besvare problemformuleringen, findes det anvendeligt med et longitudinalt forskningsdesign, fordi der indgår et tidselement heri, hvormed det er muligt at undersøge ændring over tid (de Vaus 2001:113). I dette speciale ønsker vi netop at forudsige forekomsten af somatiske kroniske sygdomme. Det betragtes, at sundhedsprofilerne er foretaget som gentagne tværsnitsstudier, da der er indsamlet information over tid og med sammenlignelige samples frem for på de samme individer

på tværs af år (ibid.). Koblingen til registerdataene muliggør imidlertid et panelstudie, hvor de samme personer følges over tid (ibid.). Eftersom vi i dette speciale undersøger forekomsten af nyopstået somatisk kronisk sygdom de efterfølgende tre år fra måletidspunktet for de forklarende variable, er det vigtigt at være opmærksom på, at undersøgelsens resultater ikke nødvendigvis kan reproducere, såfremt der eksempelvis er et ønske om blot at forudsige noget i det første efterfølgende år, hvilket prædiktative studier inden for medicinsk- og epidemiologisk forskning ofte har til hensigt (Perkins et al. 2004; Fleishman & Cohen 2010; Orueta et al. 2013; Charlson et al. 2014; Shen et al. 2014; Lu et al. 2015; Tamang et al. 2017; Maidman & Wang 2018).

Fordi vi kun har data over somatiske kroniske sygdomme til og med år 2016, betyder det, at vi ikke på nuværende tidspunkt er i stand til at undersøge, hvorvidt forklarende variable fra spørgeskemaet i 2017 har en effekt på, om personen udvikler en kronisk sygdom i årene 2018-2020. Med supervised machine learning er det imidlertid muligt at konstruere en algoritme på baggrund af de observerede sammenhænge fra 2010 og 2013, som gør os i stand til at forudsige udfaldet på uobserverede data frem i tiden - i hvert fald såfremt algoritmen viser sig at have en høj grad af generaliserbarhed (Raschka & Mirjalili 2017:12; Molina & Garip 2019:1). Med andre ord kan supervised machine learning give et kvalificeret bud på, hvem der vil optræde i Region Nordjyllands patientadministrative system i årene 2018-2020. Denne forudsigelse resulterer i en prædikeret sandsynlighed for fremtidig sygdomsforekomst for det enkelte individ, der har besvaret spørgeskemaet i 2017. Fordi machine learning ikke egner sig til at fastslå kausale sammenhænge (Hain & Jurowetski 2018:3), vil målet for de prædikterede sandsynligheder efterfølgende undersøges som afhængig variabel i en logistisk regression, for også for dette år at kunne fastslå, hvilke sociale og psykologiske determinanter der har betydning for udfaldet herpå. Dette vil blive uddybet yderligere i afsnit "6.5. Forudsigelse af somatiske kroniske sygdomme". På figur 4 nedenfor ses en model over specialets undersøgelsesdesign.

Som det ses i figuren, anvendes de forklarende variable i datasættene fra 2010 og 2013 til at undersøge, hvorvidt respondenterne ifølge det patientadministrative system får en nyopstået somatisk kronisk sygdom i henholdsvis årene 2011-2013 eller 2014-2016.

Figur 4: Undersøgelsesdesign.



Specialets afsluttes med en kortlægning af forekomsten af somatiske kroniske sygdomme i Region Nordjylland. Denne sidste analysedel vil bestå af tre kort, hvor de to første illustrerer henholdsvis den faktiske andel med en somatisk kronisk sygdom i bestemte geografisk afgrænsede områder i 2014-2016, samt den prædikterede sandsynlighed for forekomsten af somatisk kronisk sygdom samme årrække. Når der i dette speciale skrives “den faktiske andel” med en somatisk kronisk sygdom, henvises der til sygdomme, der fremgår af registeret i det patientadministrative system. Det sidste kort illustrerer den prædikterede sandsynlighed for forekomsten af somatisk kronisk sygdom i årene 2018-2020. De prædikterede sandsynligheder er som nævnt baseret på de forudsigelser, der opnås på baggrund af den konstruerede algoritme. En redegørelse for fremgangsmåden for denne kortlægning vil forekomme af afsnit “6.6. Kortlægning af somatiske kroniske sygdomme”.

6.2. Machine learning

Machine learning udgør i dag en grundpille inden for mange erhverv - eksempelvis business-, finans-, produktions- og detailsektoren, og bruges til mange forskellige formål. Dette er lige fra diagnosticering af sygdomme til vejrmødeller (Nielsen 2010:99+103; Kreatsoulas & Subramanian 2018:347). De succesfulde eksempler har åbnet øjnene og skabt interesse for machine learning, også inden for det samfundsvidenskabelige felt, og anvendelsen finder således i større og større udstrækning vej ind til både økonomien,

statskundskaben og sociologien (Molina & Garip 2019:1; Kreatsoulas & Subramanian 2018:347).

6.2.1. Hvad er machine learning?

Overordnet set kan det betragtes, at machine learning er en særlig form for kunstig intelligens, der placerer sig et sted mellem statistik og computervidenskab (Nielsen 2010:92; Molina & Garip 2019:1). Det særlige ved denne form for kunstig intelligens er den adaptive teknik, der anvendes, som gør, at maskinen kan finde sammenhænge og udarbejde forudsigelser uden at blive eksplicit programmeret til, hvorledes dette skal gøres (Educba 2019). Den adaptive teknik går ud på, at modellernes præcision gradvist forbedres gennem evalueringer, idet maskinen hele tiden lærer sig selv at blive mere præcis. I stedet for at mennesker manuelt skal udlede regler og konstruere modeller ud fra analyser, tilbyder machine learning altså mere effektive metoder til at indsamle viden fra store datamængder (Raschka & Mirjalili 2017:2).

Information udvindes fra data ved hjælp af en machine learning *algoritme*. En algoritme kan groft sagt betegnes som opskriften på en løsning til et problem, og hver algoritme er ofte konstrueret til at løse ét specifikt problem. Det kan eksempelvis være: "Find det største hele positive tal, der går op i både 60 og 24" (Clausen & Rasmussen 2012; Hjorth 2017).

På baggrund af den information, der udvindes, estimeres komplekse funktioner, der enten forudsiger data ud fra et output, der er kendt på forhånd, eller induktivt finder mønstre i data (Jordan & Mitchell 2015:255). Således skelnes der mellem supervised machine learning på den ene side, hvor outputtet (Y) observeres ud fra kendte inputs (X). Ud fra disse observationer lærer maskinen at forudsige outputtet på nye input-data, hvor outputtet er ukendt. På den anden side findes unsupervised machine learning, hvor maskinen udelukkende observerer input-data og finder mønstre i dette uden en "lærer", der kan give de rigtige svar - ofte ud fra en antagelse om, at der ikke findes "korrekte svar" (Molina & Garip 2019:1).

6.2.2. Prædiktive modeller

Som vi har nævnt i indledningen til dette speciale, anvender vi prædiktiv modellering til besvarelse af specialets problemformulering. I prædiktive modeller er formålet at konstruere en model, der kan forudsige et givent udfald så præcist som muligt (Kuhn & Johnson 2013:2). Der er tale om et klassifikationsproblem, hvis der er et ønske om at konstruere en model, der kan forudsige udfaldet på en kategorisk output-variabel (Brownlee 2017a). I dette speciale er der tale om et binært klassifikationsproblem, eftersom vi ønsker at klassificere, hvorvidt den enkelte person i undersøgelsen får en somatisk kronisk sygdom i de følgende tre år eller ej. Prædiktive modeller hører ind under kategorien supervised machine learning, eftersom forudsigelserne af (fremtidige) uobserverede output-data baseres på data, hvor outputtet er kendt på forhånd (Brownlee 2017a). I det følgende afsnit vil vi derfor give en mere uddybende redegørelse af supervised machine learning.

6.2.3. Supervised machine learning

Som nævnt er skellet mellem supervised og unsupervised machine learning angivet ved, at der i førstnævnte indgår en læringsproces, hvor maskinen gradvist bliver klogere, hvilket ikke er gældende for sidstnævnte. Denne læringsproces er baseret på en opdeling i henholdsvis et trænings- og testdatasæt. Træningsdatasættet bruges til at træne modellen, og testdatasættet bruges til at evaluere den. Hvis der er tilfredshed med resultatet i evalueringsfasen, kan modellen efterfølgende bruges til at forudsige det ønskede output på nye, uobserverede data - det vil sige data, hvor det ønskede output er ukendt (Raschka & Mirjalili 2017:12). For at kunne konstruere en model til at forudsige et bestemt outcome på nye input-data, er det en forudsætning, at der i både trænings- og testdatasættet er tilgængelige data om det output, der er et ønske om at kunne forudsige i fremtiden (Brownlee 2017a; Raschka & Mirjalili 2017:3f).

Det bliver af nogle problematiseret, hvis modellen både trænes og testes i samme datasæt. Det skyldes, at hvis datasættet har en skæv fordeling, og modellen trænes og testes med samme datasæt, vil fejlen ikke opdages i testfasen, eftersom testdatasættet er lige så skævt fordelt som træningssættet. For at opnå den største sikkerhed for at den konstruerede model også har gyldighed i andre populationer, er det derfor bedst at teste

modellen på et nyt datasæt (Kreatsoulas & Subramanian 2018:349). Det er almindeligt at inddele træningsdatasættet yderligere i et trænings- og valideringsdatasæt, som udgør grundlaget for konstruktionen af modellen, for så til sidst at anvende et separat testdatasæt til evaluering af den endelige model (Molina & Garip 2019:3; Murphy 2012:23). I en sådan opdeling anvendes træningsdatasættet således til at træne modellen, mens det er valideringsdatasættet - og ikke testdatasættet - der bruges til at udvælge forskellige kriterier for modellen. Ofte bruges i sådan en opdeling cirka 80 procent til træning, og de resterende 20 procent til validering (Murphy 2012:23). Hvilke kriterier, der anvendes i den neurale netværksmodel i dette speciale, vil blive præsenteret i afsnit "6.3.2. Neural netværksanalyse" samt "8.2.1. Konstruktion af modellen".

Evalueringen i et klassifikationsspørgsmål baseres oftest på målet "prediction accuracy". Dette mål er givet ved andelen af korrekte prædiktioner divideret med det totale antal prædiktioner (Kotsiantis 2007:250; Brownlee 2017). De fire mulige udfald for en prædiktion er True Positive (TP), True Negative (TN), False Negative (FN), False Positive (FP), som angiver henholdsvis korrekte og forkerte forudsigelser af de to værdier på den afhængige variabel, hvor 0 er negativ og 1 er positiv (Saito & Rehmsmeier 2015:4). "Prediction accuracy" - ofte forkortet AUC - kan således beregnes ved hjælp af formelen:

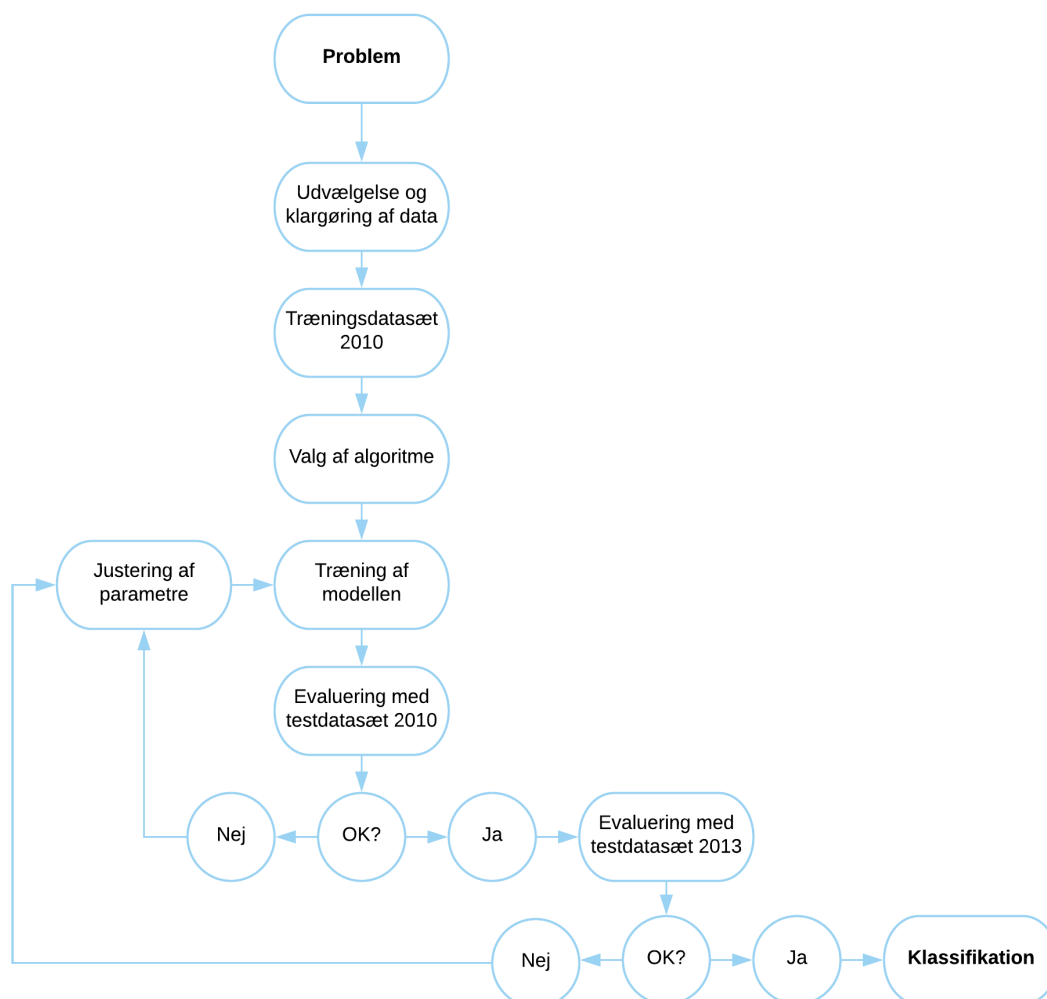
$$AUC = \frac{TP+TN}{TP+TN+FN+FP}$$

Kilde: Navlani (2018)

AUC kan antage alle værdier mellem 0 og 1, hvor en score på 1 betyder, at klassificeringen er perfekt, mens den ved en score på 0,5 betragtes som ubrugelig (ibid.). Hvis prædiktionsevnen ikke er tilfredsstillende, justeres modellen, hvorefter trænings-, validerings- og testproceduren gentages (Kotsiantis 2007:250). Der er således tale om en iterativ fremgangsmåde, hvor det er muligt at justere og prøve sig frem, indtil et tilfredsstillende resultat opnås. Fordi AUC-scoren er et udtryk for, hvor præcist en model kan forudsige, betragter vi det således, at AUC-scoren i testdatasættet er et udtryk for, hvor generaliserbar modellen er.

I dette speciale er prædiktionsmodellerne konstrueret ved hjælp af datasættet fra 2010, og dette datasæt udgør således vores træningsdata. Dette datasæt er, som netop beskrevet, inddelt i henholdsvis et trænings- og valideringsdatasæt, hvor vi ved hjælp af den iterative proces gennem evaluering og parameterjusteringer har fundet frem til de endelige prædiktionsmodeller. Slutteligt har vi anvendt datasættet fra 2013 som testdatasæt til at teste præcisionen af den endelige prædiktionsmodel. Grundet inddragelsen af samme spørgeskema i forskellige år har det således været muligt at teste modellens præcision til et nyt tidspunkt, hvilket betragtes som en fordel, når det ønskes at fastslå, hvor god modellen er til at forudsige det ønskede outcome på fremtidige data. Efter konstruktionen vil modellen som nævnt anvendes til at forudsige, hvem der får en somatisk kronisk sygdom i årene 2018-2020, hvilket i nedenstående figur illustreres som "klassifikation". Figur 5 illustrerer processen i supervised learning, der anvendes i dette speciale. Modellen er inspireret af Kotsiantis (2007).

Figur 5: Supervised learning: Anvendelse i dette speciale.



Som det kan ses i figuren, anvendes en algoritme i konstruktionen af modellen. Der vil blive redegjort for forskellige typer af algoritmer i det følgende afsnit.

6.2.4. Forskellige typer af algoritmer

Der findes mange forskellige typer af algoritmer, og de kan ligeledes anvendes til at løse mange forskellige typer af problemer. Valget af algoritme afhænger af de tilgængelige data samt typen af problemstilling, der skal løses (Brownlee 2013). Inden for machine learning feltet refereres ofte til en klassisk artikel af Wolpert & Macready fra 1997, som diskuterer anvendelsen af forskellige algoritmer. Artiklens overskrift: "*No free lunch theorems for optimization*" (Wolpert & Macready 1997) er blevet et populært citat til at beskrive det faktum, at der ikke findes én bestemt algoritme, der er den universelt bedste til at løse enhver opgave (Kreatsoulas & Subramanian 2018; Domingos 2012; Molina & Garip 2019; Murphy 2012). Alle algoritmer har fordele og ulemper, og de antagelser, der er en fordel inden for ét domæne, kan være en ulempe inden for et andet (Murphy 2012:24).

Supervised machine learning involverer et behov for en algoritme, der kan forudsige et bestemt output ved ud fra et givent input (Molina & Garip 2019:1; Kotsiantis 2007:249; Murphy 2012:245). Det er det, som Pedro Domingos og Sotiris B. Kotsiantis betegner som en *læringsalgoritme* (Kotsiantis 2007; Domingos 2012). Domingos redegør i sin artikel for, at læringsalgoritmer er, når en specifik algoritme kombineres med et evaluerings- og optimeringsled (Domingos 2012). I dette speciale vil både logistisk regression og neuralt netværk blive evalueret med en accuracy test, som beskrevet tidligere. Hvordan optimeringsleddet anvendes, vil blive redegjort for under det enkelte afsnit om specialets undersøgelsesmetoder. De to modeller er som nævnt i afsnit "3.1. En problemformulering med et prædiktivt formål" valgt, fordi de repræsenterer henholdsvis en simpel og en kompleks udgave af prædiktive modeller, der har forskellige formål og metodiske fordele og ulemper i forhold til besvarelsen af specialets problemformulering. En mere uddybende diskussion af metodernes fordele og ulemper og disses betydning for undersøgelsens resultater vil forekomme i afsnit "6.4. Diskussion af de to undersøgelsesmetoder".

6.3. Specialets undersøgelsesmetoder

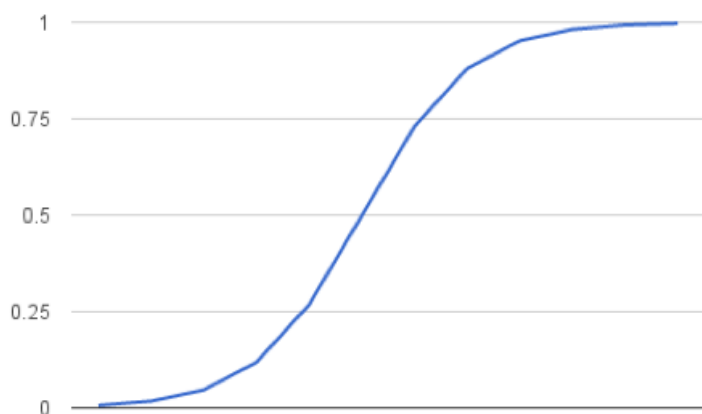
Dette afsnit vil indeholde en redegørelse for henholdsvis logistisk regressionsanalyse og neural netværksanalyse samt, hvordan de to undersøgelsesmetoder konkret er blevet anvendt i dette speciale. Under hver metode vil vi desuden redegøre for de optimeringsteknikker, der er anvendt. Slutteligt vil vi komme ind på nogle potentielle udfordringer ved anvendelsen af den enkelte metode, hvilket er risikoen for henholdsvis multikollinearitet og overfitting.

6.3.1. Logistisk regressionsanalyse

Når den afhængige variabel er binær, og det ønskes at udføre en regressionsanalyse, kan enten anvendes en lineær sandsynlighedsmodel eller logistisk regression. Valget af logistisk regression skyldes blandt andet, at det ved denne blot er muligt at opnå en sandsynlighed mellem 0 og 1, svarende til 100 procent sandsynlighed, mens der ved en lineær sandsynlighedsmodel også kan findes sandsynligheder under 0 og over 1. Ofte bruges termene *failure* og *succes* om de to værdier 0 og 1 på en binær afhængig variabel (Agresti & Finlay 2009:483f). I denne undersøgelse haves værdien 1, hvis der forekommer en eller flere somatiske kroniske sygdomme hos en person i fremtiden, og der er således tale om succes, hvis en person i løbet af den tre-årige periode får en somatisk kronisk sygdom, også selvom det ud fra en indholdsmæssig fortolkning betragtes som et negativt udfald. Overordnet set er fokus i logistisk regression på kausale sammenhænge (Hain & Jurowetzki 2018:12).

I logistisk regression anvendes Maximum Likelihood Estimation (MLE) til at beregne koefficienterne i en model for at finde den ligning, der passer bedst på det anvendte datasæt (Tuft 2000:24f). Koefficienterne beregnes ved hjælp af statistiske redskaber såsom gennemsnit og varians således, at der er størst sandsynlighed for at opnå de observerede outputværdier (Tuft 2000:24f; Navlani 2018). Sandsynligheden for succes estimeres på baggrund af en eller flere forklarende variable. Forholdet mellem afhængig og uafhængig variabel antager et S-formet kurvelineært forhold (Agresti & Finlay 2009:484). Det kurvelineære forhold i en logistisk regression er illustreret i figur 9:

Figur 6: Logistisk regression.



Kilde: Brownlee (2016a)

I en simpel model med én forklarende variabel er det kurvelineære forhold beskrevet med formelen:

$$\log \left[\frac{P(y=1)}{1-P(y=1)} \right] = \alpha + \beta x$$

hvor α er koefficienten for konstanten, β er parameterestimatet for en given uafhængig variabel x . $P(y = 1)$ angiver sandsynligheden for succes og $\left[\frac{P(y=1)}{1-P(y=1)} \right]$ er odds ratio, altså sandsynligheden for succes i forhold til sandsynligheden for fiasko (Agresti & Finlay 2009:484).

I dette speciale inddrages som nævnt i afsnit "5.2.2 Uafhængige variable" flere uafhængige variable og derfor anvendes multipel logistisk regression (Agresti & Finlay 2009:288). Formlen for en multipel logistisk regression er den samme som ovenstående, hvor der dog tilføjes ét yderligere parameterestimat for inddragelsen af hver forklarende variabel:

$$\log \left[\frac{P(y = 1)}{1 - P(y = 1)} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Kilde: Agresti & Finlay (2009:488)

Potentielle udfordringer i logistisk regressionsanalyse

Der kan opstå nogle udfordringer ved brugen af logistisk regression, når der haves et ønske om at undersøge et stort antal forklarende variable. Det skyldes, at det ikke er muligt at inddrage for mange forklarende variable, eftersom det kan medføre *overfitting*. Overfitting er en betegnelse, der bruges om kunstigt forhøjet forklaret varians. Begrebet beskriver det faktum, at det bør undgås at inddrage mange forklarende variable i en model, eftersom det øger sandsynligheden for, at variationer i den beregnede model skyldes, at modellen tager højde for tilfældig støj frem for faktiske variationer (Murphy 2012:22). Med andre ord vil det sige, at modellen tager højde for de særegne faktorer, der er gældende for det respektive sample, hvilket medfører, at modellens generaliserbarhed til andre populationer udfordres betragteligt (Molina & Garip 2019:3). Samtidig skal det siges, at ved anvendelse af multipel logistisk regression, såvel som andre multiple regressioner, er der risiko for multikollinearitet. Problemet indebærer, at der er en stærk lineær korrelation mellem to eller flere uafhængige variable, og risikoen for multikollinearitet bliver større, jo flere forklarende variable, der inddrages (Agresti & Finlay 2009:334). Hvis der er multikollinearitet i modellen, kan det medføre, at koefficienterne udregnes forkert (Bagheri et al. 2010:2086). Hvordan vi tager højde for overfitting og multikollinearitet vil blive forklaret i det følgende afsnit.

Logistisk regression anvendt som læringsalgoritme

Selvom vi inden for samfundsvidenskaben almindeligvis ikke bruger betegnelser som supervised machine learning, prædiktiv modellering eller algoritme i sammenhæng med logistisk regression, vil vi alligevel anvende metoden som en læringsalgoritme i dette speciale.

Når vi i dette speciale anvender logistisk regression som en læringsalgoritme, kræver det - som inden for al anden supervised machine learning - en inddeling af data i trænings-, validerings- og testdatasæt. Trænings- og valideringsdatasættet er brugt til at udarbejde modellen, mens test-datasættet er benyttet til at evaluere modellen ved hjælp af en AUC-score, som vi tidligere har redegjort for.

Når logistisk regression anvendes som en læringsalgoritme, kombineres den med optimeringsværktøjer til at forbedre modellens prædiktionsevne (Kreatsoulas &

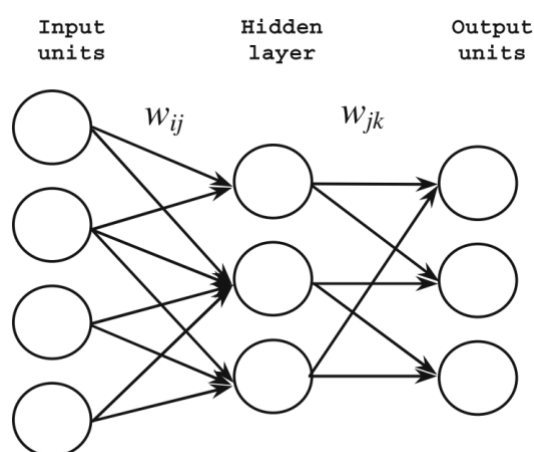
Subramanian 2018:348). Som optimeringsværktøj har vi i dette speciale anvendt en trinvis selektionsmetode, hvor formålet er at forbedre modellens præcision med inddragelse af så få variable som muligt og samtidig, at problemet med multikollinearitet undgås (Kim et al. 2015:1345f). En automatisk trinvis selektionsmetode er en almindelig måde at undgå dette problem inden for induktive analyser (Agresti & Finlay 2009:458). Ved multikollinearitet vil modellens forklaringskraft ikke øges betragteligt, ved inddragelse af en supplerende forklarende variabel, der korrelerer stærkt med øvrige variable i modellen. Det betyder ikke nødvendigvis, at variabelen ikke korrelerer med den afhængige variabel, men blot at inddragelse af variabelen ikke bidrager yderligere til modellens forklaringskraft (Agresti & Finlay 2009:334). Ud fra samme princip kan multikollinearitet undgås ved kun at inddrage få variable, der foreslås af den trinvise selektionsmetode, der har til formål at højne AUC-scoren mest muligt, idet AUC-scoren kun øges betragteligt ved inddragelse af variable, der ikke korrelerer med dem, der indgår i modellen i forvejen (Verbiest 2019). Som en sikkerhed for, at der ikke findes multikollinearitet mellem de inddragede forklarende variable, har vi desuden testet herfor og fundet, at der ikke eksisterer multikollinearitet i nogle af specialets logistiske regressionsmodeller. Det endelige antal forklarende variable i modellen vurderes ud fra det princip, at ved to modeller med samme AUC-score, men forskelligt antal forklarende variable, betragtes den model med færrest forklarende variable som den bedste model (ibid.). I dette speciale anvendes den trinvise selektionsmodel således, at metoden starter med at inddrage én variabel i modellen, og derefter beregner den, hvilken variabel der bør inddrages som den næste for at højne modellens AUC-score mest muligt. Denne proces gentages, indtil der ikke længere er nogen væsentlig forbedring i AUC-scoren. Hvilke forklarende variable, der har bidraget mest til prædiktionssevnen af den endelige logistiske regressionsmodel i dette speciale, vil blive gennemgået i analysen. Det skal afslutningsvist siges, at den logistiske regressionsmodel ikke kan anvendes til forudsige fremtidigt ukendt data, og denne model vil derfor ikke anvendes til at forudsige, hvem der får en somatisk kronisk sygdom i årene 2018-2010. Dette er imidlertid muligt med den neurale netværksmodel, som vi vil redegøre for i det følgende afsnit.

6.3.2. Neural netværksanalyse

Et neuralt netværk er en algoritme, der er inspireret af centralnervesystemet i den menneskelige hjerne på den måde, at den bruger et vægtningssystem, der minder om de synaptiske processer, der sker i hjernens nervebaner (Kreatsoulas & Subramanian 2018:348). Det overordnede fokus er at kunne udføre pålidelige prædiktioner (Hain & Jurowetzki 2018:13). Det er en algoritme, der er funderet i ikke-lineære regressionsteknikker, hvilket indebærer, at de kan håndtere problemstillinger, hvor klassificeringen ikke er lineær (Kuhn & Johnson 2013:141; Kotsiantis et al. 2006:169). Et neuralt netværk består af tre forskellige typer af lag af neuroner. Dette er input-neuroner, output-neuroner og neuroner derimellem. Input-neuronerne modtager den information, der skal behandles, mens output-neuronerne er der, hvor resultaterne af behandlingen findes. Neuronerne derimellem er skjulte enheder, og disse kan bestå af ét eller flere lag (Kotsiantis et al. 2006:169; Kuhn & Johnson 2013:144). Der er to centrale aspekter ved neurale netværk, der er afgørende for at forstå, hvordan modelleringen ved neurale netværk foregår. Dette er de enkeltstående neuroners funktion samt opbygningen af den sekventielle proces, der sker inden for neurale netværk. De enkelte neuroner henter oplysninger fra forrige lag, hvor det første lag er input-laget. Fra input-laget sker der altså en bevægelse mod output-laget gennem de skjulte lag. Denne proces fra input til output kaldes *forward propagation*. Alle variable i input-laget er fra starten tildelt en tilfældig vægtning, som gradvist justeres ved hjælp af den information, der genereres for hver gang sekvensen fra input til output gentages. Forskellen i den faktiske og den udregnede værdi af den afhængige variabel anvendes til at opdateres vægtene. Vægten er af markant betydning, da denne bestemmer styrken, hvormed signaler kommer frem i netværket (Hain & Jurowetzki 2018:23f).

Den simpleste form for opbygning af et neuralt netværk er med ét enkelt skjult lag, hvor signalerne blot sker i én retning fra input til output (Kuhn & Johnson 2013:144). Denne arkitektur kan ses i nedenstående figur:

Figur 7: Single-layer feedforward netværk.



Kilde: Kotsiantis et al. (2006:170)

Neurale netværk kan også indeholde loops, hvor signalerne sendes begge veje igennem lagene, og som tidligere nævnt kan der ligeledes indgå flere skjulte lag i et netværk alt efter, hvordan netværket konstrueres (Kuhn & Johnson 2013:144). Når der inddrages flere lag i det neurale netværk, er der tale om anvendelse af deep learning, der er en særlig form for machine learning (Brownlee 2016b). I følgende afsnit vil vi redegøre for, hvordan den neurale netværksmodel, der anvendes i dette speciale, er opbygget.

Modellens arkitektur

I dette afsnit vil vi nævne to justeringsparametre, der angår modellens arkitektur. Det første er antallet af skjulte lag, der anvendes i modellen. Et neuralt netværk uden skjulte lag – et såkaldt "single-layer network", bør kun anvendes ved meget simple, binære klassifikationsproblemer, hvor de to kategorier kan adskilles ved hjælp af en linje (Brownlee 2018a). Selvom vi i dette speciale har et binært klassifikationsproblem, vurderer vi det ikke som så simpelt, at de to grupper af henholdsvis personer med en somatisk kronisk sygdom og personer uden, kan adskilles ved en enkelt linje. Den bedste måde at vurdere antallet af lag, er en blanding mellem intuition og en eksperimenterende tilgang, hvor antallet af lag gradvist kan øges, hvis det ikke findes, at modellens prædiktionssevne er tilfredsstillende (ibid.). Den endelige model i dette speciale indeholder to skjulte lag.

Det andet justeringsparameter omfatter antallet af neuroner i hvert lag. Såvel som ved antallet af lag, vurderes antallet af neuroner bedst ud fra en blanding af intuition og en

fremgangsmåde, hvor tallene justeres på baggrund af ønsket om så høj prædiktionsværdi som muligt (Brownlee 2018a). Fordi vi har et ønske om en klassificering i output-laget, er det valgt, at der er 1 neuron i dette lag. De to skjulte lag i den endelige model i dette speciale indeholder henholdsvis 40 og 20 neuroner, mens input-laget indeholder 31 neuroner, svarende til antallet af forklarende variable.

Aktiveringsfunktioner

I en neural netværksmodel er aktiveringsfunktionen ansvarlig for at transformere de vægtede input fra den ene neuron til den næste. I dette speciale anvendes ikke-lineære aktiveringsfunktioner, da neuronerne hermed kan lære mere komplekse strukturer i data (Brownlee 2019). I dette speciale anvendes et neuralt netværk med fire lag - et input-lag, to skjulte lag og et output-lag, og dermed skal der inddrages aktiveringsfunktioner tre gange. I det sidste led vil anvendes Sigmoid, der er anvendelig, når det ønskes at opnå en sandsynlighed, hvilket skyldes, at denne kun kan tildele værdier mellem 0 og 1. Dermed at det muligt at finde sandsynligheden for at få en somatisk kronisk sygdom. Et problem med denne aktiveringsfunktion er dog, at værdier, der ellers ville være over 1 og under 0, alligevel tildeles disse værdier, hvilket øger mængden af fejl i netværk med mange lag. Dette betegnes også *the vanishing gradient problem* (Brownlee 2019). Dog betragtes dette problem ikke som af markant betydning, når funktionen blot anvendes i output-laget, fordi der ikke sendes signaler videre fra dette lag.

I de første to lag vil der i stedet anvendes ReLU (Rectified Linear Activation) som aktiveringsfunktion. Denne ligner og opfører sig som en lineær regression, selvom den ikke er det, og på den måde undgås *the vanishing gradient problem*. Dette skyldes, at vægtene forbliver proportionale i forhold til det forrige lag, og ReLU er både let at træne og medfører som regel høj prædiktionsværdi. Derfor er den efterhånden blevet standardfunktionen indenfor mange forskellige typer af neurale netværk (Brownlee 2019). ReLU anvendes ikke i output-laget, fordi denne funktion ikke egner sig til det sidste led i et binært klassifikationsproblem, hvor man er interesseret i et udfald, der angiver en sandsynlighed. I sådanne tilfælde bør ReLU kun bruges i de øvrige lag, mens Sigmoid bruges i output-laget (Walia 2017; Sharma 2017). Det er netop, hvad vi gør i dette speciale.

Optimeringsteknikker

En optimeringsfunktion anvendes til at opdatere vægtene i et neuralt netværk, og den har indflydelse på, hvor lang tid algoritmen tager om at opnå gode resultater i form af en høj prædiktionssevne. I dette speciale anvendes funktionen 'Adam' (Adaptive Moment Estimation), der er en såkaldt "stokastisk optimeringsalgoritme". Disse er gode at anvende, når der arbejdes med store datamængder, men der ikke haves så stor hukommelse på ens computer. Desuden er stokastiske optimeringsalgoritmer gode til data, der indeholder mange observationer med værdien 0 (Brownlee 2017b). Det er netop tilfældet i dette speciale, hvor andelen med en somatisk kronisk sygdom er lav, og således er andelen med værdien 0 på den afhængige variabel relativt høj. 'Adam' er en populær funktion i deep learning, fordi den opnår gode resultater på relativt kort tid, og virker godt sammenlignet med andre stokastiske optimeringsteknikker, idet den kombinerer fordelene ved forskellige andre metoder (Brownlee 2017b). Samtidig er det empirisk bevist, at 'Adam' effektivt løser deep learning problemer på store datasæt med mange inddragede variable (Kingma & Ba 2015:5). I en anden artikel af Sebastian Ruder, der undersøger forskellige metoder til optimering, anbefales 'Adam' i øvrigt som det bedste "*overall choice*" (Ruder 2017:10).

Ved inddragelse af optimeringsalgoritmer er der desuden to justeringsparametre, det er nødvendigt at forholde sig til. For det første antallet af iterationer, som maskinen skal foretage, før den kommer frem til det endelige resultat, hvilket også kaldes *epoker* (Brownlee 2018b). En redegørelse for antallet af iterationer i den endelige neurale netværksmodel i dette speciale, vil fremgå af afsnit "8.2. Analyse med neuralt netværk". Derudover skal en "batch-size" vælges, hvilket refererer til det antal observationer fra træningsdatasættet, der bruges i hver iteration, inden signalet sendes videre fra input-laget til det næste lag. Hvis der haves en stor datamængde, kræver det en høj kapacitet i computerens hukommelse, når modellen skal benytte information for samtlige observationer i hver iteration (Brownlee 2018b). Derfor anvendes ofte et såkaldt "mini-batch" i stedet, hvor maskinen i hver iteration anvender en selekteret del af observationerne. Det vil sige, at det totale antal observationer bliver inddelt i mindre grupper, hvorefter maskinen bruger én gruppe ad gangen til beregning og opdatering af vægtene. Det er en fordel at inddele i et antal grupper, der går lige op med det totale antal observationer, eftersom der ellers vil indgå færre observationer i den sidste iteration

(Brownlee 2018b). I dette speciale er den neurale netværksmodel konstrueret på datasættet fra år 2010, der består af 22.335 observationer. Disse observationer har vi inddelt i 1489 grupper, hvilket betyder at der indgår 15 observationer i hver gruppe, fordi $1489 \cdot 15 = 22.335$. Således har vi en batch-size på 15. Altså udregnes vægten for 15 observationer ad gangen, inden signalet sendes videre, hvilket betyder, at processen foretages 1489 gange for at komme igennem det fulde datasæt.

Udfordringer i neural netværksanalyse

I afsnit 6.3.1. redegjorde vi for to potentielle udfordringer, der kan opstå ved anvendelsen af logistisk regressionsanalyse, hvilket er risikoen for henholdsvis multikollinearitet og overfitting. I forhold til risikoen for multikollinearitet, beskriver flere studier, hvordan dette ikke betragtes som et problem i neurale netværksanalyser (de Veaux & Ungar 1994; Gou & Fyfe 2004; Dumancas & Bello 2015). Det skyldes strukturen og arkitekturen af de neurale netværk i forhold til, at de kan studere ikke-lineære sammenhænge og har tendens til at være overparameteriserede, hvilket medfører at "*neural networks tend to be fairly insensitive to problems of multicollinearity*" (de Veaux & Ungar 1994:2). At en model er overparameteriseret betyder, at der kan indgå flere parametre, end der er variable (Green et al. 1999:214). I neurale netværksmodeller kan antallet af neuroner i hvert lag eksempelvis være højere end antallet af input-variable, og dermed er de estimerede vægte ikke bundet op på den enkelte variabel. Det er samtidig grunden til, at det ikke er muligt at fastslå det kausale forhold mellem den afhængige og de forklarende variable i en sådan model (de Veaux & Ungar 1994:2). Dette vil blive uddybet yderligere i det kommende afsnit "6.4. Diskussion af de to undersøgelsesmetoder".

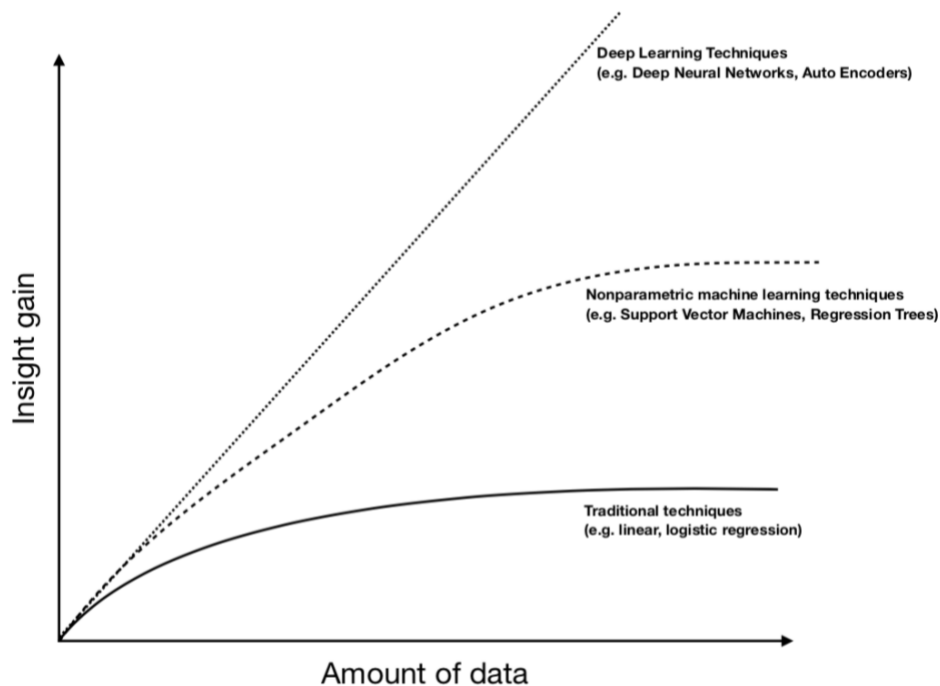
I neurale netværksanalyser er der en høj risiko for overfitting, hvilket kan betegnes som den største udfordring ved denne model (Domingos 2012:81; Kreatsoulas & Subramanian 2018:348). Det betyder som nævnt, at den konstruerede model tager højde for den særegne varians, der findes i træningsdatasættet, og dermed har en lav grad af generaliserbarhed i forhold til andre datasæt. Derfor kan vi se, om vores model overfitter, ved at der haves en lav AUC-score, når den evalueres og testes i validerings- og testdatasæt (Domingos 2012:81; Molina & Garip 2019:3). Vi vil i afsnittet "8.3. Diskussion af analyseresultater" inddrage, hvorvidt der kan være tale om overfitting i den neurale netværksmodel i dette speciale.

6.4. Diskussion af de to undersøgelsesmetoder

Som nævnt er formålet i logistisk regression at fastslå kausale sammenhænge, og metoden er populær på grund af den simple fortolkning, den lægger op til. Modellen er altså god, såfremt der haves et ønske om at fastslå, hvorvidt en bestemt forklarende variabel (X) har en statistisk signifikant effekt på sandsynligheden for et bestemt udfald på den afhængige variabel (Y) (Kuhn & Johnson 2013:286f). Logistisk regressionsanalyse betragtes således som særligt anvendelig til den indholdsmæssige fortolkning af de forklarende mekanismer, der er på spil i denne undersøgelse. Modsat kan neurale netværksmodeller betegnes som *black boxes* på grund af den måde, de er opbygget på. Black box-betegnelsen bruges om de skjulte lag, der er karakteristisk ved de mere komplekse machine learning modeller (Kreatsoulas & Subramanian 2018:348). Det betyder, at det er problematisk at gennemskue, hvorfor der er fremkommet en bestemt prædiktions fremfor en anden (Hain & Jurowetzki 2018:25). De fortolkningsmæssige vanskeligheder, som modellernes opbygning medfører, udgør den største svaghed for det neurale netværk (Kreatsoulas & Subramanian 2018:348). Sammenhængen mellem muligheden for at fortolke og at prædiktere forklarer Max Kuhn og Kjell Johnson således: *"While the primary interest of predictive modeling is to generate accurate predictions, a secondary interest may be to interpret the model and understand why it works. The unfortunate reality is that as we push towards higher accuracy, models become more complex and their interpretability becomes more difficult. This is almost always the trade-off we make when prediction accuracy is the primary goal."* (Kuhn & Johnson 2013:4)

I mange prædiktionsanalyser er selve forklaringen underordnet, hvilket eksempelvis er, når en mail enten skal eller ikke skal klassificeres som spam (Kuhn & Johnson 2013:4). Dog er der i dette speciale et ønske om at kunne påvise statistisk inferens og kunne forklare, hvorfor nogle variable er af særlig stor betydning, og derfor betragtes de fortolkningsmæssige vanskeligheder som en udfordring. Sammenlignet med en logistisk regressionsmodel er resultaterne af en neural netværksmodel måske nok vanskelige at fortolke, men modsat skal det huskes, at læringskapaciteten, i forhold til at kunne prædiktere et ukendt (fremtidigt) udfald, er stor i den neurale netværksmodel, mens denne mulighed er ikke-eksisterende i en logistisk regression (Hain & Jurowetzki 2018:2f). En yderligere forskel på de to metoder er den indsigt, de opnår ved anvendelse af store datamængder, og dette er illustreret i nedenstående figur:

Figur 8: Opnået indsigt som funktion af samplestørrelse for forskellige algoritmer.



Kilde: Hain & Jurowetski (2018:3)

Som det ses i figuren, er der en begrænsning på den indsigt, der kan opnås ved anvendelse af logistiske regressioner, når mængden af data øges, mens neurale netværk blot opnår en større indsigt des mere data, der haves. I og med der haves en stor mængde data i dette speciale, betragtes dette som en fordel for neurale netværk. En anden fordel ved det neurale netværk er, at der ikke er nogle antagelser om, hvilken form sammenhængen mellem den afhængige og de uafhængige variable skal antage. Derimod formodes det ved en logistisk regression som nævnt, at sammenhængen antager en S-formet kurve (Hastie et al. 2017:390; Agresti & Finlay 2009:484). Derfor anskues det, at det er nemmere at være tro mod data ved det neurale netværk.

Der altså både fordele og ulemper ved henholdsvis logistiske regressionsmodeller og neurale netværksmodeller. Det anskues, at det væsentligste er, hvorvidt metoden kan anvendes til at besvare første del af vores problemformulering, og denne er som nævnt: *“Hvilke sociale og psykologiske determinanter på individniveau har størst betydning for sandsynligheden for at udvikle en somatisk kronisk sygdom i fremtiden?”*. Hvilken af de to metoder, der findes bedst til at besvare denne del af problemformuleringen, vil vurderes ud fra to overordnede principper. Det første er, hvor god den enkelte model er til at give en præcis forudsigelse af somatiske kroniske sygdomme, vurderet på baggrund af AUC-

scoren. Det andet er, hvilken fortolkning af de kausale sammenhænge mellem vores uafhængige og afhængige variable den enkelte model giver anledning til.

6.5. Forudsigelse af somatiske kroniske sygdomme

Efter konstruktion og test af den neurale netværksmodel, anvendes modellen som nævnt til at forudsige, hvem af de personer, der indgår i datasættet fra 2017, der har størst sandsynlighed for at få en ny somatisk kronisk sygdom. Denne forudsigelse giver et output i form af en variabel for prædikeret sandsynlighed for hvert individ. Denne variabel er efterfølgende kodet som en binær variabel, hvor vi har inddelt samplet i to grupper. Den ene gruppe udgør de 10 procent, der har størst prædikeret sandsynlighed for at udvikle en somatisk kronisk sygdom, mens den anden gruppe udgør de resterende 90 procent af samplet. Vi har valgt denne fordeling for på den måde at isolere dem, der har størst sandsynlighed for at få en somatisk kronisk sygdom. Samtidig finder vi det anvendeligt, at det er en relativt lille andel, hvis resultaterne eksempelvis skal bruges til forebyggelse eller planlægning.

Det betragtes som nævnt, at den neurale netværksmodel i sig selv ikke kan bidrage til forståelsen af sammenhængen mellem sociale og psykologiske determinanter og mulig fremtidig sygdomsforekomst. Derfor vælges det at kombinere resultaterne fra den neurale netværksmodel med en multipel logistisk regression. På den måde udnyttes både det fulde potentiale af den høje generaliserbarhed i neurale netværksmodeller, der er mere præcise i sine forudsigelser af, hvilke personer der har størst sandsynlighed for at udvikle sygdom i fremtiden, og samtidig kan vi efterfølgende anvende denne forudsigelse som afhængig variabel i en logistisk regressionsanalyse til at undersøge statistisk inferens. Vi har inddraget samtlige af de uafhængige variable i en samlet model, hvor alle variable har en signifikant effekt på den prædikerede sandsynlighed for somatisk kronisk sygdom, med undtagelse af respondentens fysiske aktivitetsniveau i fritiden, i hvor høj grad de stoler på personer i deres lokalområde samt, hvor ofte der anvendes fedtstof i madlavningen. For at udvælge de mest relevante har vi set på odds ratio. Hvilke der er udvalgt, samt den indholdsmæssige fortolkning af denne analyse, vil fremgå i afsnit 8.2.3. og 8.2.4. Resultaterne af samtlige analyser er desuden vedlagt i bilag 7.

Derudover har vi ud fra samme fremgangsmåde ligeledes lavet forudsigelser og binær opdeling heraf på datasættet fra 2013. Disse forudsigelser er imidlertid ikke lavet med henblik på en logistisk regressionsanalyse, men de har til formål at indgå i besvarelsen af problemformuleringens sidste led, hvor den geografiske ulighed i forekomsten af nyopstået somatisk kronisk sygdom studeres. Her anses det som en fordel, at der kan laves to kort for 2013, da det dermed er muligt at sammenligne den geografiske fordeling af den faktiske sygdomsforekomst med den geografiske fordeling af den prædikterede sandsynlighed. En nærmere beskrivelse af fremgangsmåden for kortlægningen af somatiske kroniske sygdomme vil fremgå af det følgende afsnit.

6.6. Kortlægning af somatiske kroniske sygdomme

For at besvare den sidste del af dette speciales problemformulering: *“Hvordan vil forekomsten af disse fordele sig geografisk i Region Nordjylland?”*, vil variablene for både faktisk og prædikteret forekomst af disse sygdomme deskriptivt fremstilles på regionale kort afslutningsvist i specialets afsnit “9. Kortlægning af somatiske kroniske sygdomme”. I dette afsnit vil vi først redegøre for den områdeinddeling, der er anvendt, og derefter hvilken fremgangsmåde der i øvrigt er anvendt til kortlægningen af somatiske kroniske sygdomme.

6.6.1. Inddeling af områder

Til inddeling af områder i dette speciale anvendes Rolf Lyneborg Lunds konstruktion af områder i Region Nordjylland, og han skriver om denne inddeling, at: *“The main point is to create areas that have a simple logic in their creation and offer a much better model to locate microsocial enclaves in a wide variety of social measurements thus focusing on homogeneity”* (Lund 2018:13). Hans inddeling er udført ud fra den betragtning, at administrative grænser ikke er meningsfulde indikatorer for de typer af mennesker, der bor i et givent område og derfor ikke er optimale i forhold til studiet af geografisk segregering (Lund 2018). Helt konkret er områderne konstrueret ud fra en teoretisk antagelse om, at sociale områder formes ud fra fysiske barrierer såsom veje og jernbaner. Derfor har dette været første kriterium for dannelsen af områderne (Lund 2018:3f). Størstedelen af disse områder var dog for små til at leve op til de formelle krav om, at der

minimum skal være 100 husstande inden for hvert område, og derfor er områderne efterfølgende lagt sammen med udgangspunkt i, at 1) det er områder, der deler den procentvise største grænse, at 2) der sker færrest mulige sammenlægninger samt, at 3) der i de nye områder er færrest mulige indbyggere, hvis der er flere mulige sammenlægninger (Lund 2018:4). Denne områdekonstruktion er valgt i stedet for at tage udgangspunkt i en administrativ inddeling som eksempelvis kommuner eller sogne, fordi der i højere grad tages højde for, hvordan sociale områder konstrueres og er konstrueret, og derfor er beboerne i områderne mere homogene. Dette vises i artiklen for indkomst, uddannelse og etnicitet (Lund 2018). Et generelt diskussionspunkt i definitionen af lokalområder er, hvorvidt disse ansues som statiske eller emergerende størrelser. Det mest almindelige standpunkt er, at lokaliteter ikke er givne på forhånd men er socialt konstruerede gennem menneskelige processer af afgrænsningsdefinitioner (Savage et al. 2005:7). Da områder i dette speciale betragtes som dynamiske størrelser, der er foranderlige over tid, er der en risiko for, at områdeinddelingen ikke er gyldig på tværs af år. Eftersom Lunds inddeling af områder er foretaget på baggrund af data fra flere år, henholdsvis 2000, 2005, 2010 og 2015 (Lund 2018), vurderes det imidlertid, at disse områdeinddelinger har relativt høj gyldighed på tværs af år, og derfor også er mere valide i forhold til at gælde efterfølgende år, herunder også specialets undersøgelsesår. Derudover ses det som en fordel, at inddelingen er udført ud fra fysiske barrierer, da det antages, at disse er relativt stabile over tid.

6.6.2. Fremgangsmåde for kortlægning

Der er lavet tre kort i dette speciale, der hver er en deskriptiv fremstilling af andelen med en nyopstået somatisk kronisk sygdom i de førnævnte områder. De tre variable, der kortlægges på hvert sit kort, er som nævnt henholdsvis den faktiske andel med en nyopstået somatisk kronisk sygdom i årene 2014-2016 samt andelen med en prædikterede sandsynlighed i top 10-procent gruppen for samme år. Yderligere er det andelen med prædikteret sandsynlighed i top 10-procent gruppen for årene 2018-2020. Kortlægningen er foretaget ved hjælp af længde- og breddegradskoordinater for bopælen til de individer, der indgår i Sundhedsprofilen i enten 2013 eller 2017. De områder, hvor der indgik færre end tre individer, er udeladt af analysen på grund af datasikkerhed. For alle tre kort gælder det, at farvegradueringen er inddelt i følgende kategorier: 1) Til og

med 10 procent, 2) Mellem 11 og 15 procent, 3) Mellem 16 og 25 procent, 4) Mellem 26 og 30 procent og 5) Over 30 procent. Kortene præsenteres, beskrives og analyseres i afsnit "9. Kortlægning af somatiske kroniske sygdomme".

6.7. Kvalitetskriterier

I dette afsnit vil vi kommentere på kvalitetskriterierne ekstern validitet, intern validitet og reliabilitet i forhold til dette speciale.

6.7.1 Ekstern validitet

Begrebet ekstern validitet dækker over, i hvor høj grad en undersøgelses resultater er generaliserbare til andre populationer (de Vaus 2001:28). Formålet med supervised learning er at konstruere en algoritme, der har en høj gyldighed på fremtidige, uobserverede data, og vi anskuer derfor, at AUC-scoren er et mål for ekstern validitet. For at opnå en høj generaliserbarhed er det vigtigt, at det datasæt, der anvendes til at konstruere modellen, er repræsentativt (de Vaus 2001:29). Det skyldes, at en algoritme selvsagt kun kan konstrueres og trænes ud fra de input-data, den "fodres" med, og hvis disse data er biased, vil modellen også blive det (Kreatsoulas & Subramanian 2018:348). Vi vurderer ikke, at datasættet fra 2010, der anvendes til at konstruere modellerne, er biased, eftersom det er baseret på random sampling og har en høj svarprocent blandt alle grupper. Det er imidlertid et kendt problem, at det ikke er alle personer i en sådan sampling, der svarer på spørgeskemaet, og selvom samplingen er repræsentativ, kan manglende svar fra bestemte typer af respondenter medføre såkaldt *non-response bias* (de Vaus 2014:81). Som vi har set i afsnit "5.1.1. Sundhedsprofilerne fra 2010, 2013 og 2017" er der også en forholdsvis stor andel af det udvalgte sample, der ikke har besvaret de tre respektive spørgeskemaundersøgelser under Sundhedsprofilen. Til at korrigere for denne fejl er der udviklet vægte, der kan inddrages i undersøgelser foretaget på Sundhedsprofilernes data. Disse vægte har det dog ikke været muligt at benytte i dette speciale, eftersom vi har fundet, at neurale netværksmodeller ikke kan håndtere disse på samme vis, som de er tiltænkt i eksempelvis regressionsanalyser. I stedet indkluderer den neurale netværksmodel blot vægtene på lige fod med de forklarende variable. Vægtene er ligeledes ikke anvendt i de logistiske regressionsmodeller, eftersom det

vurderes, at det ville være et problem i forhold til sammenligneligheden på tværs af de to metoder.

Derudover har samplestørrelsen også betydning for den eksterne validitet (de Vaus 2001:78). I dette speciale betragtes det ikke som en udfordring, eftersom der anvendes et datasæt med et relativt stort antal observationer. Derimod anser vi det netop som en fordel, at det fulde potentiale af den store datamængde i højere grad udnyttes til generalisering i en neural netværksmodel i forhold til i en logistisk regressionsmodel (Hain & Jurowetski 2018:3). Graden af generaliserbarhed, der er opnået ved anvendelse af de to metoder i dette speciale, vil blive diskuteret yderligere i afsnit "8. Analyse af determinanternes betydning".

6.7.2. Intern validitet

Intern validitet refererer til, hvorvidt undersøgelsesdesignet giver anledning til at bekræfte de kausale konklusioner, det har til hensigt (de Vaus 2001:27). I dette speciale ønsker vi som nævnt at påvise en effekt af en række sociale og psykologiske determinanter i forhold til forekomsten af somatisk kronisk sygdom. For at kunne omtale en sammenhæng som kausal skal de følgende tre kriterier være opfyldte: 1) Der er en sammenhæng mellem variablene, 2) Der er en passende tidsorden, og 3) Alternative forklaringer udelukkes (Agresti & Finlay 2014:302).

At der er en statistisk sammenhæng mellem variablene vurderes i den logistiske regressionsmodel ud fra signifikansniveauet for de variable, der indgår i den endelige model efter den trinvis variabelselektion. Resultaterne heraf vil fremgå af afsnit "8.1. Analyse med logistisk regression". Som vi har nævnt, er det netop udfordringen ved den neurale netværksmodel, at den ikke tilbyder signifikansniveauer for de inddragede variable, og dermed ikke lægger op til en tolkning af de kausale sammenhænge. Derfor er den interne validitet udfordret i en neural netværksmodel. Idet vi med anvendelsen af en neural netværksmodel ikke er i stand til at påvise statistisk inferens, har vi valgt at supplere resultaterne af denne analyse med en logistisk regressionsanalyse, hvor det undersøges, hvilke af de forklarende variable der har betydning for udfaldet på den prædikterede sandsynlighed for at få en somatisk kronisk sygdom i årene 2018-2020. En

uddybende forklaring af fremgangsmåden herfor vil blive beskrevet i afsnit “6.5. Forudsigelse af somatiske kroniske sygdomme”. Som vi har nævnt i afsnit “5.2.1. Afhængige variable: Nyopstået somatisk kronisk sygdom” er den interne validitet i dette speciale ligeledes udfordret af, at de afhængige variable består af samlede mål for flere forskellige sygdomme, hvor de sociale og psykologiske determinanter kan tænkes at have forskellige effekter på de enkelte sygdomme. De påviste effekter i dette speciale kan altså være biased af, at forskellige effekter kan udligne hinanden, men dertil kan effekterne også være svære at give en konsistent forklaring på.

I forhold til punkt 2) indebærer det tidsmæssige element, at de uafhængige variable kommer før den afhængige (Agresti & Finlay 2014:302). For at kunne fastslå den kausale sammenhæng er det derudover nødvendigt, at sammenhængen blot går den ene vej, og det ikke skyldes det omvendte forhold, hvilket ville være, at forekomsten af somatisk kronisk sygdom har betydning for de sociale og psykologiske determinanter. På grund af det longitudinelle design, der anvendes i dette speciale, er der imidlertid ingen fare for en kausal fejlslutning på baggrund af et eventuelt omvendt kausalforhold.

Angående alternative forklaringer i punkt 3) indebærer dette, at der ikke er andre mulige forklarende variable, der er den afgørende faktor for ændringen af den afhængige variabel i stedet for de inddragede uafhængige variabel. I dette speciale betragtes den sociale virkelighed som kompleks, og blandt andet bestående af flere underliggende, uobserverbare mekanismer (jf. afsnit “3.2. Den videnskabelige placering”), og derfor kan kriteriet om alternative forklaringer være svært at opfylde. Der er altid en risiko for, at udfaldet på den afhængige variabel kan skyldes andre faktorer, som det ikke har været muligt at inddrage i modellen. Begrebet *omitted variable bias* beskriver netop dette forhold (Treiman 2009:364). Fordi det ikke er muligt at tage højde for samtlige af de faktorer, der påvirker den afhængige variabel, kan et kausalforhold aldrig endeligt bevises (Agresti & Finlay 2009:303). Dog vil vi alligevel tilstræbe at komme så tæt som muligt på at fastslå et kausalforhold, ved at inddrage kontrolvariable i den logistiske regressionsmodel. På denne måde er det muligt at se effekten af den enkelte variabel, når effekten af de resterende uafhængige variable holdes konstante (de Vaus 2001:104). Fremgangsmåden for inddragelsen af kontrolvariable og deres indflydelse på

resultaterne i den endelige regressionsmodel præsenteres i analyseafsnittet. Hvordan den interne validitet vurderes, vil indgå i konklusionen.

6.7.3. Reliabilitet

Reliabilitet (...) *defineres som den grad, hvormed man ved gentagne målinger af samme fænomen får samme resultat.*" (Andersen 2012:102). Det handler dermed om konsistens. I en artikel i tidsskriftet Science skriver Matthew Hutson: "*The booming field of artificial intelligence (AI) is grappling with a replication crisis*" (Hutson 2018:725). I artiklen fremhæves det både som et problem, hvis koden, der er anvendt til konstruktion af AI-modeller, ikke er offentligt tilgængelig, men også at koden i sig selv ikke er nok. Der må også have en argumentation for til- og fravalg og en udførlig beskrivelse af fremgangsmåden for at sikre gennemsigtighed så andre har mulighed for at reproducere resultaterne (ibid.). I forbindelse med reliabilitet er der i dette speciale fokus på at opnå gennemsigtighed med både datamanagement og analyse. Yderligere skal det siges, at hvis den beregnede model fremadrettet skal kunne forudsige somatiske kroniske sygdomme kræver det, at der stilles de samme spørgsmål i spørgeskemaet. Herudfra vurderes det, at der er en høj grad af reliabilitet i dette speciale.

7. Teori om determinanter

I det følgende afsnit vil vi redegøre for de sociologiske teorier, der anvendes til fortolkning af de sociale og psykologiske determinanter på individniveau, der har været udslagsgivende i specialets første analysedel. Som vi har nævnt indledningsvist i dette speciale, ønsker vi at belyse problemstillingen med et fokus på ulighed i sundhed. I dette speciale er teoriernes rolle som nævnt under afsnit "3.2. Den videnskabelige placering", at de skal være behjælpelige til at kunne give en afdækket forklaring af, hvad der har betydning for, hvem der har en særlig stor risiko for at få en somatisk kronisk sygdom. I analysen fremstår ikke alle temaer som betydningsfulde i forhold til at kunne forudsige, hvem der har størst sandsynlighed for dette, og de teoretiske perspektiver er derfor udvalgt på baggrund af fund i analysen. Vi har fundet, at temaer af betydning er socioøkonomiske forhold, sundhedsrelateret adfærd, mental trivsel og tillid, mens temaerne omhandlende socialt samvær og lokalområde ikke har vist sig betydningsfulde.

I dette speciale har vi valgt at anvende Pierre Bourdieus begrebsapparat til at indfange de forskellige forklaringsmekanismer, der er på spil i forhold til sammenhængen mellem somatiske kroniske sygdomme og socioøkonomiske faktorer samt sundhedsrelateret adfærd. Dette vil beskrives i afsnit 7.1, og i afsnittet vil Simon J. Williams applicering af Bourdieus teori på sundhedsfeltet ligeledes inddrages. Til at forklare sammenhængen mellem forekomsten af somatiske kroniske sygdomme og henholdsvis mental trivsel og tillid, anvendes i afsnit 7.2. en uddybende beskrivelse af den psykosociale forklaringsmodel formuleret af Jon Elstad. I sidstnævnte afsnit vil de konkrete spørgsmål fra Sundhedsprofilen nævnes og kobles på det teoretiske perspektiv, da det i højere grad findes nødvendigt at argumentere for, hvorfor disse spørgsmål kan forstås under rammen af den psykosociale forklaringsmodel. De teoretiske redegørelser vil løbende blive underbygget med eksempler fra eksisterende forskning.

7.1. Socioøkonomiske faktorer og sundhedsrelateret adfærd

I analysen fremkommer det, at flere livsstilmæssige og socioøkonomiske faktorer har betydning for risikoen for at få en somatisk kronisk sygdom. Til at forklare denne sammenhæng vil Pierre Bourdieus teoretiske perspektiv inddrages. Herunder vil vi

redegøre for hans habitusbegreb, udvalgte kapitalformer samt symbolsk vold. Ligeledes vil Simon J. Williams applicering af Bourdieus teoretiske perspektiver på sundhedsfeltet inddrages. De teoretiske redegørelser vil desuden løbende blive sammenholdt med de sygdoms- og sundhedssociologiske forklaringsmodeller til ulighed i sundhed.

Bourdieu betragtes ofte som en klasseteoretiker, idet social klasse udgør en fundamental analytisk kategori i meget af hans forskning og teoretiske begrebsapparat (Wright 2005:82f). Hans teorier er funderet i en opdeling mellem *et rum af sociale positioner* og *et rum af sociale positioneringer* (Bourdieu 1997:21ff). I rummet af sociale positioner er individer eller grupper relationelt placeret med udgangspunkt i to dimensioner, hvilket er 1) deres mængde af henholdsvis økonomisk og kulturel kapital og samtidig 2) den samlede volumen af de to kapitalformer (Bourdieu 1997:21). Det overordnede kapitalbegreb defineres som "*the set of actually usable resources and powers*" (Bourdieu 1984:114). Økonomisk kapital er et udtryk for de materielle ressourcer, individet har tilgængeligt, som ifølge Bourdieu "(...) *immediately and directly convertible into money and may be institutionalized in the forms of property rights*" (Bourdieu 1986:243). Kulturel kapital eksisterer i både en kropsliggjort form, en objektiveret form samt en institutionaliseret form. Den kropsliggjorte form angår langvarige dispositioner i krop og sind, mens den objektiverede form indebærer kulturelle ejendele såsom bøger, instrumenter og så videre. Den institutionaliserede form er eksempelvis uddannelse, officielle titler eller lignende (ibid.).

Afstanden mellem individerne eller gruppernes placering i rummet af sociale positioner sidestilles med faktiske sociale afstande, hvor kapitalvolumen er den mest afgørende dimension (Bourdieu 1997:21). Rummet af sociale positioneringer indebærer individers eller grupperes livsstil og adfærd, og dette rum svarer til rummet af sociale positioner (Bourdieu 1997:23). Hvis det skal oversættes til dette speciales kontekst vil det sige, at individets eller gruppens sundhedsadfærd er tilsvarende den sociale position i samfundet. Derudfra kan det siges, at Bourdieus perspektiv på sammenhængen mellem social position og livsstil kan relateres til den kulturelle forklaring på ulighed i sundhed, da denne netop kobler social position med sundhedsrelateret adfærd (Nettleton 2013:162).

Rummet af sociale positioner bindes sammen med rummet af sociale positioneringer gennem habitus. Dette begreb og dets funktion som mellemlid beskriver Bourdieu således:

“Habitus er dette generative og samlende princip, der omsætter de indre og relationelle karakteristika i en position til en samlet livsstil, dvs. til en samlet sæt af personvalg, valg af goder, valg af praktikker.” (Bourdieu 1997:24)

Habitus er forskelliggjort på baggrund af de sociale positioner, og dermed indgår der et passivt differentieringselement heri. Samtidig indebærer habitus en aktiv form for differentiering, da habitus også er forskelliggørende, hvilket medfører forskellige adfærdsmønstre og livstile blandt individer og grupper (Bourdieu 1997:24). Det bemærkes, at habitus er kropsligt funderede erfaringer, der er med til at forme individets handlinger (Bourdieu 1984:166; Bourdieu 2000:130). Eftersom individets handlinger er funderet i dets habitus, forstår Bourdieu altså menneskelige handlinger som et produkt af forskellige (kropslige) disponeringer, frem for handlinger genereret af rationel kalkulerende (Bourdieu 1997:24+227; Bourdieu 1984:166; Bourdieu 2000:130). Derudfra kan habitusbegrebet ligeledes kobles til livsforløbet som forklaringsmodel, hvor det fremlægges, at der igennem et menneskeliv ophober sig en akkumulation af positive og negative eksponeringer forstået som summen af samtlige betydelige faktorer, et individ bliver eksponeret for i løbet af sit liv (Nettleton 2013:165). Habitus kan ses som den bagvedliggende faktor, der påvirker denne akkumulation. Fordele og ulemper kan eksempelvis være rygning eller motion.

Disse betragtninger om habitus og handlinger har sociologen Simon J. Williams appliceret på sundhedsfeltet med den hensigt at give en teoretisk forklaring på sammenhængen mellem klasse, sundhed og livsstil (Williams 1995:581). Fordi handlinger i Bourdieus optik er rutineprægede og ubevidste, kan sundhedsrelateret adfærd ifølge Williams ligeledes *“be seen as a largely routinised feature of everyday life which is guided by a practical or implicit logic”* (Williams 1995:583). Derfor er sundhedsrelateret adfærd altså ej heller et bevidst valg, men det skal ses som en manifestation af både klassetilhørsforhold og habitus.

Ifølge Bourdieu udgør uddannelsesniveau og beskæftigelsesposition de primære faktorer, der influerer individets position i det sociale rum, men derudover har øvrige socioøkonomiske faktorer såsom køn, alder og etnicitet også en betydning. Individets adfærd er desuden påvirket af sådanne faktorer, eftersom de er med til at forme habitus (Bourdieu 1984:107ff). Som eksempel taler Bourdieu og Loïc Wacquant om, hvordan habitus er "kønnet". Det skyldes, at de normer og opfattelser, der knytter sig til køn, er et produkt af både sociale og kognitive strukturer. Køn skal altså forstås som en samfundsmæssig institution, der gennem mange årtiers sociale processer er indlejret i de sociale strukturer på bestemt vis. De kognitive strukturer refererer til de kønsopfattelser og normer, der er indlejret i individets krop og sind (Bourdieu & Wacquant 1992:171).

Sociale positioner er ikke blot forskellige, men er indlejrede i de magtstrukturer, der findes i samfundet. I relation hertil bruger Bourdieu begrebet *symbolsk vold*. Begrebet indebærer, at de dominerendes smag og adfærd opfattes som den "rigtige" af både de dominerende og dominerede grupper i samfundet, omend dette magtforhold miskendes (Bourdieu & Passeron 2006:9+24). Den symbolske vold er udtryk for en klassekamp, hvor særligt den dominerende gruppe tager afstand til de lavere stillede grupper og forsøger at definere disse gruppers adfærd som primitiv (Williams 1995:588f). De dominerende klasser har gennem deres kapitaler flere muligheder for at vælge en smag og adfærd, som ikke er tilgængeligt for de lavere klasser, hvilket netop er udtryk for det magtforhold, som begrebet symbolsk vold beskriver. Fælles for alle grupper er, at de udvikler en smag for det, der er tilgængeligt for dem (Williams 1995:590). Mens Bourdieu oftest selv bruger begreberne om økonomisk og kulturel kapital som de elementer, igennem hvilke disse magtforhold kommer til udtryk, demonstrerer Williams, hvordan symbolsk vold ligeledes kan forstås i en sundhedsmæssig kontekst (Williams 1995:588f). Hvis begrebet om symbolsk vold kobles sammen med KRAM-faktorerne og den generelle forståelse af, hvad der anses som en god sundhedsrelateret adfærd, kan det tolkes således, at det ofte er personer, der tilhører dominerende grupper i samfundet, der udvikler og tilskynder KRAM-faktorerne og lignende fænomeners betydning. Denne tolkning er baseret på en antagelse om, at viden om sundhed og god sundhedsrelateret adfærd er funderet i forskning foretaget af eksempelvis læger og andre fagprofessioner med relevans for det sundhedsvidenskabelige felt, der ofte tilhører netop denne samfundsgruppe. Med udgangspunkt i den socialkonstruktivistiske betragtning, hvor

viden anses som værende socialt konstrueret (Collin 2015:326), kan det siges, at KRAM-faktorerne som videnskabeligt produkt er social konstrueret. I forlængelse af begrebet symbolsk vold og fagprofessionernes rolle i udviklingen af sundhedsmål findes artefakt forklaringen på social ulighed i sundhed relevant at inddrage. Som nævnt er der i denne forklaring fokus på de sociale processer i målinger og herunder, at målinger af sundhed netop er socialt konstruerede (Nettleton 2013:161). Herudfra vil vi have en kritisk tilgang til de resultater, som analyserne viser på den måde, at de vil ikke blot anses som sande forskelle.

Der indgår ikke forklarende variable, der kan studeres som symbolsk vold, som det er tilfældet med Bourdieus begreber om livsstil og social position, der i højere grad direkte kan forstås som sundhedsrelateret adfærd og socioøkonomiske faktorer. I stedet vil begrebet om symbolsk vold indgå som teoretisk begrundelse for nogle af de fund, der gøres i analysen i forhold til både de socioøkonomiske faktorer og variablene om sundhedsrelateret adfærd.

7.2. Mental trivsel og tillid

Variablene under temaet om mental trivsel omhandlende respondentens oplevelse af depression og angst, deres energiniveau samt hvor ofte de føler sig rolige og afslappede, har givet udslag i analyserne. Det samme har variabelen omhandlende forsigtighed med andre mennesker, der indgår under temaet om tillid. Deres sammenhæng til somatisk kronisk sygdom mener vi, kan forklares ud fra den psykosociale forklaringsmodel. Derfor vil der i det følgende gives en uddybende redegørelse for denne, på baggrund af Jon Elstads beskrivelse, og der vil samtidig argumenteres for sammenhængen til variablene.

I den psykosociale forklaringsmodel anskues ulighed i sundhed som et produkt af social interaktion, og således er sociale miljøer og sociale relationer det primære fokus. Mennesker interagerer med deres omgivelser, og resultatet er psykologiske reaktioner. Inden for den psykosociale forklaring anskues det, at individets mentale reaktion på dets livsbegivenheder og sociale omgivelser er den største determinant for dets sundhedsstatus i den moderne vestlig verden. Følelsesmæssige reaktioner kan medføre depression og angst hos individet, men derudover kan det også medføre somatiske

reaktioner som eksempelvis forhøjet blodtryk, hjertebanken samt en række somatiske sygdomme. Dertil kommer desuden, at der kan ses en indirekte sammenhæng mellem følelsesmæssige og fysiske reaktioner forstået på den måde, at mentale reaktioner kan påvirke individets sundhedsrelaterede adfærd, som har en indflydelse på den overordnede sundhedsstatus (Elstad 2000:69ff). Som nævnt i problemfeltet har flere undersøgelser været i stand til at påvise denne sammenhæng mellem følelsesmæssige reaktioner og somatiske kroniske sygdomme (jf. problemfeltet).

Ifølge den psykosociale forklaringsmodel kan denne sammenhæng forklares med, at menneskekroppen består af flere forskellige dynamiske kropssystemer, som alle påvirkes af nervesystemet, og på den måde påvirkes disse kropslige systemer altså også af den aktivitet, der sker mentalt i form af følelsesmæssige reaktioner. Disse reaktioner i kroppens systemer som følge af ydre påvirkninger kaldes stressrespons, og de ydre påvirkninger, der sætter gang i denne stressrespons, kaldes eksterne stressorer. Det skal bemærkes, at denne stressrespons ikke er en usund eller farlig, men er en nødvendig funktion for at menneskekroppen fungerer korrekt under visse omstændigheder. Problemer opstår imidlertid, hvis stressreaktionerne ikke stoppes (Elstad 2000:72f).

Spørgsmålene omhandlende hvorvidt respondenterne inden for de sidste fire uger har følt sig henholdsvis fulde af energi samt rolige og afslappede betragtes at relatere sig til den psykosociale forklaring på den måde, at hvis respondenterne oplever at være fulde af energi samt rolige og afslappede, anskuer vi det således, at stress enten ikke er til stede eller i lav grad er til stede. Dermed anser vi, at spørgsmålene er et udtryk for et omvendt proportionalitet forhold til respondenternes stressniveau.

I den psykosociale forklaringsmodel anskues ulighed i sundhed ligeledes ud fra et klasseperspektiv, og der haves forskellige forklaringer på, hvorfor lavere sociale klasser i højere grad oplever stress (Elstad 2000:74). Elstad omtaler sårbarhedsteser, hvor det anskues, at lavere sociale positioner har færre coping ressourcer til at håndtere eksterne stressorer, og derfor er nogle personer mere sårbare end andre (Elstad 2000:74f). En anden forklaring inden for den psykosociale forklaring bygger på, at lavere sociale positioner udsættes for en større mængde eksterne stressorer (Elstad 2000:75f). En tredje forklaring handler om, at selve relationen mellem forskellige sociale positioner i

det sociale hierarki skaber stress, hvor der er fokus på, at der ikke blot er forskel mellem top og bund i samfundet, men der findes en social gradient angående sundhed. Herunder betragtes det, at der i lavere sociale positioner haves mindre kontrol over egen skæbne. Samtidig kan individer i lavere sociale positioner føle sig mindre værd relativt til personer i højere sociale positioner, og relationen mellem sociale positioner kan yderligere medføre diskrimination og stigmatisering. Alle disse faktorer kan have betydning for fordelingen af stress i samfundet (Elstad 2000:76ff). I den sidste forklaring inden for den psykosociale forklaring på, hvorfor der er ulighed i sundhed, står social kapital og social sammenhængskraft centralt. Der er fokus på, at i velstående lande har den materielle ulighed større betydning for ulighed i sundhed end absolut deprivation, forstået som materielt afsavn, og denne ulighed kan variere på tværs af miljøer, regioner og lande. Dette har både de tidligere nævnte konsekvenser forbundet med det sociale hierarki, men det har samtidig også en negativ betydning for social integration og social sammenhængskraft (Elstad 2000:78f). Elstad beskriver det således:

“Inequality (...) fosters lower levels of trust and higher levels of hostility, widespread frustration, and feelings of inferiority and lack of dignity.” (Elstad 2000:79)

Det er selvsagt, at der i højere grad er stress i sådanne samfund, hvor det dog særligt rammer individer i lavere sociale positioner, fordi der forekommer en synergieffekt ved, at der både er en lav sammenhængskraft og samtidig også frustration over den relative deprivation (Elstad 2000:79). Det er denne forklaring på, hvorfor der er ulighed i sundhed, vi finder relevant i forhold til variabelen, der indfanger, hvorvidt respondenterne mener, man ikke kan være forsigtig nok, når man har med andre mennesker at gøre, da det betragtes, at spørgsmålet indfanger graden af tillid til andre mennesker. Med udgangspunkt i denne del af den psykosociale forklarings tankegang forklarer Richard Wilkinson og Kate Pickett (2010), at tillid har betydning for individers levetid (Wilkinson & Pickett 2010:57). Ligeledes har Ichiro Kawachi et al. (1997) fremvist, at der er sammenhæng mellem økonomisk ulighed og mistillid, og samtidig mellem mistillid og dødelighed (Kawachi et al. 1997:1494f). Derudfra betragtes altså en sammenhæng mellem tillidsniveau og sundhed. Yderligere kan det ud fra denne forklaring anskues således, at hvis der i forskellige områder i Danmark er forskelligt niveau af materiel

ulighed, kan dette tænkes at have betydning for tillidsniveau og derigennem for sundheden blandt borgerne i disse områder.

8. Analyse af determinanternes betydning

Dette analyseafsnit er inddelt i to særskilte afsnit. Det første afsnit indeholder den del af analysen, der er foretaget ved hjælp af logistisk regression, mens det andet afsnit indeholder analyse foretaget ved hjælp af neuralt netværk. Som nævnt vil der dog tilføjes en logistisk regressionsanalyse til den neurale netværksanalyse for, at vi kan undersøge sociale og psykologiske determinanternes betydning for nyopstået somatisk kronisk sygdom. Den første del af analysen med logistisk regression vil primært omhandle en analyse af modellen, der konstrueres på baggrund af datasættet fra år 2010, mens den anden del med neural netværksanalyse suppleret med logistisk regression primært vil fokusere på analyse af de prædikterede sandsynligheder for nyopstået sygdom i 2018-2020 på baggrund af datasættet fra 2017. Disse forudsigelser er baseret på den neurale netværksmodel, der ligeledes er konstrueret ved hjælp af data fra år 2010. Modellen konstrueret med logistisk regression giver ikke mulighed for at forudsige på datasættet fra 2017. Derfor vil resultaterne fra denne model i stedet anvendes til at fastslå, hvilke determinanter fra 2010 der har haft betydning for konstruktionen af modellen, og som vi derudfra også formoder vil have betydning for sandsynligheden for at udvikle en somatisk kronisk sygdom i fremtiden. Når der i analysen refereres til forskelle mellem grupper, henvises der til de forskelle, der findes i modellerne, og som tidligere nævnt fastslår vi ikke, at disse er udtryk for sande forskelle. I stedet betragter vi det således, at disse resultater angiver, at det tyder på, at der er forskelle herimellem. Når vi henviser til data fra årene 2010, 2013 og 2017, inkluderer det også de afhængige mål for somatisk kronisk sygdom i henholdsvis årene 2011-2013, 2014-2016 og 2018-2020, hvilket også blev nævnt i afsnit "5. Beskrivelse af data".

8.1. Analyse med logistisk regression

Analysen med logistisk regression er inddelt i tre underafsnit. I afsnit 8.1.1. gennemgås fire steps, der er anvendt i konstruktionen af modellen. Dernæst redegøres i afsnit 8.1.2. for resultaterne af den test, der er foretaget ved hjælp af testdatasættet fra 2013. I afsnit 8.1.3 præsenteres den endelige logistiske regressionsmodel med signifikanstest og odds ratio, der i sidste afsnit 8.1.4 anvendes til at give en teoretisk fortolkning af resultaterne.

8.1.1. Konstruktion af model

Konstruktionen af modellen er som nævnt foretaget på baggrund af fire steps, som vil blive gennemgået i dette afsnit.

Første step: Automatisk variabelseleksion

Det første step blev foretaget på det fulde datasæt fra 2010. Det vil sige inden opdelingen i trænings- og valideringsdatasæt, som blev udført i step 2. Som vi har redegjort for i metodeafsnittet, blev de forklarende variable i selektionsmetoden inddraget trinvist ud fra en beregning af, hvilken variabel der højnede AUC-scoren mest muligt. Vi har valgt at stoppe efter inddragelse af 15 variable, eftersom vi ikke vurderer, at sandsynligheden for, at der bør inddrages flere variable end dette i den endelige model, er til stede. De 15 variable, der er udvalgt i en automatisk selektion for modellen er følgende (i nævnt rækkefølge): alder, BMI, hvor deprimeret eller ængstelig respondenteren føler sig, køn, rygning, hvor ofte respondenteren har kontakt til sine naboer, hvor ofte respondenteren drikker alkohol, respondenterens overordnede livstilfredshed, hvor ofte han eller hun føler sig fuld af energi, følelsen af tilknytning til lokalområdet, hvor ofte respondenteren føler sig rolig og afslappet, hvorvidt respondenteren har en at tale med ved behov, etnicitet, samlet score på den validerede PSS-skala (Perceived Stress Scale) samt, hvor ofte respondenteren spiser fedtstof på sit brød. En detaljeret beskrivelse af variablerne findes i bilag 2.

Andet step: Opdeling af datasættet fra 2010

I andet step inddelte vi datasættet fra 2010 i henholdsvis et trænings- og et valideringsdatasæt. Træningsdatasættet udgør 80 procent, mens valideringsdatasættet udgør 20 procent af det samlede datasæt fra 2010. Inddelingen er foretaget tilfældigt. Dog har vi valgt, at andelen, der får en somatisk kronisk sygdom ($Y = 1$), skal være lige stor i de to datasæt. Vi testede om dette kriterium var opfyldt, inden vi gik videre. Andelen var 0,07 procent i både trænings- og valideringsdatasættet, hvilket også svarer til andelen i det samlede datasæt for 2010.

Det er i denne forbindelse vigtigt at nævne, at eftersom datasættet opdeles tilfældigt, vil denne opdeling også være forskellig for hver gang, analysen foretages. Dette har indvirkning på både modellens samlede prædiktionssevne og AUC-værdierne ved inddragelse af de enkelte variable. Vi har derfor foretaget testen fem gange for at sikre os,

at resultaterne er sammenlignelige hver gang, hvilket de er. I resten af afsnittet vil beskrivelserne tage udgangspunkt i den analyse, der blev udført på baggrund af den sidste af de fem inddelinger, der blev foretaget.

Tredje step: Anvendelse af funktionen for beregning af AUC-score efter opdeling

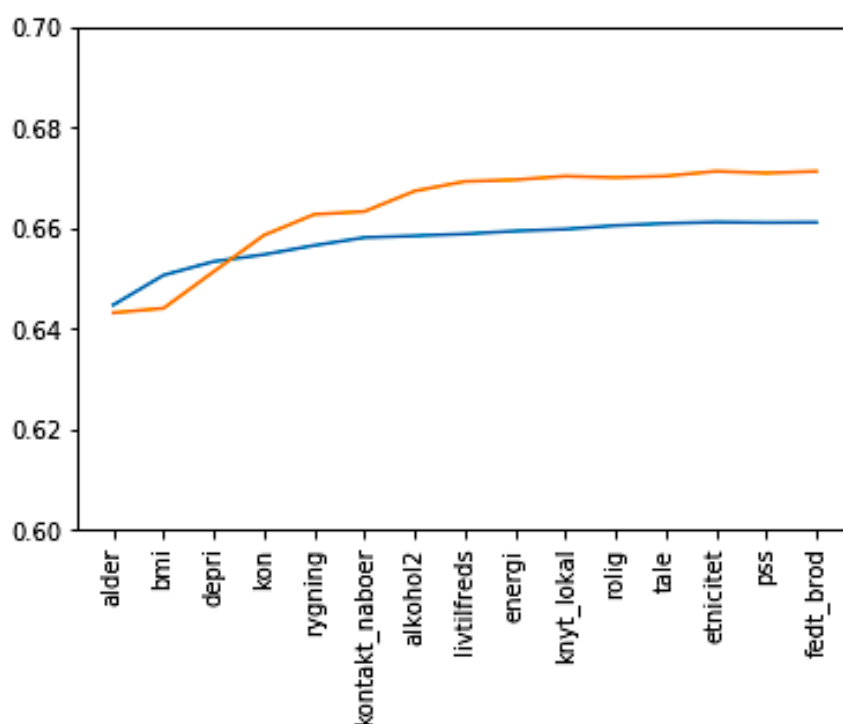
Her anvendes kun de 15 variable, som blev udvalgt i den trinvis selektionsmodel i første step. Ved inddragelse af alle 15 variable opnåede vi i valideringsdatasættet en AUC-score på 0,67. På baggrund af denne værdi vurderes det ikke, at modellen har en høj generaliserbarhed i forhold til nye data, idet modellen kun prædikerer korrekt i cirka to tredjedele af tilfældene. I metodeafsnittet redegjorde vi netop for, at en AUC-score på 1 indikerer en perfekt klassificering, mens en AUC-score på 0,5 betragtes som ubrugelig (jf. afsnit "6.2.3. Supervised machine learning"). Resultatet er imidlertid ikke overraskende, idet modellens prædiktionsværdi i testdatasættet ikke er højere, da vi her opnåede en AUC-score på 0,66.

Fjerde step: Valg af antal variable i den endelige model

Antallet af variable i den endelige model er udvalgt på baggrund af den tidligere argumentation for, at det er ønskværdigt med en model, der indeholder så få variable som muligt samtidig med, at der er et ønske om en så høj AUC-score som muligt. Valget angående antal variable er truffet på baggrund af figur 9 og tabel 6 nedenfor. Den blå linje indikerer AUC-værdien for træningsdatasættet, mens den orange indikerer AUC-værdien for valideringsdatasættet.

Som det kan ses af tabel 6, stiger AUC-scoren til og med variablen, der angiver, i hvor høj grad respondenterne føler sig knyttet til sit lokalområde. Dog fremgår det af figur 9, at den orange kurve, der indikerer, i hvor høj grad modellen kan generaliseres til valideringsdatasættet, flader ud efter variablen, der angiver respondenterens rygevaner. Ud fra ønsket om at inddrage så få variable som muligt, men have den højeste mulige AUC-score, har vi valgt i den endelige model at inddrage de første fem variable: alder, BMI, i hvor høj grad respondenterne føler sig ængstelig eller deprimeret, køn samt rygning.

Figur 9: Kurvediagram over AUC-score for den logistiske regressionsmodel.



Tabel 6: Tabel over AUC-score for de forklarende variable.

Variabel	AUC-score
Alder	0,643077
BMI	0,643973
Ængstelig eller deprimeret	0,651271
Køn	0,658502
Rygning	0,662640
Kontakt til naboer	0,663217
Alkoholindtag: 5 genstande eller mere	0,667284
Livstilfredshed	0,669195
Energyniveau	0,669520
Tilknytning til lokalområde	0,670244
Rolig og afslappet	0,669964
...	...

8.1.2. Test af modellens præcision med datasættet fra 2013

Ved anvendelse af det fulde datasæt fra 2010 som træningsdatasæt og det fulde datasæt fra 2013 som testdatasæt har den endelige model med de fem ovenstående variable en AUC-score for testdatasættet på 0,66. Ud fra dette anskues det igen, at modellen har en lav grad af generaliserbarhed. Ligesom i afsnit 8.1.1. er det ikke overraskende, eftersom der ved konstruktionen af modellen på det fulde datasæt fra 2010 blev opnået en AUC-score på 0,67.

8.1.3. Præsentation af den endelige model

Den endelige logistiske regressionsmodel præsenteres i dette afsnit med odds ratio og tilhørende signifikanstests, da disse værdier giver anledning til en fortolkning af årsagssammenhænge. Af tabel 7 nedenfor fremgår resultaterne af logistisk regressionsanalyse for både datasættet i 2010 og 2013. I resten af afsnittet forholder vi os dog kun til modellen for år 2010, fordi det er denne model, vi ønsker at sammenligne med en neural netværksmodel konstrueret ud data for samme år. Årsagen til, at vi i tabel 7 nedenfor også har medtaget odds ratio og signifikansniveauer for modellen i 2013, er for at fastslå, om der i de to modeller er overensstemmelse i forhold til størrelsen og retningen for de fundne sammenhænge. Det kan ses, at dette er tilfældet.

Som det fremgår af tabel 7, er alle fem variable i modellen statistisk signifikante både i datasættet fra 2010 og 2013. Det betragtes, at modellernes pseudo R^2 er relativt lave, idet forklaringskraften blot er henholdsvis 3,6 procent i 2010 og 4,5 procent i 2013. Dette kan skyldes, at den trinvis selektionsmetode inkluderer variable i modellen, der forøger prædiktionsniveauet, og der samtidig ikke forekommer multikollinearitet. Resultaterne af denne model er således ikke nødvendigvis et udtryk for, hvilke variable der har den største forklaringskraft i forhold til den afhængige variabel, men er i stedet netop den kombination af færrest mulige variable, der kan forudsige somatisk kronisk sygdom bedst muligt.

Tabel 7: Logistisk regressionsmodel for henholdsvis 2010 og 2013.

	Model 1: 2010	Model 2: 2013
Uafhængig variabel	OR og P> z 	OR og P> z
Alder	10,3***	14,2***
BMI	4,8***	5,6***
Ængstelig eller deprimeret	1,5**	1,6***
Køn	1,3***	1,2**
Rygning	1,3***	1,4***
Konstant	0,013***	0,018***
N	22.335	19.014
Pseudo R²	0,036	0,045

Afhængige variable: Forekomsten af nyopstået somatisk kronisk sygdom 2011-2013 og 2014-2016 (0 = Nej, 1 = Ja)

*** p < 0,000

** p < 0,010

* p < 0,050

Som vi har nævnt i metodeafsnittet, har vi foretaget en analyse med kontrolvariable for at styrke undersøgelsens interne validitet. Fremgangsmåden for denne analyse har været, at samtlige af de resterende uafhængige variable enkeltvist er blevet inddraget sammen med de fem variable i den endelige regressionsmodel. Af de resterende 26 variable var de fleste insignifikante med undtagelse af følgende fem variable: Uddannelse, indkomst, hyppighed af kontakt til naboer, alkoholforbrug målt som hyppighed i forhold til at drikke minimum 5 genstande ved samme lejlighed, samt i hvor høj grad respondenterne føler, at man ikke kan være forsigtig nok, når man har med andre mennesker at gøre. Tabel 8 nedenfor viser resultaterne af den endelige logistiske regressionsmodel ved inddragelse af disse fem variable, både enkeltvist (model 2-6), og som en samlet model (model 7):

Tabel 8: Den endelige logistiske regressionsmodel med kontrolvariable.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Uafhængig variabel	OR og P> z 	OR og P> z 	OR og P> z 	OR og P> z 	OR og P> z 	OR og P> z 	OR og P> z
Alder	10,3***	9,54***	10,19***	8,98***	9,09***	9,69***	7,38***
BMI	4,8***	4,45***	4,68***	4,63***	4,81***	4,65***	4,28***
Ængstelig og deprimeret	1,5**	1,43**	1,45**	1,50***	1,46**	1,44**	1,44**
Køn	1,3***	1,27***	1,26***	1,28***	1,22***	1,26***	1,24***
Rygning	1,3***	1,27***	1,28***	1,28***	1,32***	1,28***	1,28***
Uddannelse	-	0,61***	-	-	-	-	0,70**
Indkomst	-	-	0,56**	-	-	-	i.s.
Kontakt til naboer	-	-	-	1,34***	-	-	1,31**
Alkoholforbrug	-	-	-	-	0,73*	-	0,75*
Forsigtighed	-	-	-	-	-	1,27*	i.s.
Konstant	0,013***	0,018***	0,016***	0,012***	0,015***	0,012***	0,018***
N	22.335	22.335	22.335	22.335	22.335	22.335	22.335
Pseudo R²	0,036	0,038	0,037	0,038	0,037	0,037	0,040

Afhængig variabel: Forekomsten af nyopstået somatisk kronisk sygdom 2011-2013 (0 = Nej, 1 = Ja)

*** p = 0,000

** p < 0,010

* p < 0,050

Som det fremgår af tabellen, er der ingen af modellerne med kontrolvariable, der ændrer på signifikansniveauet af de fem oprindelige variable. De påviste sammenhænge skyldes altså ikke en effekt mellem de øvrige inddragede faktorer og somatisk kronisk sygdom. Nu hvor dette er påvist, vil vi fortolke på modellen med de oprindelige fem variable udvalgt med den trinvis selektionsmetode. Dog vil kontrolvariablen for uddannelse også indgå i tolkningen, eftersom denne variabel knytter sig til tolkningen af sammenhængen mellem somatisk kronisk sygdom og sundhedsrelateret adfærd, jævnfør den teoretiske redegørelse af sammenhængen mellem social position og livsstil (Bourdieu 1984; Bourdieu 1997; Bourdieu 2000). De følgende fortolkninger er ikke udtryk for prædiktioner i fremtiden. De skal i stedet som nævnt forstås således, at det er sociale og

psykologiske determinanter, der har betydning for den konstruerede model på det givne tidspunkt. Når modellen testes ved hjælp af datasættet fra år 2013, ved vi desuden fra AUC-scoren, at den samme grad af prædiktionssevne er gældende for et nyt datasæt, omend prædiktionssevnen, som vi har set, hverken er høj i 2010 eller 2013. Vi har også set, at sammenhængen mellem de forklarende variable og forekomsten af nyopstået sygdom er den samme for 2013, idet odds ratio-værdierne er sammenlignelige for de to år. Ud fra dette antager vi, at de samme faktorer ligeledes vil være af betydning i nye datasæt indenfor en overskuelig årrække. Derudfra anskuer vi, at fortolkningerne kan anvendes til, hvilke sociale og psykologiske determinanter på individniveau der i fremtiden kan have betydning for, hvem der får en somatisk kronisk sygdom.

8.1.4. Fortolkning ud fra odds ratio

Aldersvariablen har den højeste odds ratio på 10,3. Det betyder, at sandsynligheden for at have en somatisk kronisk sygdom er lidt mere end 10 gange større hos personer med værdien 1 på aldersvariablen sammenlignet med personer, der har værdien 0. Eftersom alder er en intervallskaleret variabel rangerende fra 16 til 99 år, hvor 0 således angiver 16 år, mens 1 angiver 99 år, giver det mening, at den relative forskel fra 0 til 1 er høj. Det er i øvrigt heller ikke et overraskende resultat, at alder i denne undersøgelse har en signifikant effekt på forekomsten af somatisk kronisk sygdom, fordi det i eksisterende forskning er vist, at netop alder udgør en særligt stor risikofaktor i forhold til forekomsten af somatiske kroniske sygdomme (jf. problemfeltet). Ses det i lyset af livsforløbet som forklaringsmodel til ulighed i sundhed, antages det, at der gennem et menneskeliv ophober sig en akkumulation af ulemper og fordele, som alle er med til at påvirke individets overordnede sundhedsstatus (Nettleton 2013:165). Ud fra denne betragtning vil også risikoen for at udvikle en somatisk kronisk sygdom øges jo flere år, et menneske har levet.

BMI indgår som den variabel, der har den næsthøjeste odds ratio i modellen. Sandsynligheden for at have en somatisk kronisk sygdom er 4,8 gange højere for personer, der har undersøgelsens højeste BMI i forhold til personer, der har undersøgelsens laveste BMI. Igen skal det bemærkes, at der er stor afstand mellem de personer, der har undersøgelsens laveste BMI (13,7) og dem, der har undersøgelsens

højeste BMI (68,4). Ifølge Sundhedsstyrelsens anbefalinger bør ens BMI ligge i intervallet mellem 18,5 og 25, før man betegnes som "normalvægtig" (Sundhedsstyrelsen 2018a:5). Derfor er vi opmærksomme på, at der i begge af variabelens yderpunkter er tale om et BMI, der kan tænkes at være forbundet med negative sundhedsmæssige konsekvenser. Dog ser det alligevel ud til, at det at have et BMI, der ifølge Sundhedsstyrelsen er for lavt, i mindre grad har en negativ effekt på forekomsten af somatiske kroniske sygdomme end det at have et BMI, der betragtes som for højt. I forhold til spørgsmålet "Ryger du?" ses det ligeledes, at risikoen for at få en kronisk sygdom er 1,3 gange højere for dem, der ryger hver dag, frem for dem, der aldrig har røget. Eftersom både BMI og rygning hører til den del af de forklarende variable, der handler om sundhedsrelateret adfærd, vil der i det følgende foretages en samlet teoretisk fortolkning af de resultater, der knytter sig til de to variable.

Med udgangspunkt i Bourdieus to rum rummet af sociale positioner og rummet af sociale positioneringer er der sammenhæng mellem individers sociale position og livsstil (Bourdieu 1997:24). Williams forklarer, at den sociale position ligeledes hænger sammen med sundhedsrelateret adfærd, der kan anses som livsstil appliceret på sundhedsfeltet (Williams 1995:583). Derfor kan sundhedsrelateret adfærd forklares ud fra en forståelse af livsstil som en social praksis, der hænger sammen med individets klassetilhørsforhold som eksempelvis uddannelsesniveau. Studier af ulighed i sundhed peger på, at det især er den lavtuddannede del af befolkningen, der har en sundhedsrelateret adfærd, der påvirker den overordnede sundhedstilstand i en negativ retning (jf. problemfeltet). Bourdieus begreb om kulturel kapital i institutionaliseret form (Bourdieu 1986:243) tilbyder den forklaring, at det kan være mangel på viden blandt den lavtuddannede del af befolkningen, der medfører en uhensigtsmæssig sundhedsrelateret adfærd. Læner man sig i stedet op ad den kulturelle forklaringsmodel og Bourdieus begreb om den kropsliggjorte kulturelle kapital, kan ulighed i sundhedsrelateret adfærd skyldes, at der knytter sig forskellige normer for social praksis til de forskellige sociale positioner og sociale lag (Nettleton 2013:161f; Bourdieu 1986:243). Eksempelvis kan det forestilles, at det i højere grad er socialt acceptabelt at ryge, spise usundt og fravælge motion blandt visse befolkningsgrupper, end det er tilfældet i andre. Disse tolkninger er overensstemmende med odds ratio for kontrolvariablen for uddannelse, der netop indikerer, at personer med en lang videregående uddannelse har lavere risiko for at få en

somatisk kronisk sygdom, end det er tilfældet for personer uden uddannelse, idet odds ratio for denne variabel er 0,7. Det kan således tyde på, at der er en positiv korrelation mellem uddannelsesniveau og BMI og rygning således, at individer med en højere social position i form af et højere uddannelsesniveau både har et lavere BMI og ryger mindre.

Uligheden i BMI og rygevaner kan ligeledes anskues som et udtryk for symbolsk vold, hvor begrebet indfanger, at det er den dominerende klasse, der fastsætter sundhedsnormer i forsøget på at tage afstand fra lavere stillede klasser (Williams 1995:588f). Det kan således i denne analyse tolkes som om, at de personer med størst risiko for at udvikle en ny somatisk kronisk sygdom, er personer fra lavere sociale lag, der ikke efterlever den dominerende klasses normer og forventninger til en passende sundhedsrelateret adfærd.

Ud fra begrebet om økonomisk kapital kan uligheden imidlertid også tilskrives økonomiske årsager, som eksempelvis, at "usunde" madvarer oftest er billigere, hvor de billigere usunde madvarer eksempelvis kan være kød med et højt fedtindhold. Det økonomiske argument holder dog ikke, når det kommer til rygning, da cigaretter med stigende afgifter er blevet dyrere, og derfor må forklaringen her nærmere tænkes at være kulturelt betinget. Generelt kan det diskuteres i hvor stor udstrækning, det økonomiske aspekt har betydning for folks sundhedstilstand i Danmark. Modsat mange andre lande har vi blandt andet en offentlig sygesikring, og der ydes tilskud til medicin og forskellige sundhedstilbud. Der er imidlertid stadig egenbetaling på en lang række sundhedstilbud, hvilket eksempelvis er tandlæge, fysioterapeut og psykolog. Dertil kommer eksempelvis også, at medlemskaber af sportsklubber og fitnesscentre kan være dyre, hvilket måske ikke er en økonomisk prioritering for lavt bemidlede familier.

Bourdieu's teoretiske perspektiv kan ligeledes bruges til at forklare uligheden mellem mænd og kvinder. Som nævnt i teori afsnittet er hverken habitus eller individets position i det sociale rum uafhængige af individets køn. Derfor har også individets køn en indvirkning på sundhedsrelateret adfærd og derigennem risikoen for at udvikle diverse sygdomme (Bourdieu 1984:107ff; Williams 1995:583). Som det ses i tabel 8 har kvinder (tildelt værdien 1) 1,3 gange større sandsynlighed for at udvikle en somatisk kronisk sygdom i forhold til mænd (tildelt værdien 0). Ud fra Bourdieu og Wacquants

betragtninger om det socialt konstruerede køn kan det tolkes således, at forskellen mellem mænd og kvinder skyldes de forskellige normer og forventninger til sundhedsrelateret adfærd, der knytter sig til hvert køn (Bourdieu & Wacquant 1992:171). Det vil sige, at individets køn eksempelvis kan have indflydelse på hans eller hendes mad- eller motionsvaner, og disse faktorer kan have betydning for sandsynligheden for at udvikle en somatisk kronisk sygdom. Det kan imidlertid også forklares med, at det muligvis er mere acceptabelt for kvinder at opsøge sundhedssystemet, og at de dermed oftere end mænd fremgår af registrene, på trods af at den reelle forskel i forekomsten af sygdomme måske ikke er til stede.

I den sidste variabel spørges overordnet: "Hvordan vurderer du din sundhedstilstand?", og her er der fokus på angst og depression. Respondenten har haft mulighed for at svare følgende: "Jeg er ikke ængstelig eller deprimeret", "Jeg er moderat ængstelig eller deprimeret" eller "Jeg er ekstremt ængstelig eller deprimeret". I denne analyse er variablen kodet således, at værdien 1 er tildelt, hvis respondenten har svaret, at han eller hun er ekstremt ængstelig eller deprimeret, og modsat har værdien 0, hvis det modsatte er angivet. Odds ratio er 1,5, hvilket vil sige, at personer, der har angivet, at de er ekstremt ængstelige eller deprimerede, har 1,5 gange større sandsynlighed for at få en somatisk kronisk sygdom. Som vi tidligere har nævnt i teoriafsnittet, udgør individets mentale reaktion på livsbegivenheder og sociale omgivelser den største determinant for sundheden inden for den psykosociale forklaringsmodel (Elstad 2000:69ff). Derfor er denne forklaringsmodel særligt anvendelig til at forklare denne sammenhæng mellem den mentale sundhed og somatiske kroniske sygdomme. Ud fra denne forklaring kan oplevelsen af angst og depression tænkes både at have en direkte og en indirekte effekt på forekomsten af kronisk sygdom hos individet (ibid.). For det første kan mentale reaktioner i sig selv føre til sygdom. Det vil sige oplevelsen af angst og depression kan være direkte udslagsgivende i forhold til forekomsten af somatiske kroniske sygdomme. Derudover kan livsstil også fungere som medierende led, hvor angst og depression medfører en dårligere sundhedsrelateret adfærd, som igen medfører større risiko for udviklingen af en somatisk kronisk sygdom (Elstad 2000:69ff). Yderligere kan den psykosociale forklaringsmodel også anvendes til at forklare uligheden i somatiske kroniske sygdomme ud fra et klasseperspektiv. Tolkningen af variablen for individets følelse af angst og depression kan dermed ses i relation til kontrolvariablen for

uddannelse. Kontrolvariablen viser, at personer med grunduddannelse har større risiko for somatiske kroniske sygdomme end personer med lang videregående uddannelse. Det kan herudfra tænkes, at der er en negativ korrelation mellem uddannelse og oplevelsen af angst og depression. Oplevelsen heraf kan i denne optik forstås som en mental reaktion på en ekstern stressor, som nogle mennesker udsættes for, og som kan medføre dårligere fysisk sundhed blandt denne gruppe. Omvendt kan der også argumenteres for, at gruppen af personer, der oplever angst og depression, ikke har samme muligheder for at håndtere disse stressorer, idet deres ressourcer kan være mere begrænsede i forhold til personer, der ikke oplever angst og depression. Disse tolkninger er i overensstemmelse med betragtninger inden for den psykosociale forklaring (Elstad 2000:74ff).

8.2. Analyse med neuralt netværk

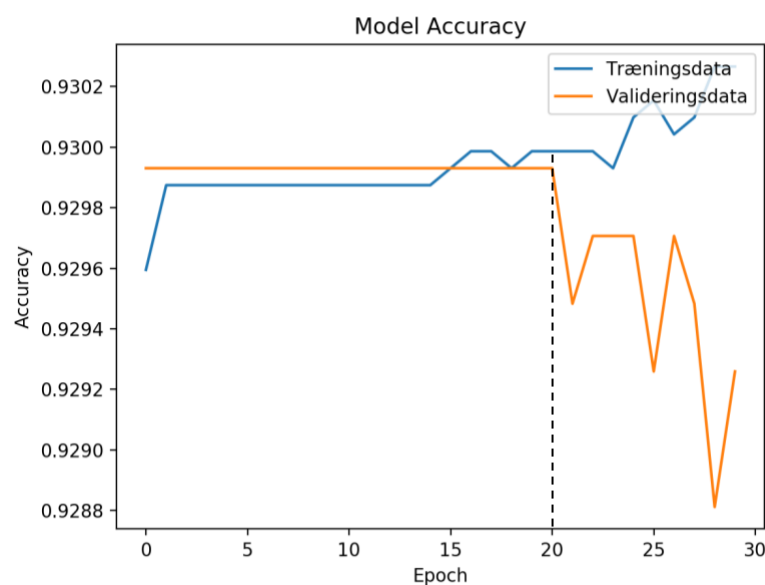
I dette afsnit beskrives den analyse, der er foretaget ved hjælp af neuralt netværk. Afsnittet indeholder først en beskrivelse af fremgangsmåden for konstruktionen af modellen ved hjælp af datasættet fra 2010. Derefter evalueres modellen ved hjælp af datasættet fra 2013. Til sidst appliceres den endelige konstruerede model på datasættet fra 2017, hvor udfaldet på den afhængige variabel for somatisk kronisk sygdom er ukendt. Ved hjælp af den konstruerede model opnås en prædikeret sandsynlighed for forekomsten af nyopstået somatisk kronisk sygdom for det enkelte individ i datasættet for 2017. For at fastslå hvilke af specialets forklarende variable der har haft betydning for den prædikterede sandsynlighed, afsluttes dette afsnit som nævnt med en logistisk regressionsanalyse, hvor den prædikterede sandsynlighed anvendes som afhængig variabel.

8.2.1. Konstruktion af modellen

Såvel som ved logistisk regression er den neurale netværksmodel konstrueret ved hjælp af data fra 2010. Det fulde datasæt fra 2010 blev igen tilfældigt inddelt i trænings- og valideringsdata ud fra samme fordeling i 80/20. Ligeledes testede vi for, om andelen af personer med en somatisk kronisk sygdom var lige stor i både trænings- og valideringsdatasættet. Vi fandt igen, at andelen i begge datasæt var 0,07. I træningen af modellen har vi anvendt de aktiverings- og optimeringsfunktioner, der er redegjort for i

afsnit “6.3.2. Neural netværksanalyse”. Som det ligeledes er nævnt i metodeafsnittet, har vi anvendt følgende justeringsparametre i forhold til modellens arkitektur: Fire lag hvor der indgår et input-lag, to skjulte lag og et output-lag. Antallet af neuroner i hvert lag er som følger: 31 (svarende til antallet af forklarende variable), 40, 20 og 1. Derudover har vi som nævnt i samme afsnit anvendt en batch-size på 15, der er det ene af de justeringsparametre, der knytter sig til optimering af modellen. I dette afsnit vil vi redegøre for vores valg i forhold til det andet justeringsparameter, der knytter sig hertil, hvilket er antallet af epoker i den endelige model. Det vurderes ud fra to figurer, der angiver henholdsvis modellens præcision samt, hvor meget information der tabes gennem lagene ved det pågældende antal epoker. Igen er det vigtigt at nævne, at grundet den tilfældige inddeling i trænings- og valideringsdata, kan resultatet være forskelligt for hver gang, modellen køres. Derfor har vi foretaget fem gentagelser med 30 epoker, og vurderet det endelige antal epoker ud fra dette. I de fem gentagelser var modellerne i høj grad sammenlignelige, og figurerne for de fem gentagelser fremgår af bilag 6. Nedenfor bruges de to figurer fra sidste kørsel til at illustrere et eksempel på, hvordan antallet af epoker er udvalgt.

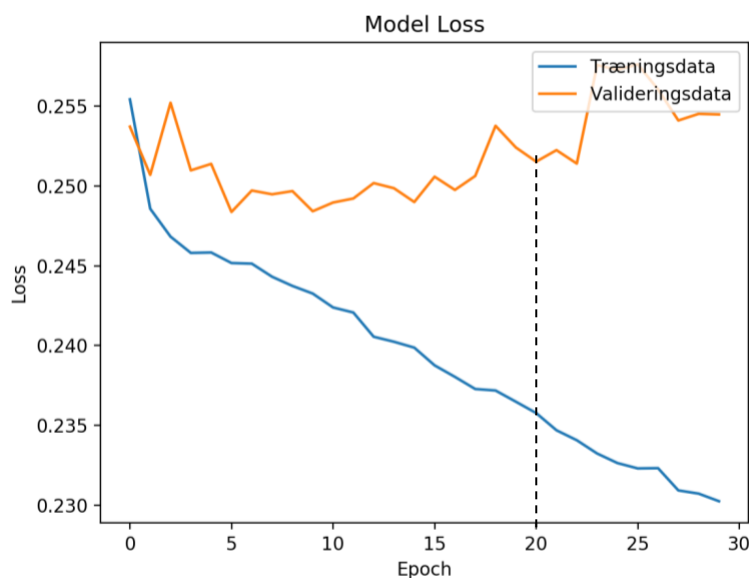
Figur 10: AUC-score som funktion af antal epoker.



Figur 10 angiver modellens præcision (AUC-score) for hver epoke i henholdsvis trænings- og valideringsdatasættet. Fordi modellen konstrueres på træningsdatasættet, vil præcisionen forøges i dette datasæt jo flere epoker, der foretages. Vi har et ønske om

så høj en prædiktionsværdi som muligt i valideringsdatasættet, eftersom det er ved høj præcision på nye data, at modellens generaliserbarhed vurderes (jf. argumentation i afsnit “6.2.3. Supervised learning”). Som det fremgår af figuren, falder AUC-scoren i valideringsdatasættet ved cirka 20 epoker. Figur 11 nedenfor angiver mængden af information, der tabes gennem modellens lag for hver epoke. Denne ønskes så lav som muligt i valideringsdatasættet. Ud fra denne figur kan vi se, at tabet stiger langsomt efter 5 epoker.

Figur 11: Modellens tab som funktion af antal epoker.



Sammenholder vi de to figurer og ligeledes sammenligner med de resterende figurer for de i alt 5 kørsler, vurderer vi, at 20 epoker er det optimale antal. Det skyldes, at vi først og fremmest går efter den højest mulige prædiktionsværdi og desuden er opmærksomme på, at tabet af information ikke stiger markant inden for dette antal epoker.

8.2.2. Evaluering af modellen

Ved evaluering af den endelige model opnås en AUC-score i valideringsdatasættet fra 2010 på 0,93. Derefter har vi gemt den endelige model og evalueret den med testdatasættet fra 2013. I denne test fastslås modellens præcision i forhold til et nyt datasæt med en AUC-score på 0,89. Ud fra disse tal kan det umiddelbart betragtes, at den konstruerede algoritme har en høj prædiktionssevne, idet den er i stand til at forudsige, hvorvidt en person får en somatisk kronisk sygdom med 89 procent sikkerhed, når den

appliceres på det nye data. Den følgende analyse af den konstruerede model giver imidlertid anledning til at betvivle brugbarheden af den klassifikation, som AUC-scoren er udregnet på baggrund af, og derigennem modellens generaliserbarhed. Dette vil blive diskuteret i afsnit "8.3. Diskussion af analyseresultater".

8.2.3. Forudsigelse af somatiske kroniske sygdomme 2018-2020

Efter appliceringen af den konstruerede model på datasættet fra år 2017 har vi opnået forudsigelser af sandsynligheden for somatiske kroniske sygdomme hos det enkelte individ i årene 2018-2020. Det findes, at den højeste prædikterede sandsynlighed, der opnås, er 31,3 procent, som altså er sandsynligheden hos den person med størst prædikteret risiko for at udvikle en somatisk kronisk sygdom. Derudover ses det, at den gennemsnitlige prædikterede sandsynlighed blandt alle personer i datasættet fra år 2017 er 9,4 procent. Dette er sammenligneligt med den faktiske andel med en nyopstået somatisk kronisk sygdom i datasættet fra år 2010, som modellen er konstrueret på baggrund af, hvor andelen i disse år udgjorde 7 procent. Det er ikke en forudsætning for modellens validitet, at andelen af sygdomstilfælde er sammenligneligt på tværs af de to datasæt, eftersom man kan forestille sig, at der over tid kan være forskellige andele, der får en somatisk kronisk sygdom. Det betragtes dog alligevel som en indikator for, at beregningen af de prædikterede sandsynligheder er meningsfuld, fordi vi mener, det virker plausibelt, at andelen med nyopstået somatisk kronisk sygdom er forholdsvis stabil over tid. Det betragtes, at dette resultat fra den neurale netværksanalyse i sig selv kan bruges til at prædiktere, hvem og hvor mange der bliver syge, men det er ikke muligt herudfra at vide, hvad der har betydning for, hvem der bliver syge. Derfor vil vi i næste afsnit benytte de prædikterede værdier i en logistisk regressionsanalyse for at kunne sige noget om, hvilke sociale og psykologiske determinanter på individniveau der har betydning.

8.2.4. Analyse af de prædikterede værdier med logistisk regression

I dette afsnit foretager vi analysen med logistisk regression, med den afhængige variabel for prædikterede sandsynligheder for 2018-2020. Denne variabel er som nævnt inddelt i to grupper, hvor den ene gruppe udgør de 10 procent af det samlede sample, der har størst prædikteret sandsynlighed for en somatisk kronisk sygdom, mens den anden

gruppe udgør de resterende 90 procent. Den gennemsnitlige sandsynlighed for at udvikle en somatisk kronisk sygdom i de to grupper er henholdsvis 18,8 procent i top-10-procent gruppen, og 8,4 procent sandsynlighed i blandt de øvrige 90 procent af samplet. Det er denne binære variabel, der anvendes som den afhængige variabel i den logistiske regressionsanalyse.

Først har vi inddraget samtlige af specialets 31 forklarende variable i en samlet model, hvor det viste sig, at alle disse havde en statistisk signifikant effekt på den prædikterede sandsynlighed med undtagelse af variabelen for motion, det sammensatte mål for mængden af fedtstof, der anvendes i madlavningen samt variabelen for, i hvor høj grad respondenterne stoler på personer i sit lokalområde (jf. afsnit "6.5. Forudsigelse af somatiske kroniske sygdomme"). Resultaterne af denne analyse fremgår af tabel 9 nedenfor. For at fremvise en overskuelig analyse har vi valgt ikke at kommentere på alle signifikante variable men i stedet at udvælge de mest centrale baseret på størrelsen af odds ratio. Således tolker vi fra tabel 9 kun på de variable, hvor forholdet mellem yderkategorierne (med værdierne 0 og 1) er større end 1:10 - det vil sige odds ratio på henholdsvis over 10,0 og under 0,10 afhængigt af, hvordan forholdet vender. En enkelt undtagelse er dog, at vi også har inddraget køn, selvom odds ratio for denne variabel i den samlede model blot er 5,7. Det skyldes, at eftersom køn er en binær variabel, er den reelle afstand mellem de to køn større, end det er tilfældet for de enkelte kategorier i de øvrige variable, der er enten ordinal- eller intervallskalerede.

Som det kan ses, har modellen med de udvalgte variable en høj forklaringskraft, sammenlignet med den, vi så i afsnit 8.1.3. Pseudo R^2 er nemlig 0,42, hvilket vil sige at de inddragede variable forklarer 42 procent af den samlede varians på den afhængige variabel.

Tabel 9: Logistisk regressionsmodel for binær prædikeret sandsynlighed.

Uafhængig variabel	OR og P> z
Køn	5,7***
Alder	3855,7***
Uddannelse	10,8***
Etnicitet	0,07***
Indkomst	0,69*
Rolig og afslappet	0,06***
Energiniveau	26,1***
Trist til mode	5,8***
Ængstelig eller deprimeret	0,35***
Overordnet livstilfredshed	1,8***
Percieved Stress Scale	0,15***
Kontakt til familie	1,6***
Kontakt til venner	3,4***
Kontakt til naboer	1,2*
Uønsket alene	0,11***
Nogen at tale med	2,0***
Fedtstof på brød	4,1***
Fedtstof i mad	1,6 (i.s.)
Indtag af grøntsager	1,6**
Indtag af frugt	7,5***
Rygning	0,31***
Alkoholforbrug, antal dage om ugen med alkohol	0,14***
Alkoholforbrug, 5 genstande ved samme lejlighed	0,49***
Fysisk aktivitetsniveau i fritiden	1,2 (i.s.)
BMI	0,12***
Stoler på personer i lokalområde	0,76 (i.s.)
Tilknytning til lokalområde	0,23***
Tryk ved at gå alene i lokalområde	0,33***
Stoler på andre mennesker	2,8***
Forsigtighed i forhold til andre mennesker	0,09***
Tillid til samfundsmæssige instanser	7,4***
Konstant	0,0001***
N	22.583
Pseudo R²	0,42

Afhængig variabel: Prædikeret sandsynlighed (1 = Top 10 procent, 0 = Øvrige 90 procent)

*** p = 0,000

** p < 0,010

* p < 0,050

8.2.5. Fortolkning ud fra odds ratio

Det mest bemærkelsesværdige ved resultaterne i tabel 9 er den høje odds ratio for aldersvariablen. Det ses, at den prædikterede sandsynlighed for at udvikle en ny somatisk kronisk sygdom er 3855,7 gange større for personer, der har den højeste værdi på aldersvariablen (99 år) i forhold til personer med samplets laveste alder (16 år). Det er ingen overraskelse, at risikoen er større, jo ældre en person er, eftersom det både er overensstemmende med resultater fra tidligere forskning på området (jf. problemfeltet), men også i overensstemmelse med resultaterne af den logistiske regressionsanalyse på datasættet fra år 2010, der fremgår af tabel 8 tidligere i analyseafsnittet. I den tidligere analyse var odds ratio for aldersvariablen imidlertid blot 10,3. I denne analyse af den prædikterede sandsynlighed har aldersvariablen således markant større betydning, end det er tilfældet i analysen af den faktiske sygdomsforekomst i datasættet fra 2010, som modellen blev konstrueret på. Det vil sige, at der i træningen af modellen formodentlig sker en overdimensionering af betydningen af aldersvariablen baseret på de vægte, der sendes videre gennem de skjulte lag i den neurale netværksmodel. Det er i øvrigt gældende for samtlige af de inddragede variable i tabel 9, at størrelsen af odds ratio er relativt høj i forhold til tabel 8. Det er vigtigt at understrege, at vi ikke med sikkerhed kan udtale os om de faktiske størrelser af vægtene i modellens skjulte lag, eftersom disse netop betegnes som *black boxes* (jf. afsnit "6.4. Diskussion af de to undersøgelsesmetoder"). Antagelsen, om at det er en overdimensionering i det neurale netværks vægtning, er derfor blot vores tolkning ud fra den relativt markante stigning i odds ratio.

Foruden aldersvariablen indgår af øvrige variable fra temaet "socioøkonomiske faktorer" også uddannelse, etnicitet og køn i modellen. Retningen af sammenhængen for kønsvariablen er overensstemmende med det tidligere resultat, hvor vi fandt, at kvinder har størst sandsynlighed for at udvikle en ny somatisk kronisk sygdom i årene 2011-2013 (jf. afsnit 7.1.3). Ligeledes har kvinder i tabel 9 større prædikteret sandsynlighed for nyopstået sygdom i 2018-2010, hvor de har 5,7 gange større sandsynlighed end mænd. Ligesom tidligere kan denne forskel forklares med Bourdieus habitusbegreb, idet udformningen af individets habitus afhænger af socioøkonomiske faktorer, herunder køn og etnicitet (Bourdieu 1984:107ff). Derudfra kan forskellen mellem mænd og kvinder eksempelvis forklares med, at der blandt kvinder og mænd findes forskellige normer for

passende sundhedsrelateret adfærd, og risikoen for somatisk kronisk sygdom således er en afledt effekt heraf.

At den prædikterede sandsynlighed for at få en somatisk kronisk sygdom med en odds ratio på 0,07 er lavere for efterkommere (med værdien 1 på variablen), i forhold til etniske danskere (med værdien 0 på samme variabel), kan jævnfør ovenstående betragtning om, at habitus er influeret af etnicitet, ligeledes forklares ud fra habitusbegrebet. I denne optik kan uligheden i forekomsten af somatiske kroniske sygdomme skyldes forskellige adfærdsmønstre i de to etniske grupper, hvor det tyder på, at den sundhedsrelaterede adfærd blandt danskerne i højere grad leder til en dårlig sundhed. Dertil kan knyttes, at der kan være forskellige normer for passende sundhedsrelateret adfærd i de to grupper.

I forhold til uddannelsesvariablen ses det, at den prædikterede sandsynlighed for nyopstået somatisk kronisk sygdom er 10,8 gange højere for personer, der har en lang videregående uddannelse, end det er tilfældet for personer med grunduddannelse. Dette tal er overraskende, hvilket for det første skyldes, at det er modstridende med resultaterne i tabel 8 i afsnit 8.1.3, idet sandsynligheden for at have en nyopstået sygdom her var 0,7 for de højtuddannede i forhold til personer med grunduddannelse. Derudover strider det også mod resultaterne af den eksisterende forskning på området (jf. problemfeltet). Ved anvendelse af Bourdieus og Williams' teoretiske betragtninger (Bourdieu 1997:21ff; Williams 1995:588ff), vil det i forhold til uddannelsesvariablen i denne model kunne tolkes således, at individer med en højere social position har tilegnet sig vaner i forhold til sundhedsrelateret adfærd, der fremmer en dårligere sundhed blandt denne befolkningsgruppe. At den prædikterede sandsynlighed er større blandt højtuddannede kan omvendt tolkes ud fra begrebet om symbolsk vold således, at den dominerende klasse gennem kapitaler har adgang til, og dermed praktiserer en adfærd, der ikke er tilgængelig for de lavere stillede klasser, hvor denne adfærd både af de dominerende og de dominerede grupper betragtes som den "rigtige" (Bourdieu & Passeron 2006:9+24). Ligesom det er tilfældet for "dårlig" sundhedsrelateret adfærd, som eksempelvis rygning blandt den lavere stillede klasse, kan en bestemt adfærd hos den dominerende klasse ligeledes være sundhedsskadelig. Som eksempel kan man forestille sig, at højtuddannede har et andet forbrug af alkohol, idet det kan tænkes at

være almindeligt inden for denne gruppe at drikke et par glas vin i løbet af ugen eller tage en whisky eller cognac på direktørens kontor. Dette forklarer imidlertid ikke, hvordan forholdet kan være modsatrettet i den prædikterede sandsynlighed i forhold til de kausale sammenhænge vi har påvist i datasættet fra år 2010, som den neurale netværksmodel er konstrueret ud fra. I denne sammenhæng findes det væsentligt at nævne det bias, der som nævnt i afsnit "5.2.1. Afhængig variabel: Nyopstået somatisk kronisk sygdom" potentielt forekommer i denne undersøgelse. Fordi det afhængige mål er et udtryk for nyopståede somatiske kroniske sygdomme, er sandsynligheden for at få værdien 1 på denne variabel større, jo færre af de udvalgte sygdomme, der haves i forvejen. Således kan det også formodes, at sandsynligheden for en ny kronisk sygdom er højere blandt personer med en lang videregående uddannelse, som ud fra både tidligere forskning og de teoretiske betragtninger må antages at have færre af disse sygdomme i forvejen (jf. problemfeltet). Dette bias vil blive diskuteret i afsnit "8.3. Diskussion af analyseresultater".

I forhold til temaet om mental trivsel indgår to variable herfra i modellen. Det drejer sig om variablene, der omhandler, hvor ofte respondenter inden for de seneste 4 uger har følt sig henholdsvis fuld af energi og følt sig rolig og afslappet. Det ses, at den prædikterede sandsynlighed for en nyopstået somatisk kronisk sygdom er mindre, jo højere værdi der haves på variabelen angående rolig og afslappethed, da odds ratio er 0,06. Personer, der føler sig rolige og afslappede "Hele tiden" (1), har altså lavere risiko for sygdom, end personer, der "På intet tidspunkt" (0) føler sådan. Dette er i overensstemmelse med de teoretiske betragtninger inden for den psykosociale forklaringsmodel, hvor det som nævnt er individets mentale reaktion på dets livsbegivenheder og sociale omstændigheder, der udgør den største determinant for dets sundhedsstatus (Elstad 2000:69ff). Ud fra dette perspektiv kan det anskues, at personer der føler sig rolige og afslappede også har lavere sandsynlighed for, at deres mentale reaktioner (eller mangel på samme) har en negativ indflydelse på den overordnede sundhed. Dermed haves også lavere sandsynlighed for forekomsten af de udvalgte somatiske kroniske sygdomme. Hvad angår variabelen, der omhandler respondentens energiniveau, ses imidlertid det modsatte forhold. Personer, der "Hele tiden" har oplevet at være fulde af energi, har nemlig 26,1 gange større prædikteret sandsynlighed for at få en somatisk kronisk sygdom, end det er tilfældet for personer, der "På intet tidspunkt" i

løbet af 4 uger op til besvarelsen af spørgeskemaet følte sig fulde af energi. Dette er umiddelbart modstridende med de antagelser, der haves inden for den psykosociale forklaringsmodel. Forklaringen skal måske igen findes i det før omtalte bias, der skyldes, at sandsynligheden for at få en ny kronisk sygdom bliver mindre, jo flere af disse sygdomme, der haves i forvejen. Sammenstilles resultaterne af variabelen for energiniveau med resultaterne af uddannelsesvariabelen, kan det betragtes, at det inden for den psykosociale forklaringsmodel anskues, hvordan social position har betydning for sammenhængen mellem mental trivsel og sundhedsstatus. Det skyldes, at personer med lav social status for det første udsættes for flere stressorer men også, at de ofte i mindre grad er i stand til at håndtere de eksterne stressorer, de udsættes for grundet færre ressourcer til netop dette (Elstad 2000:69+73ff). Fordi det fremgår, at både et højt energiniveau og et højt uddannelsesniveau medvirker til større sygdomsrisiko, kan det tænkes, at disse to variable korrelerer således, at der samlet set er tale om en gruppe bestående af mennesker med både et højt uddannelsesniveau og et højt energiniveau. Det omtalte bias kan bruges til at forklare, hvorfor både variabelen for energiniveau og variabelen for uddannelsesniveau vender omvendt i forhold til vores forventninger, men det kan imidlertid ikke forklare, hvorfor dette forhold ikke gælder for de øvrige variable, og ej heller er gældende i den tidligere logistiske regressionsmodel.

Yderligere betragtes det, at variabelen under temaet "tillid" omhandlende respondenternes holdning til, hvorvidt man ikke kan være forsigtig nok, når man har med andre mennesker at gøre har en odds ratio på 0,09. Dermed fremgår det af modellen, at respondenter, der er meget enige i dette har en markant større sandsynlighed for at få en somatisk kronisk sygdom. I dette speciale betragtes det at være meget forsigtig i forhold til andre mennesker som en indikator for en lav grad af tillid, og derfor kan også dette resultat fortolkes ved hjælp af den psykosociale forklaringsmodel. Selvom vi i Danmark generelt har en høj levestandard, er der imidlertid stadig materiel ulighed, og dette medfører ifølge den psykosociale forklaringsmodel både mangel på social samhørighed og tillid mellem mennesker (Elstad 2000:78f). Med udgangspunkt i den psykosociale forklaring kan det anskues, at forskellen i somatisk kronisk sygdom opstår ved, at økonomisk ulighed i et samfund kan medføre mistillid generelt set, men særligt rammer individer i lavere sociale positioner. Det skyldes, at denne gruppe udover at blive ramt af

den lave sammenhængskraft, ligeledes er frustrerede over deres relative position i det sociale hierarki (Elstad 2000:78f).

8.3. Diskussion af analyseresultater

I det følgende afsnit vil vi opsummere og diskutere de to potentielle bias, der er nævnt i analysen, og som kan tænkes at påvirke analyseresultaterne.

For det første omfatter dette problemet ved at undersøge nye sygdomstilfælde, hvor den binære variabel indeholder de to kategorier: 0) raske og syge personer der ikke får en ny somatisk kronisk sygdom og 1) raske og syge personer der får en ny somatisk kronisk sygdom. Således sidestilles personer, der ikke får sygdommen, fordi de har den i forvejen, med personer, der ikke får sygdommen, og heller ikke tidligere har haft den. I analysen har vi anvendt dette bias til at forklare, hvorfor nogle af de kausale forhold ikke stemmer overens med vores forventninger. Ifølge eksisterende forskning og teoretiske betragtninger burde både et højere uddannelsesniveau og et højere energiniveau være forbundet med en lavere risiko for at få en somatisk kronisk sygdom (jf. problemfeltet; Elstad 2000; Bourdieu 1984; Williams 1995). I analysen af den prædikterede sandsynlighed forholder det sig dog omvendt. Dette bias kan være med til at forklare retningen for de kausale sammenhænge, men det kan imidlertid ikke forklare, hvorfor det kausale forhold vender sig om fra den ene analyse til den anden. I analysen af datasættet fra år 2010 fremgår det, at personer med en lang videregående uddannelse har mindre sandsynlighed for at få en somatisk kronisk sygdom i årrækken 2011-2013, mens analysen af datasættet fra år 2017 viser, at den prædikterede sandsynlighed derimod er større for personer med lang videregående uddannelse. Samtidig kan det heller ikke forklare, hvorfor dette forhold kun gælder for nogle af de inddragede variable. Eksempelvis stemmer variabelen for, hvor ofte individet føler sig rolig og afslappet, overens med de teoretiske betragtninger og eksisterende forskning (jf. problemfelt; Elstad 2000).

Det andet bias, der nævnes i analysen, relaterer sig til den forhøjede værdi af odds ratio for især aldersvariablen, der har ændret sig fra 10,3 i analysen af datasættet fra 2010 til lidt mere end 3855,7 i analysen af den prædikterede sandsynlighed for datasættet fra

2017. Selvom odds ratio for aldersvariablen er den mest iøjenfaldende, ses det, at odds ratio generelt er disproportionalt høje i modellen i år 2017 i forhold til modellen fra år 2010. I modellen fra 2010 er odds ratio for kønsvariablen eksempelvis 1,3, mens den i 2017 er steget til 5,7. Da der er tale om forskellige år, skal værdierne ikke nødvendigvis at være ens, men alligevel anskues det, at de burde være nogenlunde overensstemmende, fordi det antages, at variablenes betydning ikke vil ændres så markant over en periode på 7 år. Værdierne for den logistiske regressionsmodel betragtes som mest valide, fordi de er udregnede på baggrund af faktisk nyopstået sygdom, mens værdierne i den logiske regressionsmodel for år 2017 i stedet er beregnet ud fra prædikterede værdier. Det tyder altså på, at der nærmere er tale om en generel overestimering af betydningen af de inddragede variable i konstruktionen af den neurale netværksmodel, hvilket kan tænkes at være forårsaget af overfitting. Som vi tidligere har nævnt i metodeafsnittet om det neurale netværk, er en af de potentielle problematikker for denne metode netop overfitting (jf. afsnit "6.3.2. Neural netværksanalyse"). Problemet indebærer som nævnt, at modellen er god til at indfange de komplekse mønstre i det træningsdatasæt, som den er konstrueret på, men at modellen i ringe grad lader sig generalisere til nye datasæt (Murphy 2012:22; Molina & Garip 2019:3). En overfitting af modellen vil således typisk vise sig i en lav AUC-score, når modellen testes på nye data (Domingos 2012:81). Ved anvendelse af testdatasættet fra 2013 opnåede vi i dette speciale en AUC-score på 0,89, hvilket umiddelbart må tolkes som en høj grad af generaliserbarhed. Som det er vist i figur 10 i afsnit "8.2.1. Konstruktion af modellen", opnås der allerede fra første epoke en høj AUC-værdi, og denne forbedres ikke markant ved flere epoker. Dette anskues at kunne kobles til, at modellen for år 2010 forudsiger 93 procent af tilfældene korrekt, mens dette er gældende for 89 procent i 2013, da dette er svarende til andelen i de respektive år, der ikke får en somatisk kronisk sygdom. Derudfra betragtes det, at det kan skyldes, at modellen blot hurtigt lærer at tildele alle individer værdien 0, eftersom ingen har en prædikeret værdi over 0,5. Dette skyldes sandsynligvis, at der er en relativt lille andel, der får en somatisk kronisk sygdom, hvor vi har en skæv fordeling i forholdet mellem de to værdier 0 og 1 på den afhængige variabel. Herom kan bruges betegnelsen *ubalanceret klasseinddeling*, hvilket refererer til tilfælde, hvor en af klasserne i outputvariablen er i overtal i forhold til den anden (Godase & Attar 2012:4). Årsagen, til at en ubalanceret klasseinddeling kan være et problem, skyldes, at læringen fra sådanne skæve datainddelinger ofte resulterer i en klassificering, der er biased mod majoritetsklassen,

hvilket gør, at modellen har en tendens til at fejlklassificere observationer, der tilhører minoritetsklassen:

“When such algorithms are applied to skewed data streams then their accuracy of classifying majority examples is good but the accuracy of classifying the minority examples is poor. This happens because learning from such imbalanced/skewed streams causes the learner to become biased towards the majority class; thus the minority examples are likely to be misclassified.” (Godase & Attar 2012:5)

Der findes utallige lignende empiriske eksempler på ubalancerede klasseinddelinger inden for en lang række videnskabelige områder. Problematikken er især hyppig inden for studier af levende organismer som eksempelvis mennesker og herunder menneskelig adfærd. Eftersom anvendelsen af machine learning er en forholdsvis ny tendens inden for disse videnskaber, er udfordringen i forhold til ubalancerede data ligeledes en relativt ny betragtning (Saito & Rehmsmeier 2015:2). I de seneste år har man imidlertid erkendt problemet som et af de største inden for machine learning (Godase & Attar 2012:5). De mest almindelige måder at løse problemet på er ved enten over- eller undersampling, der begge fører til et mere balanceret forhold mellem de to klasser. I oversampling genbruges de observationer, der har minoritetsudfaldet (Godase & Attar 2012:9). I undersampling reduceres antallet af observationer, der har majoritetsudfaldet gennem tilfældig selektion (Godase & Attar 2012:10). Det findes vigtigt at nævne, at vi ikke betragter det som et problem i sig selv, at modellen prædikerer værdien 0 for samtlige respondenter på baggrund af, at ingen har en prædikeret sandsynlighed over 0,5. Det vurderes nemlig i dette tilfælde, at det virker validt, at personen med den højeste sandsynlighed har 31,3 procent risiko for at få en somatisk kronisk sygdom, mens der gennemsnitligt haves en sandsynlighed på 9,4 procent (jf. afsnit ”8.2.3. Forudsigelse af somatiske kroniske sygdomme 2018-2020”). Fordi ingen har en større sandsynlighed end dette, og dermed ikke over 50 procent, er der ikke nogen, der får tildelt værdien 1, når modellen testes med datasættet fra 2013. Problemet består derimod i, at AUC-scoren ud fra denne betragtning antages at være en dårlig indikator for modellens generaliserbarhed, fordi den fejlklassificerer observationer, der tilhører minoritetsklassen. Derfor vurderer vi efter denne diskussion, at det er problematisk at sammenligne AUC-værdierne for

testdatasættet fra 2013 på henholdsvis 0,66 i den logistiske regressionsmodel og 0,89 i den neurale netværksmodel, der er opnået i dette speciale.

Særligt i forhold til at det er en persons sandsynlighed for at få en somatisk kronisk sygdom inden for en periode på tre år, virker det rimeligt, at der ikke er nogen, der har en højere sandsynlighed for at få en somatisk kronisk sygdom. Selvom det i dette speciale antages, at der er en vis kontinuitet i folks socioøkonomiske status og/eller sundhedsrelaterede adfærd (jf. afsnit "5.2.1. Afhængig variabel: Nyopstået somatisk kronisk sygdom"), anskues det alligevel, at det er en relativt kort tidshorisont blot at se på nyopstået somatisk kronisk sygdom maksimum tre år fremtidigt, da det anskues at de inddragede somatiske kroniske sygdomme kan tage flere år at udvikle. En mulig løsning på det problem, at AUC-scoren blot er et udtryk for andelen af korrekte negative prædiktioner, kan, som nævnt ovenfor, være undersøgelsen af en afhængig variabel, hvor andelen, der har det studerede udfald (1), er højere, end det er tilfældet i dette speciale. En ændring af undersøgelsesdesignet, hvor den fremtidige sygdomsforekomst bliver undersøgt over en længere årrække, kunne være én måde at gøre det på.

Ud fra disse betragtninger om ubalanceret klasseinddeling kan vi ikke, på trods af en høj AUC-score, afvise risikoen for forekomsten af overfitting i den konstruerede neurale netværksmodel. Tværtimod tyder de overestimerede størrelser af odds ratio på, at der forekommer overfitting i modellen. De mest almindelige løsningsforslag til overfitting-problematikken er en større datamængde og en reducere af kompleksiteten i modellens arkitektur (Ruizendaal 2017). En større datamængde vurderes ikke at være muligt i Sundhedsprofilerne, eftersom samplingen til disse spørgeskemaundersøgelser i forvejen er høje. Samtidig har vi forsøgt at reducere antallet af lag fra 4 til 3 i dette speciales neurale netværksmodel, men dette medførte ikke mere plausible parameterestimer, da værdierne blev anvendt i en logistisk regression. Derfor er der behov for at afprøve andre metoder til at efterkomme denne problematik i lignende fremtidige undersøgelser.

9. Teori om geografisk segregering

I dette afsnit vil der fokuseres på sidste del af vores problemformulering, hvor der er et ønske om geografisk at kortlægge, hvordan forekomsten af somatisk kronisk sygdom vil komme til at fordele sig i Region Nordjylland. Der vil teoretisk og med udgangspunkt i eksisterende forskning redegøres for, hvorfor det er interessant at afslutte specialets undersøgelse med en geografisk områdeinddeling af, hvor risikoen for forekomst af somatiske kroniske sygdomme er fundet størst. Vi vil redegøre for, hvordan Bourdieus teoretiske begrebsapparat kan udfoldes og nuanceres til at se menneskelige dispositioner og handlen som noget, der også hænger sammen med tilknytning til bestemte fysiske placeringer. Det vil vi gøre ved hjælp af Mike Savage, Gaynor Bagnall og Brian Longhursts teoretiske betragtninger, hvor Savage og kollegers værk tager udgangspunkt i Bourdieus feltbegreb, som også vil forklares i dette afsnit. I afsnittet vil der lægges særlig vægt på, at områder ikke er tilfældige konstellationer og samtidig, at der i områder kan være forskellige normer for sundhedsrelateret adfærd.

Med udgangspunkt i Bourdieus begreb *felt* belyser Savage et al., hvordan nærområder skal forstås som særegne og komplekse størrelser med høj relevans i en globaliseret verden. Bourdieu selv har sjældent talt om globalisering, omend han bruger den rumlige metafor i form af sit feltbegreb (Savage et al. 2005:8). Med udgangspunkt i Bourdieus perspektiv kan feltbegrebet forstås som forskellige miljøer. Eksempelvis nævner Bourdieu "det kunstneriske miljø" (Bourdieu 2000:100). Feltbegrebet skal samtidig forstås som en kampplads, hvor individer fra forskellige sociale positioner konfronterer hinanden og indleder konflikter og konkurrence med hinanden (Bourdieu & Wacquant 1992:17). Uligheden bliver tydelig ved anvendelsen af dette begreb, idet nogle individer har en mere gunstig position end andre i disse konflikter. Personer med den største magt eller kapital er dem, som sætter dagsordenen for, hvilke objekter eller praksisser der er mest efterstræbelsesværdige (Bourdieu 1984:247; Bourdieu & Wacquant 1992:17ff). Ethvert felt indeholder en specifik *doxa*, som han beskriver som: "*a set of fundamental beliefs which does not even need to be asserted in the form of an explicit, self-conscious dogma*" (Bourdieu 2000:15). Doxa er altså udtryk for de normer eller opfattelser, der ubevidst og implicit eksisterer i det specifikke felt, og ethvert felt har sin egen unikke

doxa (Bourdieu 2000:100). Det er altså dette feltbegreb, som Savage et al. belyser også kan forstås som et geografisk afgrænset sted (Savage et al. 2005:8f).

Som vi har gjort rede for i afsnit "7.1. Socioøkonomiske faktorer og sundhedsrelateret adfærd", dækker Bourdieus habitusbegreb over den betragtning, at menneskers dispositioner i livet er kropsligt funderede (Bourdieu 2000:130). Derudfra betragter Savage et al., at menneskers dispositioner også nødvendigvis er knyttet til bestemte lokaliteter ud fra, hvor kroppen rent fysisk befinder sig. Centralt for koblingen mellem habitusbegrebet og fysisk placering er individets følelse af kropslig komfort i specifikke fysiske kontekster. Ved inddragelse af Bourdieus feltbegreb tydeliggør Savage et al., hvordan Bourdieu adskiller sig fra eksistentielle antagelser, der anskuer hjemstavnen som værende af størst betydning for, hvor det enkelte menneske føler sig mest i ét med sine omgivelser, og altså oplever størst kropslig komfort (Savage et al. 2005:9). Med udgangspunkt i kropslig komfort beskriver Savage et al. sammenhængen mellem habitus og felter således:

"People are comfortable when there is a correspondence between habitus and field, but otherwise people feel ill at ease and seek to move – socially and spatially – so that their discomfort is relieved." (Savage et al. 2005:9)

Som det fremgår af citatet, flytter individer sig, hvis de ikke er tilpas, hvor de er. I moderne kapitalistiske samfund bliver felter desuden mere og mere differentierede og separerede fra hinanden, end det tidligere har været tilfældet, og derfor er det enkelte individs behov for at flytte sig til et sted med kropslig komfort ikke blevet mindre. Fordi det enkelte menneske hver især er drevet af sin habitus, er det ikke tilfældigt, hvor det flytter hen, og derfor heller ikke tilfældigt, hvordan de respektive områder er sammensat (Savage et al. 2005:9f).

Denne nuancering af Bourdieus begreber om habitus og felter understøttes desuden af eksisterende forskning om, at vi bosætter os med nogen, vi ligner (jf. problemfeltet). I en dansk kontekst anskues, at individer bosætter sig i homogene enklaver, hvor individer ofte bor i områder med nogle, der ligner dem selv med udgangspunkt i økonomiske og uddannelsesmæssige baggrunde. Denne økonomiske og etniske segregering er forøget

over de seneste tre årtier (Christensen et al. 2015:18). Med udgangspunkt i sociale klasser kan den bosætningsmæssige segregering i Danmark både ses på kommune- og sogneniveau (Juul & Nawa 2012). Fordi menneskers bosætningsmønstre ikke er tilfældige konstellationer, og social position og sundhedsrelateret adfærd hænger sammen (Williams 1995:581ff), kan man også formode, at risikoen for forekomsten af forskellige sygdomme heller ikke er den samme i alle områder. Derfor er det i høj grad relevant at tilføje et geografisk perspektiv til undersøgelsen i dette speciale.

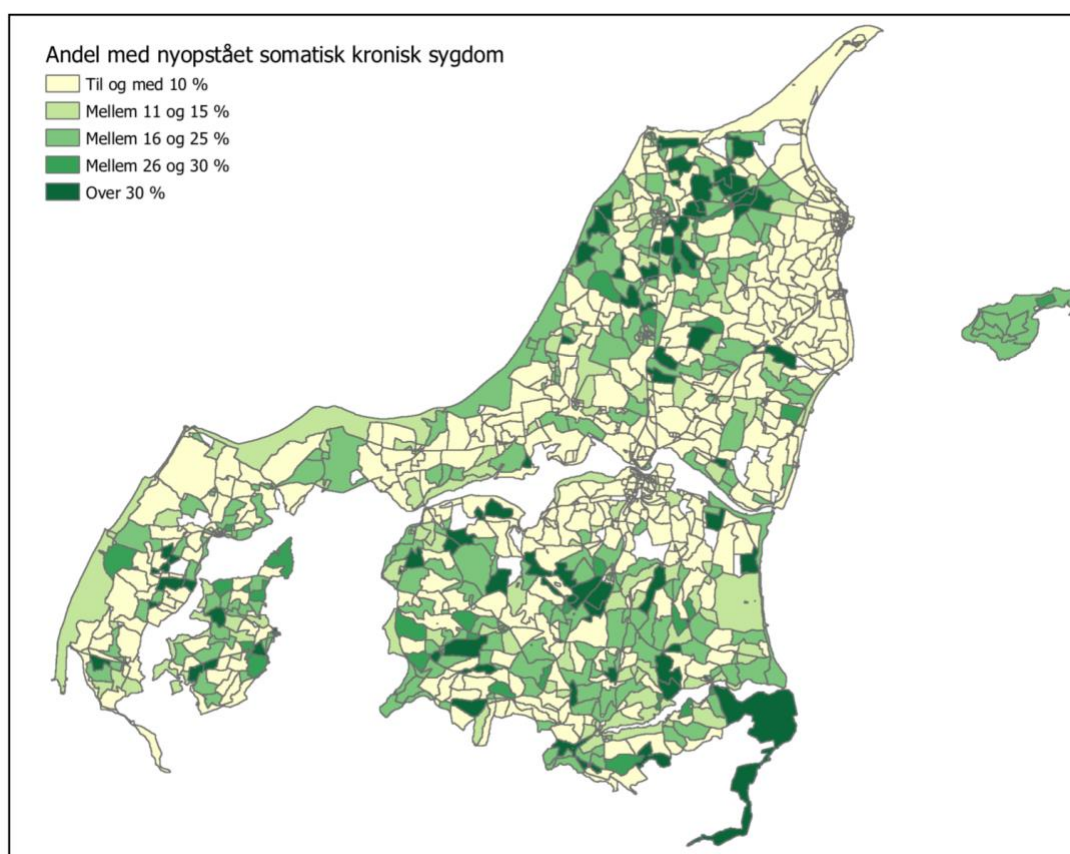
Som vi har nævnt i specialets indledning, kan ulighed i forekomsten af sygdom også skyldes områdeeffekter. Disse effekter kan forklare, hvorfor områder med samme type beboere ud fra socioøkonomiske karakteristika ikke altid er ens, hvad angår andelen med somatiske kroniske sygdomme (jf. problemfeltet). Hvad der konkret er på spil i områderne vil dog ikke undersøges her, da det ikke efterspørges i problemformuleringen. I stedet vil inddelingen bruges til undersøge, hvordan forekomsten af somatiske kroniske sygdomme fordeler sig i regionen.

10. Analyse af geografisk fordeling af sygdom

I dette analyseafsnit vil anden del af specialets problemformulering besvares. Denne del indebærer, hvordan forekomsten af somatiske kroniske sygdomme vil fordele sig geografisk i Region Nordjylland. Vi vil præsentere kortlægningen af nyopståede somatiske kroniske sygdomme for datasættene fra år 2013 og 2017. For datasættet fra år 2013 udnyttes det, at det er muligt at sammenligne den faktiske forekomst af somatiske kroniske sygdomme i perioden 2014-2016 med de prædikterede værdier opnået ved hjælp af den konstruerede algoritme fra samme år. Derfor haves der to kort for dette år, mens der for år 2017 vil vises ét kort i form af de prædikterede værdier. I dette afsnit vil vi først give en deskriptiv beskrivelse af de tre kort, hvorefter der til sidst vil forekomme en samlet teoretisk fortolkning af disse.

Af figur 12 nedenfor ses kortet med andelen med faktisk nyopstået sygdom i årene 2014-2016.

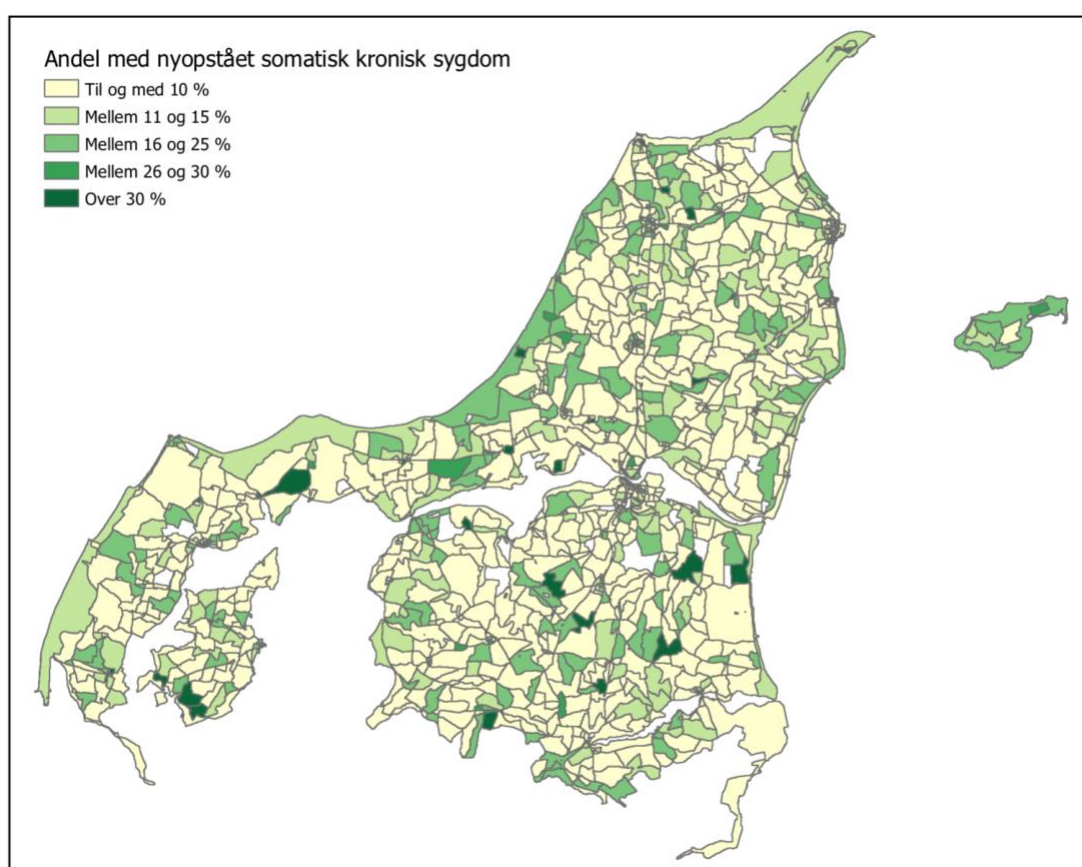
Figur 12: Kortlægning af faktisk forekomst af nyopståede somatiske kroniske sygdomme i 2014-2016.



På dette kort ses det, at områderne med den største andel af personer med en somatisk kronisk sygdom særligt findes i den nordvestlige del af Region Nordjylland. Ligeledes kan det ses, at der i den nordligste del og samtidig den nordøstlige del af Region Nordjylland er en relativt lille andel med nyopstået somatisk kronisk sygdom. Syd for Limfjorden findes der ikke samme mønster med en vest/øst opdeling af områder angående andelen med somatisk kronisk sygdom.

På figur 13 nedenfor ses kortet med de prædikterede værdier fra samme år. Ligesom det blev gjort ved den logistiske regressionsanalyse af de prædikterede sandsynligheder for datasættet for år 2017, er de prædikterede sandsynligheder, som er kortlagt på denne figur, inddelt i to grupper. I den ene indgår de 10 procent, der ifølge modellens prædiktions har størst sandsynlighed for at udvikle en somatisk kronisk sygdom, mens der i den anden indgår de resterende 90 procent af samplet.

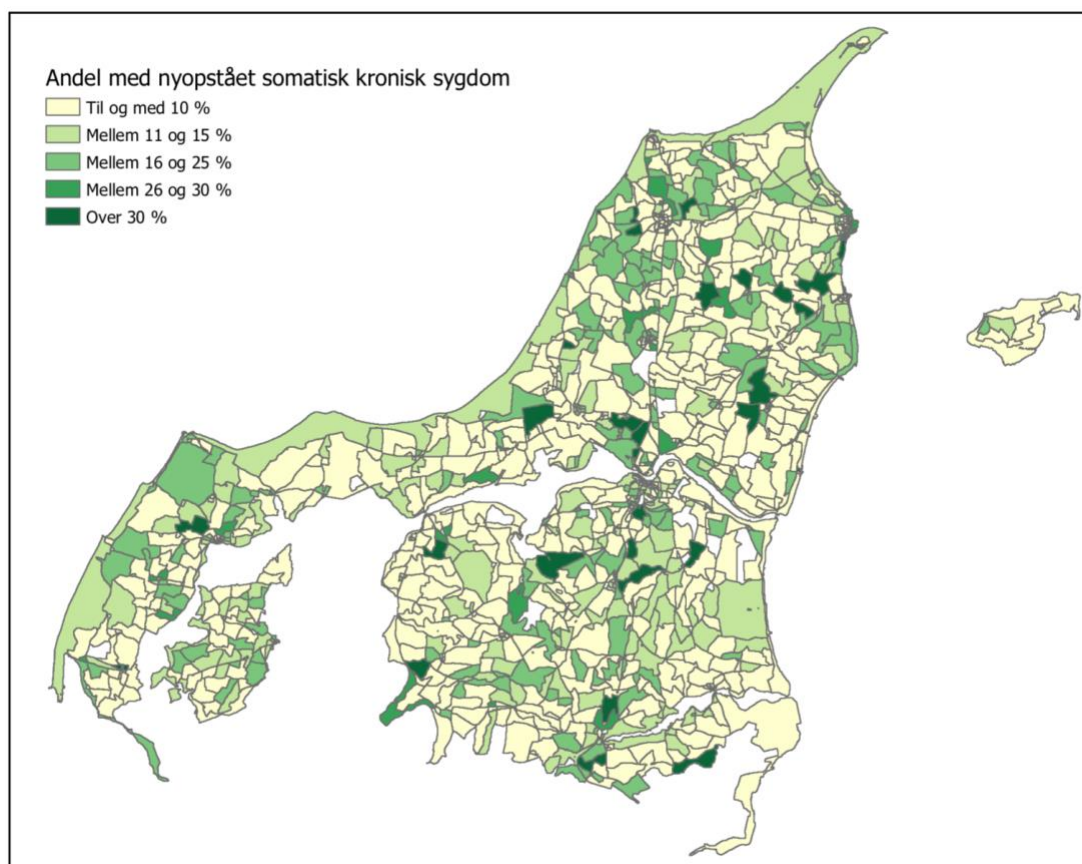
Figur 13: Kortlægning af prædikteret forekomst af nyopståede somatiske kroniske sygdomme i 2014-2016.



Sammenlignet med det foregående kort, kan det ses, at der ikke findes samme opdeling mellem den nordvestlige og den nordøstlige del af Region Nordjylland, og det betragtes, at der generelt ikke findes noget mønster i områderne. Ligeledes kan det ses, at der ikke er lige så mange områder, hvor andelen med nyopstået somatisk kronisk sygdom er henholdsvis mellem 26 og 30 procent, og over 30 procent. Forskellene mellem de to kort kan skyldes, at vores model ikke prædikerer 100 procent korrekt, og sidstnævnte forskel kan samtidig skyldes, at vi ved at se på de prædikterede 10 procent med størst sandsynlighed for at få en somatisk kronisk sygdom ikke indfanger alle de personer, der bliver syge.

Af figur 14 nedenfor ses det prædikterede kort over nyopstået somatisk kronisk sygdom i årene 2018-2020.

Figur 14: Kortlægning af prædikeret forekomst af nyopståede somatiske kroniske sygdomme i 2018-2020.



Som på det prædikterede kort for år 2014-2016 findes ej heller på dette kort et mønster i områderne over andelene, som det forudsiges har særligt stor risiko for at få en somatisk kronisk sygdom. Det anskues, at den største forskel mellem det prædikterede kort for årene 2014-2016 og det prædikerede kort for årene 2018-2020 er, at det tyder på, at andele med somatisk kronisk sygdom fremtidigt er mindre på Læsø.

Med udgangspunkt i Savage et al.s perspektiv på, at individer flytter og bosætter sig alt efter kropslig komfort (Savage et al. 2005:9), anskuer vi, at individerne på tværs af områder i højere grad er heterogene og samtidig i højere grad homogene internt angående deres sociale positioner. Yderligere kan Bourdieus begreb habitus og Williams anvendelse heraf benyttes til at forklare, at det har betydning for andelen med nyopstået somatisk kronisk sygdom, at der er en forskellig social komposition i forskellige områder. Dette skyldes blandt andet, at habitus er afgørende for individers livsstil og herunder sundhedsrelaterede adfærd (Bourdieu 1997:24; Williams 1995:581ff).

Som det ses på de forskellige kort, kan områder med berøringsflader have markant forskellige andele med nyopstået somatisk kronisk sygdom, og ud fra den teoretiske forståelse betyder dette, at der kan være forskellige homogene enklaver relativt tæt på hinanden geografisk set. Med udgangspunkt i Savage et al.s forståelse af nærområder som forskellige felter og samtidig, at hvert felt indeholder en specifik doxa, kan den forskellige andel med kroniske somatiske syge ligeledes skyldes, at der i områderne eksempelvis er forskellige normer for og forståelser af sundhedsrelateret adfærd. Når doxa anvendes på denne måde, kan begrebet kobles med begrebet områdeeffekter, der blev nævnt både i indledningen og afsnit "9. Teori om geografisk segregering". Disse effekter kan eksempelvis netop indebære holdninger til og syn på sundhed i området (Pickett & Pearl 2001:111), og som det blev nævnt i indledningen kan områdeeffekter forklare ulighed i sundhed, og herunder uligheden i nyopstået somatisk kronisk sygdom. Med udgangspunkt heri kan det forklare, hvorfor områder med samme type beboere ud fra socioøkonomiske karakteristika ikke nødvendigvis altid er ens, hvad angår andelen med somatiske kroniske sygdomme. Selvom undersøgelser af områdeeffekter, som eksempelvis arbejdsløshedsprocenten eller graden af forurening i de enkelte områder, kunne være interessante at knytte til områdeuligheden i forekomsten af somatiske kroniske sygdomme, der fremgår af resultaterne af kortlægningerne, indgår det ikke som

en del af problemformuleringen i dette speciale at undersøge determinanter på aggregeret niveau.

11. Diskuterende konklusion

I problemfeltet blev det fremlagt, at der er ulighed i sundhed og herunder ulighed i forhold til somatiske kroniske sygdomme. Samtidig blev det forklaret, at kroniske sygdomme er særligt dyre for sundhedsvæsenet. Når det for Region Nordjylland er ønskværdigt at udføre forebyggende arbejde og planlægning i forhold til dem, der især har stor risiko for at blive særligt dyre for sundhedsvæsenet, blev det fundet relevant at fokusere på disse sygdom. Derudfra opstillede vi denne problemformulering:

Hvilke sociale og psykologiske determinanter på individniveau har størst betydning for sandsynligheden for at udvikle en somatisk kronisk sygdom i fremtiden, og hvordan vil forekomsten af disse fordele sig geografisk i Region Nordjylland?

Til besvarelse af problemformuleringen anvendes både Sundhedsprofilen for år 2010, 2013 og 2017 for Region Nordjylland, hvilket er en spørgeskemaundersøgelse med fokus på trivsel, sundhed og sygdom. Samtidig har vi anvendt registerdata over kroniske sygdomme fra Region Nordjyllands patientadministrative system, og registerdata over alder, indkomst, køn og etnicitet. Fra registerdataene er udvalgt følgende sygdomsgrupperinger: allergiske og inflammatoriske sygdomme, gig, hjertekarsygdomme, KOL eller astma, osteoporose (knogleskørhed) samt diabetes.

I dette afsnit vil vi besvare den todelte problemformulering i to særskilte afsnit. Først vil vi besvare første del af problemformuleringen, der omhandler hvilke sociale og psykologiske determinanter på individniveau, der har størst betydning for sandsynligheden for at udvikle en somatisk kronisk sygdom. I forlængelse heraf vil vi give en vurdering af, hvorvidt de to metoder logistisk regression og neuralt netværk både hver for sig og i kombination kan besvare problemformuleringen. I det efterfølgende afsnit vil vi besvare anden del af problemformuleringen, hvor der er fokus på, hvordan den nyopståede sygdomsforekomst fordeler sig geografisk i Region Nordjylland. I dette afsnit vil der primært være fokus på at besvare, hvordan de prædikterede sandsynligheder for årene 2018-2020 fordeler sig, eftersom det er dette, der efterspørges i problemformuleringen. Disse vil dog sammenlignes med tolkningerne fra de øvrige kort. Tilsammen kan de to dele af problemformuleringen give et nuanceret billede af nyopstået

somatisk kronisk sygdom, hvor det både er undersøgt, hvad der har betydning for, hvem der får en sygdom og samtidig undersøgt, hvor disse personer bor. Det ønskes, at begge dele udgør vigtig information i forhold til at udvikle forebyggende indsatser og sundhedstilbud, der kan være med til både at mindske de økonomiske udgifter til sundhedsområdet samt mindske den sociale ulighed i sundhed. Slutteligt vil der være en afslutning, hvor vi fremlægger idéer til, hvordan fremtidige lignende projekter med fordel kan opbygges.

11.1. Sociale og psykologiske determinanternes betydning

Besvarelsen af første del af specialets problemformulering bygger på konstruktionen af to forskellige modeller, både en logistisk regressionsmodel og en neural netværksmodel. Begge modeller er konstrueret og valideret ved hjælp af datasættet fra år 2010 med tilhørende informationer om forekomsten af somatiske kroniske sygdomme i årene 2011-2014. Dertil er begge modeller testet ved hjælp af datasættet fra år 2013 med tilhørende informationer om somatiske kroniske sygdomme i årene 2014-2016. I dette afsnit vil resultaterne fra de to modeller fremlægges, hvor der tages højde for de metodiske og designmæssige udfordringer, der fremgår af afsnit "8.3. Diskussion af analyseresultater".

I modellen konstrueret ved hjælp af logistisk regression har følgende variable vist sig at være dem, der bedst kan forudsige, hvorvidt en person i fremtiden får en nyopstået somatisk kronisk sygdom: Alder, BMI, rygning, køn samt i hvor høj grad personen er ængstelig eller deprimeret. Mere specifikt fremgår det af modellen, at personer på 99 år 10,3 gange større sandsynlighed for at få en somatisk kronisk sygdom, end det er tilfældet for personer på 16 år. Personer med undersøgelsens højeste BMI har 4,8 gange større sandsynlighed for sygdom end personer med undersøgelsens laveste BMI, hvor det dog skal bemærkes, at ingen af de to yderpunkter betragtes som et udtryk for en "sund" BMI. Sundhedsbetegnelsen er sat i citationstegn, fordi vi tager udgangspunkt i Sundhedsstyrelsens anbefalinger og betragter, at hvad der defineres som sundt og usundt er socialt konstrueret. Ligeledes tyder det på, at personer, der ryger hver dag, har 1,3 gange større sandsynlighed for at blive syge, i forhold til personer, der ikke ryger. Sammenhængen mellem somatisk kronisk sygdom og rygning og højt BMI har vi forklaret

ved hjælp af Bourdieus teoretiske betragtninger om sammenhængen mellem social position og livsstil (Bourdieu 1986; Bourdieu 1997) samt Williams applicering af disse begreber på sundhedsfeltet (Williams 1995). Mere specifikt udleder vi den tolkning, at lavtuddannede har en sundhedsrelateret adfærd, der i højere grad leder til en større risiko for at udvikle disse sygdomme, hvilket kan skyldes en lavere grad af kulturel kapital i både institutionaliseret form (Bourdieu 1986:243), forstået som mangel på viden om sundhed, og kropsliggjort form (ibid.), der knytter sig til forskellige mønstre for sundhedsrelateret adfærd blandt højt- og lavtuddannede. Dertil har vi forklaret disse sammenhænge ud fra begrebet om økonomisk kapital (ibid.), hvor det tolkes, at lavtuddannede individer har et økonomisk incitament for en sundhedsrelateret adfærd, der medfører større risiko for et højt BMI, baseret på argumentet om at "usunde" madvarer ofte er billigere, og medlemskab af sportsklubber eller fitnesscentre er dyre. Den økonomiske forklaring er imidlertid ikke meningsfuld, hvad angår rygning. Ud fra disse tolkninger har vi desuden belyst, at lavtuddannede kan ses som en gruppe, hvis sundhedsrelaterede adfærd der tages afstand til blandt personer i højere social position. Da de højtuddannede tilmed kan betragtes som dem, der oftest fastsætter normerne for passende sundhedsrelateret adfærd gennem eksempelvis KRAM-faktorerne, har vi tolket dette som et udtryk for det, Bourdieu betegner som symbolsk vold (Bourdieu & Passeron 2006:9). Bourdieus teoretiske begrebsapparat er ligeledes anvendt til at forklare odds ratio af kønsvariablen, der angiver, at kvinder har 1,3 gange større sandsynlighed for at udvikle en somatisk kronisk sygdom end mænd. Betragtningerne om det socialt konstruerede køn (Bourdieu & Wacquant 1992:171) leder til den tolkning, at køn kan have en betydning for blandt andet normer i forhold til sundhedsrelateret adfærd og normer i forhold til at opsøge læge. Det betragtes, at begge dele kan have indflydelse på, hvorvidt personen står registreret i det patientadministrative system som havende en somatisk kronisk sygdom. Afslutningsvist fremgår det af modellen, at personer, der føler sig ekstremt ængstelige eller deprimerede, har 1,5 gange større sandsynlighed for at få en sygdom end personer, der ikke har disse følelser. Ud fra den psykosociale forklaring har vi tolket, at det både kan skyldes det direkte forhold mellem individers mentale og fysiske reaktioner, men også den indirekte sammenhæng, hvor dette forhold er medieret af sundhedsrelateret adfærd (Elstad 2000:69ff). Desuden har vi ud fra et klasseperspektiv tolket denne odds ratio, hvor det inden for den psykosociale forklaringsmodel betragtes, at mængden af stressreaktioner er socialt graderet på den

måde, at personer i de lavere sociale grupper både er udsat for flere eksterne stressorer men også har færre ressourcer til at cope med disse (Elstad 2000:74ff).

Formålet med at inddrage den neurale netværksanalyse er, at denne er god til at prædiktere, mens det med en sådan tilgang ikke er muligt at opnå viden om kausale sammenhænge (Kuhn & Johnson 2013:286f; Kreatsoulas & Subramanian 2018:348; Hain & Jurowetzki 2018:25). Derfor har vi suppleret denne model med en logistisk regressionsmodel. Vores neurale netværksmodels samt logistiske regressionsmodels prædiktionssevne vil inddrages i afsnit "11.1.1. Metodernes evne til besvarelse af problemformuleringens første del". Da de prædikterede værdier blev anvendt i en logistisk regressionsanalyse viste det sig, at samtlige af specialets 31 uafhængige variable har en statistisk signifikant effekt på den prædikterede sandsynlighed for at få en somatisk kronisk sygdom i årene 2018-2020, når de inddrages i en samlet model. Dette er dog med undtagelse af de tre variable: respondentens fysiske aktivitetsniveau i fritiden, i hvor høj grad de stoler på personer i deres lokalområde samt, hvor ofte der anvendes fedtstof i madlavningen. Vi valgte at fokusere på tolkningen af følgende 7 variable på baggrund af størrelsen af deres odds ratio: Alder, energiniveau, hvor forsigtig respondenter er i forhold til andre mennesker, etnicitet, uddannelse, køn, samt hvor ofte respondenter føler sig rolig og afslappet. Odds ratio for køns- og aldersvariablen giver i denne analyse anledning til samme teoretiske tolkninger, som i modellen konstrueret ved hjælp af logistisk regression. Dog er størrelsen af disse odds ratio med værdier på henholdsvis 5,7 og 3855,7 bemærkelsesværdige, hvilket især er gældende for størrelsen af odds ratio for aldersvariablen. Odds ratio for variabelen for uddannelse er omvendt for analysen af den prædikterede sandsynlighed, når den sammenlignes med odds ratio for kontrolvariablen i den første model. Samtidig er odds ratioen modsatrettet resultater i eksisterende forskning (Barbeau et al. 2004; Dewalt et al. 2004; Sanders et al. 2009; Leganger & Kraft 2003; Mickelson & Kubzansky 2003; Uchino 2006). Ud fra dette resultat forstås det, at individer med lang videregående uddannelse har større prædikeret sandsynlighed for at få en somatisk kronisk sygdom end individer med grunduddannelse. Dette har vi teoretisk forklaret med, at også højtuddannede kan tilegne sig vaner og sundhedsrelateret adfærd, der har en negativ indvirkning på helbredet, og som er en adfærd, der ikke er tilgængelig eller optaget blandt personer med lavere social position. Således kan Bourdieus betragtninger om både symbolsk vold og sammenhængen mellem

social position og livsstil (Bourdieu & Passeron 2006:9; Bourdieu 1986; Bourdieu 1997) også understøtte forklaringen af det modsatte forhold, end de blev anvendt til at forklare ved resultaterne af den første model. Den modsatrettede kausale sammenhæng blev ligeledes forklaret ud fra et bias, der også blev nævnt i specialets metodeafsnit. Dette bias knytter sig til konstruktionen af specialets afhængige mål for nyopstået somatisk kronisk sygdom, hvor vi sammenligner personer, der får en sygdom med personer, der ikke gør, uanset om det sidstnævnte forhold skyldes, at de har sygdommen i forvejen og derfor ikke kan få den igen, eller om de på intet tidspunkt har den. Det kan være et problem, fordi vi på den måde sidestiller personer, der er raske til begge tidspunkter, med personer der fejler én eller potentielt mange forskellige ting til udgangspunktet. Ligeledes så vi, at odds ratio for variablen, der angiver respondentens energiniveau, er bemærkelsesværdig høj. Denne indikerer, at personer med et højt energiniveau har 26,1 gange større prædikeret sandsynlighed for nyopstået somatisk kronisk sygdom, end det er tilfældet for personer med et lavt energiniveau. Dette resultat har vi ikke kunnet forklare ud fra den psykosociale forklaringsmodel som ventet, eftersom vi ud fra den teoretiske forståelse havde forventet den modsatte sammenhæng. I stedet brugte vi ligeledes forklaringen om det potentielle bias, der knytter sig til specialets afhængige variabel. Variablen for, hvorvidt respondenterne føler sig rolige og afslappede, samt variablen for forsigtighed blev imidlertid forklaret med den psykosociale forklaringsmodel. Af modellen fremgår det, at personer, der ofte føler sig rolige og afslappede, har 0,06 gange prædikeret sandsynlighed for nyopstået sygdom, hvor vi har tolket det således, at personer med denne følelse i lavere grad har stressreaktioner, hvor stressreaktioner ifølge den psykosociale forklaring er af betydning for sygdom (Elstad 2000:69ff). Ligeledes viste denne analyse, at personer, der oplever et højt tillidsniveau til andre mennesker, har 0,09 gange prædikeret sandsynlighed for en ny somatisk kronisk sygdom. Dette tolkede vi ud fra den betragtning inden for den psykosociale forklaringsmodel, at mistillid har en negativ effekt på sundhed (Elstad 2000:79; Wilkinson & Pickett 2010:57; Kawachi et al. 1997:1494f). At personer, der har en anden etnisk oprindelse end dansk, har 0,07 gange prædikeret sandsynlighed for nyopstået sygdom blev forklaret med Bourdieus habitusbegreb, hvor det betragtes, at udformningen af individets habitus blandt andet afhænger af socioøkonomiske faktorer, som eksempelvis etnicitet (Bourdieu 1984:107ff). Igen kan man ud fra dette begreb

således forestille sig, at der hersker forskellige normer og adfærdsmønstre i de to etniske grupper.

På grund af de høje odds ratio værdier i analysen, har vi både i analysen og diskussionen af analysen været inde på risikoen for overfitting af den neurale netværksmodel. Ved at være opmærksomme på den mulige overfitting af den neurale netværksmodel, der potentielt påvirker undersøgelsens validitet, bør variablene i denne model ikke danne grundlag for indsatser, før modellen er korrigeret herfor. Efter korrigering kan det dog forestilles, at andre variable vil være udslagsgivende. Desuden er det svært at konkludere noget på baggrund af variabelen for BMI grundet variabelens kodning.

Hvis vi overser resultaterne af den neurale netværksanalyse, der kan være influeret af overfitting-problematikken, og i stedet blot inddrager resultaterne fra modellen konstrueret alene ved hjælp af logistisk regression, tyder det på, at de determinanter, der har størst betydning for, hvorvidt et individ udvikler en ny somatisk kronisk sygdom, er følgende: alder, uddannelse, køn, rygevaner samt hvorvidt personen føler sig ængstelig eller deprimeret. Alle disse variable indgår i modellen konstrueret ved hjælp af logistisk regression som dem, der øger modellens prædiktionssevne mest muligt i forhold til nye data.

11.1.1. Metodernes evne til besvarelse af problemformuleringens første del

Afslutningsvist for besvarelsen af første del af specialets problemformulering vil vi vurdere i hvilken grad, de to metoder har været anvendelige til besvarelsen af denne og herunder, hvad de hver især kan bidrage med, og hvad de kan bruges til i kombination.

Ud fra specialets centrale diskussion af de forskellige matematiske antagelser og formål, der ligger til grund for to metoder, vurderer vi på trods af risikoen for overfitting i specialets neurale netværksmodel, at det er muligt at drage fordel af et neuralt netværk. Det skyldes, at denne metode tilbyder et output i form af et nuanceret mål for den prædikterede sandsynlighed. Som vi har set, er dette en variabel, hvor værdien varierer fra person til person og umiddelbart virker valid (jf. afsnit "8.3. Diskussion af

analyseresultater"). Et lignende output har vi ikke opnået ved anvendelse af logistisk regression. Derudover har vi opnået en højere prædiktionssevne ved det neurale netværk end ved den logistiske regression, hvor der dog er sået tvivl om, hvorvidt denne er til at regne med, fordi ingen tildeles værdien 1, og derfor vises det ikke, hvor god modellen er til at forudsige, hvem der bliver syge. Dette har den betydning, at det er problematisk at sammenligne den logistiske regressions prædiktionssevne med det neurale netværks. Udfordringen ved det neurale netværk, har vi vurderet, skyldes en skæv fordeling af målet for nyopstået somatisk kronisk sygdom, hvor kun 7 procent af respondenterne, der har besvaret spørgeskemaet i 2010, bliver syge i løbet af 2011-2013. I fremtidige undersøgelser kan dette muligvis løses ved enten en oversampling af minoritetsklassen eller en undersampling af majoritetsklassen, hvormed klasseinddelingen bliver balanceret (Godase & Attar 2012:9f). Hvad den neurale netværksanalyse imidlertid ikke kan tilbyde, er en undersøgelse af de kausale årsagssammenhænge, der er på spil mellem de sociale samt psykologiske determinanter og nyopstået somatisk kronisk sygdom, hvilket logistisk regression til gengæld kan (Kuhn & Johnson 2013:286f; Kreatsoulas & Subramanian 2018:348; Hain & Jurowetzki 2018:25). I forhold til de to metoders evne til hver især at besvare den første del af problemformuleringen betragtes det, at det neurale netværk i sig selv ikke har kunnet bidrage til forståelsen af sammenhængen mellem sociale samt psykologiske determinanter og mulig fremtidig sygdomsforekomst, mens vi ikke har fundet at den logistiske regressionsmodel kan forudsige fremtidige data. Det vurderes, at en kombination af de to metoder udgør et bedre grundlag for at besvare dette undersøgelsesspørgsmål, end den enkelte metode hver for sig er i stand til, såfremt overfitting-problematikken i den neurale netværksmodel løses. På den måde udnyttes det fulde potentiale af den høje generaliserbarhed i neurale netværksmodeller. Efterfølgende kan variabelen for prædikteret sandsynlighed anvendes som afhængig variabel i en logistisk regressionsanalyse til at undersøge statistisk inferens. I dette speciale vises det, hvordan den prædikterede værdi, konstrueret ved hjælp af det neurale netværk, kan udgøre grundlaget for undersøgelsen af de kausale sammenhænge ved anvendelse logistisk regressionsanalyse mellem sociale samt psykologiske determinanter og sandsynligheden for at udvikle en somatisk kronisk sygdom i fremtiden.

11.2. Geografisk fordeling af nyopståede sygdomme

I specialet har vi fremført tre kort over den geografiske fordeling af nyopståede somatiske kroniske sygdomme. Dette er to kort for datasættet for år 2013 og ét kort for datasættet for år 2017. For datasættet for år 2013 har vi udnyttet, at det er muligt at sammenligne den faktiske forekomst af somatiske kroniske sygdomme i perioden 2014-2016 med de prædikterede værdier opnået ved hjælp af den konstruerede model fra samme år. For datasættet for år 2017 vises et kort over de prædikterede værdier, der i dette afsnit vil tages udgangspunkt i, fordi der i problemformuleringen efterspørges et fremtidigt element. Der vil være særligt fokus på kortet for år 2017, men de andre to kort vil inddrages til sammenligning.

Inden kortet for år 2017 blev lavet, konstruerede vi en binær variabel for de prædikterede værdier for somatisk kronisk sygdom, hvor der i den ene kategori indgår de 10 procent, der ifølge modellens prædiktions har størst sandsynlighed for at udvikle en somatisk kronisk sygdom. I den anden kategori indgår de resterende 90 procent af samplet. Det er denne variabel, der blev kortlagt, og dermed andelen af de 10 procent med størst sandsynlighed, der blev set på i kortet. I kortet for år 2017 ses geografisk ulighed i andelen med prædikeret sygdom, men der fremkommer ikke et mønster over placeringen af områder med en relativt stor andel med særligt stor risiko for at få en somatisk kronisk sygdom. Dette er også gældende for kortet over de prædikterede værdier for år 2013, mens kortet over de faktiske værdier for år 2013 i stedet viser en opdeling i den nordlige del af Region Nordjylland, hvor områderne med en stor andel særligt findes i den nordvestlige del af Region Nordjylland, mens der er en relativt lille andel i områderne i den nordøstlige del. Forskellen i de to kort for datasættet fra år 2013 viser, at der er forskel i modellens beregninger og de faktiske værdier.

Ud fra Mike Savage, Gaynor Bagnall og Brian Longhursts teoretiske betragtninger på, at individer flytter og bosætter sig alt efter kropslig komfort (Savage et al. 2005:9), anskuer vi, at individerne på tværs af områder i højere grad er heterogene og samtidig i højere grad homogene internt angående deres sociale positioner. I og med social position og livsstil er sammenhængende (Bourdieu 1997:21ff), kan dette altså have betydning for, hvor stor en andel i et område, der bliver syge. Da der er en stor del af områder med berøringsflader, der har markant forskellige andele med nyopståede somatisk kroniske

sygdomme, betragter vi fra Savage et al.s perspektiv, at der kan være forskellige homogene enklaver, der er bosat relativt tæt på hinanden geografisk set. Ud fra Savage et al.s forståelse af geografiske nærområder som forskellige felter, hvor der til hvert felt er en specifik doxa (Bourdieu 2000:15+100; Savage et al. 2005:8f), kan forskellen mellem områder ligeledes skyldes, at der i forskellige områder for eksempel kan være forskellige normer for sundhedsrelateret adfærd. Der kan altså være forskellige forklaringer på, hvorfor der fremkommer denne ulighed mellem områder i Region Nordjylland.

11.3. Afslutning

I dette afsnit vil vi inddrage mulige ændringer til fremtidige undersøgelser inden for feltet, og dette inkluderer både designændringer og samtidig en yderligere inddragelse af faktorer, der har betydning for, hvem der får en somatisk kronisk sygdom.

I vores undersøgelse er vi kommet frem til forskellige ændringer i designet, der kunne være interessant at tage højde for i fremtidige undersøgelser inden for området. For det første kunne det være relevant blot at fokusere på én somatisk kronisk sygdom. Det skyldes, at der til forskellige sygdomme kan være forskellige udslagsgivende faktorer, og dette kan have betydning for, hvilke sociale og psykologiske determinanter der vises som af særlig betydning i en analyse. På denne måde vil det ikke være muligt at opnå et overordnet mål for somatiske kroniske sygdomme, men det vil i højere grad være muligt at få retvisende værdier for determinanternes betydning. Hvis det skal gøres for hver sygdom vil det være markant mere tidskrævende, men hvis outputtet i højere grad er retvisende, vurderer vi, at det er det værd, fordi det dermed i højere grad vil være muligt at skabe brugbare planlægningsværktøjer til eksempelvis forebyggelse.

En anden ændring til fremtidige undersøgelser kan være, at det kunne være anvendeligt at konstruere de afhængige variable anderledes end det er gjort i vores speciale. Som nævnt har vi sidestillet alle personer der ikke får en specifik sygdom, uanset om det skyldes at de har den i forvejen, eller at de aldrig har haft den. Dette ville være en fordel, fordi det kan diskuteres, hvor brugbart et mål det er, når personer med en eksisterende kronisk sygdom og personer, der aldrig har haft denne sygdom er i samme kategori, fordi de to grupper sandsynligvis er markant forskellige. Denne problematik vil være mest

meningsfuld at løse, hvis de enkelte sygdomme undersøges hver for sig, hvor alle personer, der enten har eller tidligere har haft den pågældende sygdom, udelades fra undersøgelsen. Således ville der have en afhængig variabel, hvor alle personer er raske som udgangspunkt, og hvor "raske" personer, der ikke får sygdommen sammenlignes med "raske" personer, der får sygdommen.

Derudover kunne det i fremtidige undersøgelser være interessant at undersøge betydningen af områdeeffekter, fordi det i eksisterende forskning er vist, at disse har betydning for sundhed i et område (jf. problemfeltet). I denne forbindelse kunne der for eksempel inddrages aggregerede mål for både områdets komposition ud fra socioøkonomiske forhold og kontekstuelle forhold såsom afstand til sygehus og sportspladser. Desuden kunne materialistiske vilkår som for eksempel luftforurening også være vedkommende. På denne måde vil det være muligt at opnå en endnu mere nuanceret model og dermed forståelse af, hvad der har betydning for, hvem der får en somatisk kronisk sygdom. En inddragelse og kortlægning af områdeeffekter kan desuden give analysen af geografisk ulighed i sundhed en endnu mere nuanceret forståelse heraf og samtidig yderligere understrege relevansen af at undersøge områdeulighed.

12. Litteraturliste

Acock, Alan C. (2014): *A Gentle Introduction to Stata*. Stata Press.

Agresti, Alan & Barbara Finlay (2009): *Statistical Methods for the Social Sciences*. Pearson Education Ltd.

Agresti, Alan & Barbara Finlay (2014): *Statistical Methods for the Social Sciences*. Pearson Education Ltd.

Almdal, Thomas (2017): "Type 2-diabetes".

Link: <https://www.sundhed.dk/borger/patienthaandbogen/hormoner-og-stofskifte/sygdomme/diabetes-type-2-hvad-er-det/type-2-diabetes/>. Senest opdateret 16. januar 2017. Tilgået 1. juni 2019.

Almedom, A. M. (2005): "Social capital and mental health: an interdisciplinary review of primary evidence". *Social Science & Medicine*. Vol. 61(5), pp. 943–964.

Andersen, Lotte Bøgh (2012): "Forskningskriterier", i: Lotte Bøgh Andersen, Kasper Møller Hansen & Robert Klemmesen (red.): *Metoder i statskundskab*. Hans Reitzels Forlag.

Avendano, M. & M. M. Glymour (2008): "Stroke disparities in older Americans: Is wealth a more powerful indicator of risk than income and education?" *Stroke*. Vol. 39, pp. 1533–1540.

Bagheri, A.; H. Midi & A. H. M. R Imon (2010): "The Effect of Collinearity-influential Observations on Collinear Data Set: A Monte Carlo Simulation Study". *Journal of Applied Sciences*. Vol. 10, pp. 2086-2093.

Barbeau, E.; N. Krieger & M. J. Soobader (2004): "Working class matters: socioeconomic disadvantage, race/ethnicity, gender, and smoking in NHIS 2000". *American Journal of Public Health*. Vol. 94, pp. 269–278.

Bertsimas, D.; M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala & G. Wang (2008): "Algorithmic Prediction of Health-Care Costs". *Operations Research for Health Care*. Vol. 56(6), pp. 1382-1392.

Bourdieu, Pierre (1984): *Distinction. A Social Critique of the Judgement of Taste*. Havard University Press.

Bourdieu, Pierre (1986): "The Forms of Capital", in John Richardson (ed.): *Handbook of Theory and Research for the Sociology of Education*. New York: Greenwood Press.

Bourdieu, Pierre (1997): *Af praktiske grunde*. København: Hans Reitzels Forlag.

Bourdieu, Pierre (2000): *Pascalian Meditations*. Cambridge: Polity Press.

Bourdieu, Pierre & Jean-Claude Passeron (2006): *Reproduktionen – bidrag til en teori om uddannelsessystemet*. København: Hans Reitzels Forlag

Bourdieu, Pierre & Loïc J. D. Wacquant (1992): *An Invitation to Reflexive Sociology*. Cambridge: Polity Press.

Braveman P.; S. Egerter & R. Mockenhaupt (2010): "Broadening the focus: the need to address the social determinants of health". *American Journal of Preventive Medicine*. Vol. 40(1), pp. 4-18.

Braveman, P.; C. Cubbin, S. Egerter, S. Chideya, K. Marchi et al. (2005): "Socioeconomic status in health research: One size does not fit all". *The Journal of the American Medical Association*. Vol. 294, pp. 2879–2888.

Braveman, P.; S. Egerter & D. R. Williams (2011): "The Social Determinants of Health: Coming of Age". *The Annual Review of Public Health*. Vol. 32, pp. 381–398.

Brownlee, Jason (2013): "A Tour of Machine Learning Algorithms".
Link:<https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>. Udgivet 25. november 2013. Tilgæt 12. marts 2019.

Brownlee, Jason (2016a): "Logistic Regression for Machine Learning". Link:
<https://machinelearningmastery.com/logistic-regression-for-machine-learning/>. Udgivet 1. april 2016. Tilgæt 2. juni 2019.

Brownlee, Jason (2016b): "What is Deep Learning?"
Link:<https://machinelearningmastery.com/what-is-deep-learning/>. Udgivet 16. august 2016. Tilgæt 12. maj 2019.

Brownlee, Jason (2017a): "Difference Between Classification and Regression in Machine Learning". Link:<https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>. Udgivet 11. december 2017. Tilgæt 15. marts 2019.

Brownlee, Jason (2017b): "Gentle Introduction to the Adam Optimization Algorithm for Deep Learning". Link:<https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning>. Udgivet 3. juli 2017. Tilgæt 1. maj 2019.

Brownlee, Jason (2018a): "How to Configure the Number of Layers and Nodes in a Neural Network". Link:<https://machinelearningmastery.com/how-to-configure-the-number-of-layers-and-nodes-in-a-neural-network/>. Udgivet 27. juli 2018. Tilgæt 3. maj 2019.

Brownlee, Jason (2018b): "What is the Difference Between a Batch and an Epoch in a Neural Network?". Link:<https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/>. Udgivet 20. juli 2018. Tilgæt 1. maj 2019.

Brownlee, Jason (2019): "A Gentle Introduction to the Rectified Linear Unit (ReLU) for Deep Learning Neural Networks". Link:<https://machinelearningmastery.com/rectified-linear->

[activation-function-for-deep-learning-neural-networks/](#). Udgivet 9. januar 2019. Tilgået 1. maj 2019.

Bury, Mike (2013): "Medical Model", i: J. Gabe & L. F. Monaghan (eds.): *Key Concepts in Medical Sociology*. SAGE Publications Ltd.

Carter, B. D.; A. C. Christian, D. Fishkanich et al. (2015): "Smoking and mortality— beyond established causes". *The New England Journal of Medicine*. Vol. 372(7), pp. 631–640.

Chandola, T.; E. Brunner & M. Marmot (2006): "Chronic stress at work and the metabolic syndrome: prospective study". *BMJ*. Vol. 332, pp. 521–525.

Charlson M.; M. Wells, R. Ullman, F. King, C. Shmukler (2014): "The Charlson Comorbidity Index Can Be Used Prospectively to Identify Patients Who Will Incur High Future Costs". *PLoS ONE*. Vol. 9(12), pp. 1-16.

Christensen, A. I.; M. Severin, T. Holmberg, L. Eriksen, M. Toftager, A. Zachariassen, O. Ekholm, J. S. Tolstrup, M. Grønbaek & T. Curtis (2009): *KRAM-undersøgelsen i tal og billeder*. København: Statens Institut for Folkesundhed, Syddansk Universitet.

Christensen, Gunvor; Rune C. H. Jørgensen & Malene R. Larsen (2015): *Erfaringer med at ændre socialt mix i udsatte boligområder - evaluering af brugen af anvisnings- og udlejningsredskaber som led i landsbyggefondens 2006-10-midler*. København: SFI - Det Nationale Forskningscenter for Velfærd.

Clausen, Jens & Uffe Rasmussen (2012): "algoritme". Link: [http://denstoredanske.dk/It, teknik og naturvidenskab/Informatik/Software, programmering, internet og webkommunikation/algoritme](http://denstoredanske.dk/It,_teknik_og_naturvidenskab/Informatik/Software,_programmering,_internet_og_webkommunikation/algoritme). Senest revideret 26. marts 2012. Tilgået 25. marts 2019.

Cockerham, Williams C. (2005): "Health Lifestyle Theory and the Convergence of Agency and Structure". *Journal of Health and Social Behavior*. Vol. 46, pp. 51-67.

Cockerham, Williams. C.; B. W. Hamby & G. R. Oates (2017): "The Social Determinants of Chronic Disease". *American Journal of Preventive Medicine*. Vol. 52(1), pp. 5-12.

Collin, Finn (2015): "Socialkonstruktivisme", i Michael Hviid Jacobsen, Kasper Lippert-Rasmussen & Peter Nedergaard (red.): *Videnskabsteori*. Hans Reitzels Forlag.

Daly, M. C.; G. J. Duncan, P. McDonough & D. R. Williams (2002): "Optimal indicators of socioeconomic status for health research". *American Journal of Public Health*. Vol. 92, pp. 1151–1157.

Danmarks Statistik (2019a): "Folketal". Link: <https://www.dst.dk/da/Statistik/emner/befolkning-og-valg/befolkning-og-befolkningsfremskrivning/folketal>. Tilgået 12. maj 2019.

Danmarks Statistik (2019b): "Flytninger internt i Danmark".

Link:<https://www.dst.dk/da/Statistik/emner/befolkning-og-valg/flytninger/flytninger-internt-i-danmark>. Tilgået 12. maj 2019.

Davye Smith, G.; C. Hart, G. Watt et al. (1998): "Individual social class, area-based deprivation, cardiovascular disease risk factors, and mortality: the Renfrew and Paisley study." *Journal of Epidemiology and Community Health*. Vol. 52, pp. 399-405.

de Jonge, J.; H. Bosma, R. Peter & J. Siegrist (2000): "Job strain, effort-reward imbalance and employee well-being: a large-scale cross-sectional study". *Social Science & Medicine*. Vol. 50, pp. 1317-1327.

de Vaus, David (2001): *Research Design in Social Research*. SAGE Publications Ltd.

de Vaus, David (2014): *Surveys in Social Research*. Routledge.

de Veaux R. D. & L. H. Ungar (1994): "Multicollinearity: A tale of two nonparametric regressions". In: P. Cheeseman & R. W. Oldford (eds.): *Selecting Models from Data. Lecture Notes in Statistics, vol 89*. New York: Springer.

Dewalt, D. A.; N. D. Berkman, S. Sheridan, K. N. Lohr & M. P. Pignone (2004): "Literacy and health outcomes: a systematic review of the literature". *The Journal of General Internal Medicine*. Vol. 19, pp. 1228-39.

Diez-Roux A. V.; F. J. Nieto, C. Muntaner, H. A. Tyroler et al. (1997): "Neighborhood environments and coronary heart disease: a multilevel analysis". *American Journal of Epidemiology*. Vol. 146(1), pp. 48-63.

Domingos, Pedro (2012): "A Few Useful Things to Know About Machine Learning". *Communications of the ACM*. Vol. 55(10), pp. 78-87.

Donoho, David (2017): "50 Years of Data Science". *Journal of Computational and Graphical Statistics*, vol. 26(4), pp. 745-766.

Dumancas, G. G. & G. A. Bello (2015): "Comparison of machine-learning techniques for handling multicollinearity in big data analytics and high-performance data mining". Conference-paper. Supercomputing 2015: The International Conference for High Performance Computing, Networking, Storage, and Analysis. Austin, Texas.

Educba (2019): "Machine Learning vs Predictive Modelling – 8 Awesome Differences". Link: <https://www.educba.com/machine-learning-vs-predictive-modelling/>. Tilgået 25. Marts 2019.

Elstad, Jon Ivar (2000): *Social inequalities in health and their explanations*. NOVA - Norwegian Social Research. NOVA Rapport 9/00.

Eskildsen, Anita; Vita L. Dalgaard, Kent J. Nielsen, Johan H. Andersen, Robert Zachariae, Lis R. Olsen, Anders Jørgensen, David H. Christiansen (2015): "Cross-cultural adaptation and

validation of the Danish consensus version of the 10-item Perceived Stress Scale". *Scandinavian Journal of Work, Environment & Health*. Vol. 41(5), pp. 486-490.

Evans, G. W. & P. Kim (2007): "Childhood poverty and health: cumulative risk exposure and stress dysregulation". *Psychological Science*. Vol. 18, pp. 953-957.

Flachs, Esben M.; Louise Eriksen, Mette B. Koch, Julie T. Ryd, Emily Dibba, Lise Skov-Ettrup & Knud Juel (2015): *Sygdomsbyrden i Danmark – sygdomme*. Statens Institut for Folkesundhed, Syddansk Universitet. København: Sundhedsstyrelsen.

Fleishman, J. A. & J. W. Cohen (2010): "Using information on clinical conditions to predict high-cost patients". *Health services research*. Vol. 45, pp. 532-52.

Fortin, M.; G. Bravo, C. Hudon et al. (2006): "Psychological distress and multimorbidity in primary care". *The Annals of Family Medicine*. Vol. 4, pp. 417-422.

Galobardes, B.; G. D. Smith & J. W. Lynch (2006): "Systematic review of the influence of childhood socioeconomic circumstances on risk for cardiovascular disease in adulthood". *Annals of Epidemiology*. Vol. 16, pp. 91-104.

Godase, Abhijeet B. & Vahida Attar (2012): "Classification of Data Streams with Skewed Distribution". ACM International Conference Proceeding Series. *ResearchGate*.

Gou Z. & C. Fyfe (2004): "A canonical correlation neural network for multicollinearity and functional data". *Neural Networks*. Vol. 17, pp. 285-293.

Green, S. B.; J. G. Marquis, S. L. Hershberger, M. S. Thompson & K. M. McCollam (1999): "The overparameterized analysis of variance model". *Psychological Methods*. Vol. 4(2), pp. 214-233.

Hain, Daniel & Roman Jurowetzki (2018): "The Potentials of Machine Learning and Big Data in Entrepreneurship Research - The Liaison of Econometrics and Data Science". Manuscript prepared for the *Handbook of Quantitative Research Methods in Entrepreneurship* (Edward Elgar Publishing).

Hastie, Trevor; Robert Tibshirani & Jerome Friedman (2017): "Boosting and Additive Trees", in: *The Elements of Statistical Learning*. Springer.

Herd, P.; B. Goesling & J. S. House (2007): "Socioeconomic position and health: the differential effects of education versus income on the onset versus progression of health problems". *Journal of Health and Social Behaviour*. Vol. 48, pp. 223-238.

Hjort, Poul G. (2017). "Den effektive algoritme". Link: <https://videnskab.dk/teknologi-innovation/den-effektive-algoritme>. Udgivet 27. juli 2017. Tilgængeligt 12. marts 2019.

Huntley, A.; R. Johnson, S. Purdy, J. M. Valderas & C. Salisbury (2012): "Measures of Multimorbidity and Morbidity Burden for Use in Primary Care and Community Settings: A Systematic Review and Guide". *Annals of Family Medicine*. Vol. 10(2), pp. 134-141.

Hutson, Matthew (2018): "Artificial intelligence faces reproducibility crisis". *Science*. Vol. 359(6377), pp. 725-726.

Jensen, B. (2005): "Sundhedsfremme og forebyggelse - to forskellige paradigmer?". *Tidsskrift for Forskning I Sygdom Og Samfund*. Vol. 2(3), pp. 67-87.

Jordan M. I. & T. M. Mitchell (2015): "Machine learning: Trends, perspectives, and prospects". *Science*. Vol. 349(6245), pp. 255-260.

Juul, Jonas S. & Samira Nawa (2012): *Bopæl og bolig for de sociale klasser*. Arbejdernes Erhvervsråd.

Kawachi, I.; N. E. Adler & W. H. Dow (2010): "Money, schooling, and health: mechanisms and causal evidence". *Annals of the New York Academy of Sciences*. Vol. 1186, pp. 56-68.

Kawachi, Ichiro; Bruce P. Kennedy, Kimberly Lochner & Deborah Pothrow-Stith (1997): "Social Capital, Income Inequality and Mortality". *American Journal of Public Health*. Vol. 9, pp. 1491-1498.

Kim, Y.; S. Lee, M. Kwon, A. Na, Y. Choi et al. (2015): "Developing cancer prediction model based on stepwise selection by AUC measure for proteomics data". IEEE International Conference on Bioinformatics and Biomedicine (BTBM). Dato for konference: 9-12 november 2015. Uploadet til *IEEE Xplore*: 17. december 2015.

Kingma, Diederik P. & Jimmy L. Ba (2015): "Adam: A Method for Stochastic Optimization". Conference Paper published at ICLR (International Conference on Learning Representations), maj 2015.

Kotsiantis, S. B. (2007): "Supervised Machine Learning: A Review of Classification Techniques". *Informatica*. Vol. 31(3), pp. 249-268.

Kotsiantis, S. B.; I. D. Zaharakis & P. E. Pintelas (2006): "Machine learning: a review of classification and combining techniques". *Artificial Intelligence Review*. Vol. 26(3), pp. 159-190.

Kreatsoulas, Catherine & S. V. Subramanian (2018): "Machine learning in social epidemiology: Learning from experience". *Social Science & Medicine – Population Health*. Vol. 4, pp. 347-349.

Kuhn, Max & Kjell Johnson (2013): *Applied Predictive Modeling*. Springer.

Kyed, Vickie H.; Sisse B. Johnsen, Martin Mejlby & Anker L. Vinding (2018): *Hvordan har du det? Sundhedsprofil for Nordjylland 2017*. Region Nordjylland.

Lando, J.; S. M. Williams, B. Williams & S. Sturgiset (2006): "A Logic Model for the Integration of Mental Health Into Chronic Disease Prevention and Health Promotion". *Preventing Chronic Disease*. Vol. 3(2), pp. 1-4.

Lange, Peter (2015): "KOL".

Link:<https://www.sundhed.dk/borger/patienthaandbogen/lunger/sygdomme/kol/kol/>.

Senest opdateret 25. august 2015. Tilgængeligt 1. juni 2019.

Larson, K. & N. Halfon (2009): "Family income gradients in the health and health care access of US children". *Maternal and Child Health Journal*. Vol. 4(3), pp. 332–342.

Leganger, A. & P. Kraft (2003): "Control constructs: Do they mediate the relation between educational attainment and health behavior?" *Journal of Health Psychology*. Vol. 8, pp. 361–372.

Lu, J.; E. Britton, J. Ferrance, E. Rice, A. Kuzel & A. Dow (2015): "Identifying Future High Cost Individuals within an Intermediate Cost Population". *Quality in Primary Care*. Vol. 23(6), pp. 318–326.

Lund, Rolf Lyneborg (2018): "From the dark end of the street to the bright side of the road: Automated redistricting of areas using physical barriers as dividers of social space". *Methodological Innovations*. Vol.11(3), pp: 1–15.

Mackenbach, J. P. & P. Howden-Chapman (2003): "New perspectives on socioeconomic inequalities in health". *Perspectives in Biology and Medicine*. Vol. 46, pp. 428–444.

Maidman, A. & L. Wang (2018): "New Semiparametric Method for Predicting High-Cost Patients". *Biometrics*. Vol. 74, pp. 1104–1111.

Marmot, M. (1999): "Epidemiology of socioeconomic status and health: Are determinants within countries the same as between countries?" *Annals of the New York Academy of Sciences*. Vol. 896, pp. 16–29.

Marmot, M. (2005): "Social determinants of health inequalities". *Lancet*. Vol. (365), pp. 1099–1104.

Marmot, M.; H. Bosma, H. Hemingway, E. Brunner & S. Stansfeld (1997): "Contribution of job control and other risk factors to social variations in coronary heart disease incidence". *Lancet*. Vol. 350, pp. 235–239.

McEwen, B. S. & P. J. Gianaros (2010): "Central role of the brain in stress and adaptation: links to socioeconomic status, health, and disease". *Annals of the New York Academy of Sciences*. Vol. 1186, pp. 190–222.

McPhail, S. M. (2016): "Multimorbidity in chronic disease: impact on health care resources and cost". *Risk Management and Healthcare Policy*. Vol. 9, pp. 143–156.

Mickelson, K. D. & L. D. Kubzansky (2003): "Social distribution of social support: the mediating role of life events". *The American Journal of Community Psychology*. Vol. 32, pp. 265–281.

Møhl, Bo & Peter la Cour (2008): "Nogle kritiske vinkler på EBM-logikken". *Psyke & Logos*. Vol. 29, pp. 126–151.

Molina, Mario & Filiz Garip (2019): "Machine Learning for Sociology". Forthcoming in: *Annual Review of Sociology*. Vol. 45.

Moturu, S. T.; W. G. Williams & H. Liu (2013): "Predictive risk modelling for forecasting high-cost patients: a real-world application using Medicaid data". *International Journal of Biomedical Engineering and Technology*. Vol. 3, pp. 114-132.

Murphy, Kevin P. (2012): *Machine Learning: A Probabilistic Perspective*. Massachusetts Institute of Technology: The MIT Press.

Nationalt Videnscenter for Demens (2019): "Viden om demens".
Link:<http://www.videnscenterfordemens.dk/viden-om-demens/>. Tilgået 13. marts 2019.

Navlani, Avinash (2018): "Understanding Logistic Regression in Python". Link:
<https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python#comments>. Udgivet 7. september 2018. Tilgået 25. marts 2019.

Nettleton, Sarah (2013): *The Sociology of Health and Illness*. Cambridge: Polity Press.

Nielsen, L.; T. Curtis, T. S. Kristensen et al. (2008): "What characterizes persons with high levels of perceived stress in Denmark? A national representative study". *Scandinavian Journal of Public Health*. Vol. 36, pp. 369-379.

Nielsen, Mads (2010): "Machine learning – automatisk læring med computere", side 92-103 i: Datalogisk Institut, Københavns Universitet (DIKU): *Den digitale revolution – fortællinger fra datalogiens verden*. Bogen er udgivet af Datalogisk Institut, Københavns Universitet (DIKU) i anledning af instituttets 40 års jubilæum.

Oakes, J. M.; K. E. Andrade, I. M. Biyoow & L. T. Cowan (2015): "Twenty Years of Neighborhood Effect Research: An Assessment". *Current Epidemiology Reports*. Vol. 2, pp. 80-87.

Orueta, J. F.; R.Nuño-Solinis, M. Mateos, I. Vergara, G. Grandes & S. Esnaola (2013): "Predictive risk modelling in the Spanish population: a cross-sectional study". *BMC Health Services Research*. Vol. 13(269), pp. 1-9.

Pathirana, Thanya & Caroline A. Jackson (2018): "Socioeconomic status and multimorbidity: a systematic review and meta-analysis". *Australian and New Zealand Journal of Public Health*. Vol. 42, pp. 186-194.

Pentecost C. & A. Taket (2011): "Understanding exercise uptake and adherence for people with chronic conditions: a new model demonstrating the importance of exercise identity, benefits of attending and support". *Health Education Research*. Vol. 26(5), pp. 908-922.

Perkins, A; K. Kroenke, J. Unützer, W. Katon, J. Williams, C. Hope & C. Callahan (2004): "Common comorbidity scales were similar in their ability to predict health care costs and mortality". *Journal of Clinical Epidemiology*. Vol. 57, pp. 1040-1048.

Pickett, Kate E. & Michelle Pearl (2001): "Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review". *The Journal of Epidemiology and Community Health*. Vol. 55, pp. 111-122.

Prior, A.; M. Fenger-Grøn, K. K. Larsen et al. (2016): "The association between perceived stress and mortality among people with multimorbidity: a prospective population-based cohort study". *American Journal of Epidemiology*. Vol. 184, pp. 199-210.

Psykiatrifonden (2019): "Misbrug".

Link:<http://www.psykiatrifonden.dk/viden/diagnoser/misbrug-af-alkohol-og-stoffer/misbrug.aspx>. Tilgået 13. marts 2019.

Raschka, Sebastian & Vahid Mirjalili (2017): *Python Machine Learning*. Packt Publishing Ltd.

Region Nordjylland (2011): *Sundhedsprofil 2010 - Trivsel, sundhed og sygdom i Nordjylland*. Rapport. Region Nordjylland. Kan findes på www.rn.dk.

Region Nordjylland (2014): *Sundhedsprofil 2013 – Trivsel, sundhed og sygdom i Nordjylland*. Rapport. Region Nordjylland. Kan findes på www.rn.dk.

Region Nordjylland (2018): *Hvordan har du det? – Sundhedsprofil for Nordjylland 2017*. Rapport. Region Nordjylland. Kan findes på www.rn.dk.

Region Nordjylland (2019): "Baggrund. Formålet med sundhedsprofilerne".

Link:<https://www.rn.dk/sundhed/til-sundhedsfaglige-og-samarbejdspartnere/folkesundhed/sundhedsprofil/tidligere-undersogelser>. Tilgået 25. marts 2019.

Reijneveld, S. (1998): "The impact of individual and area characteristics on urban socioeconomic differences in health and smoking." *International Journal of Epidemiology*. Vol. 27, pp. 33-40.

Robert, S. (1998): "Community-level socioeconomic status effects on adult health." *Journal of Health and Social Behaviour*. Vol. 39, pp. 18-37.

Rose, Nikolas (2001): "The Politics of Life Itself". *Theory, Culture & Society*. Vol. 18(6), pp. 1-30.

Ruder, Sebastian (2017): "An overview of gradient descent optimization algorithms". Link: <https://arxiv.org/abs/1609.04747>. Tilføjet 15. september 2016. Senest revideret 15. juni 2017. Tilgået 3. maj 2019.

Ruizendaal, R. (2017): "Deep Learning #3: More on CNNs & Handling Overfitting". Link: <https://towardsdatascience.com/deep-learning-3-more-on-cnns-handling-overfitting-2bd5d99abe5d>. Udgivet 12. maj 2017. Tilgået 22. maj 2019.

Saito, Takaya & Marc Rehmsmeier (2015): "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". *PLoS ONE*. Vol. 10(3), pp. 1-21.

Sampson, Robert J. (2008): "Moving to Inequality: Neighborhood Effects and Experiments Meet Social Structure". *American Journal of Sociology*. Vol. 114(1), pp. 189-231.

Sampson, Robert J. & Patrick Sharkey (2008): "Neighborhood Selection and the Social Reproduction of Concentrated Racial Inequality". *Demography*. Vol. 45(1), pp. 1-29.

Sanders, L. M.; S. Federico, P. Klass, M. A. Abrams & B. Dreyer (2009): "Literacy and child health: a systematic review". *The Archives of Pediatrics & Adolescent Medicine*. Vol. 163, pp. 131-140.

Savage, Mike; Gaynor Bagnall & Brian Longhurst (2005): *Globalization and Belonging*. SAGE Publications Ltd.

Schafer, I.; H. Hansen, G. Schon, S. Hofels, A. Altiner, A. Dahlhaus et al. (2012): "The influence of age, gender and socio-economic status on multimorbidity patterns in primary care. First results from the multicare cohort study". *BMC Health Services Research*. Vol. 12, p. 89.

Scott, K.M.; R. Bruffaerts, A. Tsang et al. (2007): "Depression-anxiety relationships with chronic physical conditions: Results from the World Mental Health surveys". *Journal of Affective Disorders*. Vol. 103, pp. 113-120.

Seligman, Benjamin; Shripad Tuljapurkar & David Rehkopf (2017): "Machine learning approaches to the social determinants of health in the health and retirement study". *Social Science & Medicine – Population Health*. Vol. 4, pp. 95-99.

Sharma, Sagar (2017): "Activation Functions in Neural Networks". Link: <https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>. Udgivet 6. september 2017. Tilgæet 3. maj 2019.

Shenas, S. A. I.; B. Raahemi, M. H. Tekieh & C. Kuziemyky (2014): "Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes". *Computers in Biology and Medicine*. Vol. 53, pp. 9-18.

Skivington, K.; S. V. Katikireddi, A. H. Leyland, K. Hunt & S. Mercer (2015): "Risk factors for multimorbidity: A multilevel analysis of a longitudinal cohort from Scotland". *European Journal of Public Health*. Vol. 25(3), pp. 13-14.

South, Scott J. & Kyle D. Crowder (1997): "Escaping Distressed Neighborhoods: Individual, Community, and Metropolitan Influences". *American Journal of Sociology*. Vol. 102(4), pp. 1040-1084.

Stansfeld, S. A.; E. G. S. Rael, J. Head, M. Shipley & M. Marmot (1997): "Social support and psychiatric sickness absence: a prospective study of British civil servants". *Psychological Medicine*. Vol. 27, pp. 35-48.

Stansfeld, S.; H. Bosma, H. Hemingway & M. Marmot (1998a): "Psychosocial work characteristics and social support as predictors of SF-36 health functioning: The Whitehall II Study". *Psychological Medicine*. Vol. 60, pp. 247–255.

Stansfeld, S. A.; J. Head & M. Marmot (1998b): "Explaining social class differences in depression and well-being". *Social Psychiatry and Psychiatric Epidemiology*. Vol. 33, pp 1–9.

Stansfeld, S.; M. Shipley & M. Marmot (1999): "Work characteristics predict psychiatric disorders:prospective results from the Whitehall II study." *Occupational and Environmental Medicine*. Vol. 56(5), pp. 302–307.

Stansfeld, S.; R. Fuhrer, M. Shipley et al. (2002): "Psychological distress as a risk factor for coronary heart disease in the Whitehall II Study". *The International Journal of Epidemiology*. Vol. 31, pp. 248–55.

StataCorp (2017): *Stata Multiple-Imputation Reference Manual. Release 15*. Statistical Software. College Station, TX: StataCorp LLC.

Steptoe, A. & M. Marmot (2002): "The role of psychobiological pathways in socio-economic inequalities in cardiovascular disease risk". *European Heart Journal*. Vol. 23, pp. 13–25.

Sundhedsdatastyrelsen (2018): *Profiler for patienter, der står for de højeste sundhedsudgifter i 2017*. Sundhedsdatastyrelsen.

Sundhedsstyrelsen (2017): "Kronisk sygdom". Link:<https://www.sst.dk/da/sygdom-og-behandling/kronisk-sygdom>. Senest opdateret 15. september 2017. Tilgået 27. april 2019.

Sundhedsstyrelsen (2018a): "Forebyggelsespakke - Overvægt". Rapport over retningslinjer og anbefalinger, udgivet 20. april 2018.

Sundhedsstyrelsen (2018b): "Forebyggelsespakke - Alkohol". Rapport over retningslinjer og anbefalinger, udgivet 20. april 2018.

Sundhedsstyrelsen (2018c): "Forebyggelsespakke - Tobak". Rapport over retningslinjer og anbefalinger, udgivet 20. april 2018.

Sundhedsstyrelsen (2019): "Anbefalinger om fysisk aktivitet". Link:<https://www.sst.dk/da/Viden/Fysisk-aktivitet/Anbefalinger-om-fysisk-aktivitet>. Senest opdateret 13. maj 2019. Tilgået 30. maj 2019.

Sundhedsstyrelsen & Fødevarestyrelsen (2018): "Forebyggelsespakke - Mad & Måltider". Rapport over retningslinjer og anbefalinger, udgivet 20. april 2018.

Tamang, S.; A. Milstein, H. Sørensen, L. Pedersen, L. Mackey, J. Betterton, L. Janson & N. Shah (2017): "Predicting patient 'cost blooms' in Denmark: a longitudinal population-based study". *BMJ Open*. Vol. 7, pp. 1-10.

Treiman, Donald J. (2009): *Quantitative data analysis: doing social research to test ideas*. United States of America: Jossey-Bass.

Tufte, Per Arne (2000): *En intuitiv innføring i logistisk regresjon*. Arbejdsnotat nr. 8-2000. Statens Institutt for Forbruksforskning, Lysaker.

Uchino, B. (2006): "Social support and health: a review of physiological processes potentially underlying links to disease outcomes". *The Journal of Behavioral Medicine*. Vol. 29, pp. 377–387.

Umberson, D. & J. Montez (2010): "Social relationships and health: a flashpoint for health policy". *Journal of Health and Social Behaviour*. Vol. 51, pp. 54–66.

Vassilev, I.; A. Rogers, C. Sanders et al. (2010): "Social networks, social capital and chronic illness self-management: a realist review". *Chronic Illness*. Vol. 7, pp. 60–86.

Verbiest, Nele (2019): "Foundations of Predictive Analytics in Python (Part 1)". Interactive course, DataCamp. Link: <https://www.datacamp.com/courses/foundations-of-predictive-analytics-in-python-part-1>. Tilgået 12. maj 2019.

Wad, Peter (2015): "Realitisk videnskabsfilosofi og kritisk realisme", i: M. H. Jacobsen, K. Lippert-Rasmussen & P. Nedergaard (red.): *Videnskabsteori - i statskundskab, sociologi og forvaltning*. Hans Reitzels Forlag.

Walia, Anish Singh (2017): "Activation functions and it's types – Which is better?". Link: <https://towardsdatascience.com/activation-functions-and-its-types-which-is-better-a9a5310cc8f>. Udgivet 29. maj 2017. Tilgået 3. maj 2019.

WHO (2019): "About social determinants of health". Link: <https://www.who.int/social-determinants/sdh-definition/en/>. Tilgået 1. Maj 2019

Wilkinson, Richard & Kate Pickett (2010): *The Spirit Level – Why equality is better for everyone*. London: Penguin Books.

Williams, D. R. & S. A. Mohammed (2009): "Discrimination and racial disparities in health: evidence and needed research. *The Journal of Behavioral Medicine*. Vol. 32, pp. 20–47.

Williams, Simon J. (1995): "Theorising class, health and lifestyles: can Bourdieu help us?". *Sociology of Health & Illness*. Vol 17(5), pp. 577-604.

Wolpert D. & W. Macready (1997): "No free lunch theorems for optimization". *IEEE Transactions on Evolutionary Computation*. Vol. 1(1), pp. 67–82.

Wright, Erik Olin (2005): *Approaches to Class Analysis*. Cambridge University Press.

Yurkovich, M.; J. A. Avina-Zubieta, J. Thomas, M. Gorenchtein & D. Lacaille (2015): "A systematic review identifies valid comorbidity indices derived from administrative health data". *Journal of Clinical Epidemiology*. vol. 68, pp. 3-14.

13. Oversigt over bilag

Følgende er en oversigt over de bilag, der er vedlagt separat.

Bilag 1: Sundhedsatlas Region Nordjylland

Bilag 2: Beskrivelse af forklarende variable i surveyundersøgelsen fra 2010

Bilag 3: Beskrivelse af forklarende variable i surveyundersøgelsen fra 2013

Bilag 4: Beskrivelse af forklarende variable i surveyundersøgelsen fra 2017

Bilag 5: Sammenligning af den konstruerede afhængige variabel for forekomsten af somatiske kroniske sygdomme i surveydata og registerdata for årene 2010 og 2013

Bilag 6: Figurer anvendt til vurdering af antal epoker i den endelige model

Bilag 7-11: Stata DO-filer

- Bilag 7: Sammenkobling af survey- og registerdata
- Bilag 8: Sammenligning af survey- og registerdata ift. den afhængige variabel
- Bilag 9: Klargøring af datasæt for de enkelte år
- Bilag 10: Klargøring af data fra 2013 og 2017 til kortlægning
- Bilag 11: Analyse med logistisk regression

Bilag 12-16: Python Notebook-filer

- Bilag 12: Logistisk regressionsanalyse: Træning og validering med data fra 2010
- Bilag 13: Logistisk regressionsanalyse: Træning og test med data fra 2010 og 2013
- Bilag 14: Neural netværksanalyse: Træning og validering med data fra 2010
- Bilag 15: Neural netværksanalyse: Test med data fra 2013
- Bilag 16: Neural netværksanalyse: Prædikeret sandsynlighed for data fra 2013 og 2017