Predicting mortality by comorbidity for patients with hip arthroplasty

Prospective observational register studies of a nationwide Swedish cohort

Erik Bülow

Department of Orthopaedics
Institute of clinical sciences
Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2020

Cover illustration: *Dancing in the dark* by Simon Bülow.
Collage including the *Perpetual* (Ghetu Daniel; CC BY 2.0) and
Hip replacement (National Institutes of Health, USA; public domain).

Predicting mortality by comorbidity for patients with hip arthroplasty
Prospective observational register studies of a nationwide Swedish cohort
©Erik Bülow 2020
erik.bulow@gu.se

Study II is reproduced under the CC BY-NC 3.0 license.
Study III is reproduced as an unedited, pre-publication version with permission and copyright of the British Editorial Society of Bone and Joint Surgery.

ISBN 978-91-7833-950-1 (PRINT)
ISBN 978-91-7833-951-8 (PDF)
http://hdl.handle.net/2077/64518

Typeset with LaTeX
Main font: The Scientific and Technical Information Exchange (STIX) Two

Printed in Borås, Sweden 2020
Printed by Stema Specialtryck AB

To Mom, Who took me to the library.

Predicting mortality by comorbidity for patients with hip arthroplasty

Prospective observational register studies of a nationwide Swedish cohort

Erik Bülow

Department of Orthopaedics, Institute of clinical sciences
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ABSTRACT

Introduction: Patients with total hip arthroplasty (THA) due to osteoarthritis (OA) are usually healthy, some with a remaining lifetime of several decades after surgery. Patients with hip arthroplasty due to a femoral neck fracture (FNF) are often old and frail with 13 % mortality within 90 days of surgery. To predict all-cause mortality for those groups has been considered but no prediction model has so far been widely accepted.

Patients and methods: We developed an R package to estimate comorbidity from large data sets. We used data from the Swedish Hip Arthroplasty Register (SHAR), the National patient register (NPR), the national prescription register, the Longitudinal integrated database for health insurance and labour market studies (LISA), the Swedish population register and the National Joint Registry for England, Wales, Northern Ireland, the Isle of Man and the States of Guernsey (NJR). We evaluated the discriminatory abilities of the Charlson and Elixhauser comorbidity indices to predict mortality for patients with hip arthroplasty due to OA and FNF. We also developed a new statistical prediction model for 90-day mortality after cemented THA due to OA using a bootstrap ranking procedure with logistic least absolute shrinkage and selection operator (LASSO) regression. The model was validated internally, as well as externally with patients from England and Wales. We built a web calculator for clinical usage. Finally, association between the Elixhauser comorbidity index and the restricted mean survival time (RMST) after surgery was assessed for patients with THA due to OA.

Results: The *coder* R-package provides a dynamic solution for patient classification. Neither the Elixhauser, nor the Charlson comorbidity indices accurately predicted mortality after hip arthroplasty due to OA or FNF (area under the curve (AUC) < 0.6 and AUC < 0.7; where 0.7 is a common lower threshold for an acceptable model). The new model, based on age, sex, the American Society of Anesthesiologists (ASA) physical status class, and the presence of cancer, disease of the central nervous system (CNS), kidney disease and obesity, did predict 90-day mortality with good discriminatory ability (AUC > 0.7) and was well calibrated for predicted probabilities up to 5 %. Shortening of the RMST for 10 years after surgery ranged from 315 days for patients with no comorbidity, to 1,193 days for patients with at least 3 comorbidities.

Conclusion: We found that the Charlson and Elixhauser comorbidity indices, although associated with RMST, did not predict mortality after hip arthroplasty. Our parsimonious model did predict 90-day mortality after THA due to OA.

Keywords: *Hip arthroplasty, mortality, comorbidity, osteoarthritis, femoral neck fracture, prediction, validation, web calculator, shared decision making, restricted mean survival time*

ISBN 978-91-7833-950-1 (PRINT)
ISBN 978-91-7833-951-8 (PDF)
http://hdl.handle.net/2077/64518

Sammanfattning på svenska

Introduktion: Majoriteten höftprotesoperationer föregås av antingen höftledsartros eller en fraktur på lårbenshalsen. Artrospatienter är i regel friska individer, en del med en återstående livslängd på flera decennier. Frakturpatienter å andra sidan är ofta gamla och sköra. 13 % av dem avlider inom 90 dagar efter operation. Det har länge varit önskvärt att predicera överlevnad för respektive patientgrupp. Av befintliga modeller har dock ännu ingen fått något bredare genomslag. Utvecklingen av tidigare modeller har ofta lidit av för små patientunderlag eller användning av suboptimala statistiska metoder.

Patienter och metod: Programvara med öppen källkod (ett R-paket) utvecklades för beräkning av samsjuklighet utifrån registrerade diagnoskoder i NPR. För de empiriska studierna inkluderade vi sedan patienter från det Svenska Höftprotesregistret (SHPR). Vi länkade patienternas data med hjälp av personnummer till det Nationella patientregistret (NPR), Läkemedelsregistret och den Longitudinella integrationsdatabasen för sjukförsäkrings- och arbetsmarknadsstudier (LISA). Vi validerade sedan den prediktiva styrkan av två samsjuklighetsindex, Charlson och Elixhauser, för prediktion av död avseende patienter med dels artros, dels höftledsfraktur. Vi utvecklade därefter en egen prediktionsmodell för 90-dagarsmortalitet efter höftprotesoperation till följd av artros. Vi nyttjade bootstrapping kombinerat med logistisk LASSO-regression. Modellen validerades internt och externt för patienter från England och Wales i samarbete med det brittiska nationella ledproesregistret (NJR). Modellen kompletterades med en webbkalkylator för kliniskt bruk. Slutligen undersökte vi association på gruppnivå mellan Elixhausers samsjuklighetsindex och medelvärdet för överlevnadstiden begränsad till tio år för protesopererade patienter med höftledsartros.

Resultat: R-paketet *coder* (`https://eribul.github.io/coder/`) bidrog till effektivare datorberäkningar och erbjuder ett flexibelt ramverk för patientklassifikation. Varken Elixhausers eller Charlsons samsjuklighetsindex möjliggjorde någon noggrannare prediktion av död efter vare sig elektiv operation med totalprotes till följd av artros, eller akut operation med halv- eller totalprotes efter fraktur på lårbenshalsen (AUC $< 0{,}6$ respektive AUC $< 0{,}7$; där 0,7 är ett vanligt lägre gränsvärde för en acceptabel modell). Vår föreslagna modell baserades på ålder, kön, hälsoklass enligt det Amerikanska Sällskapet för Anestesiologi (ASA) samt förekomst av cancer, neurologisk sjukdom, njursjukdom och fetma. Modellen predicerade död inom 90 dagar med ett AUC-värde på över 0,7. Modellen var också väl kalibrerad för skattade sannolikheter upp till 5 %. Medelvärdet av den begränsade överlevnadstiden under tio år efter operation var 315 dagar kortare än så för patienter utan samsjuklighet, jämfört med 1 193 dagar för patienter med tre eller fler samtida diagnoser enligt Elixhauser.

Slutsats: Användning av samsjuklighetsindex som prediktor av död efter höftprotesoperation har tidigare rekommenderats. Vi fann denna rekommendation tveksam även om vi också påvisade association mellan samsjuklighet och medelvärdet av den begränsade överlevnadstiden under tio år efter operation. Istället föreslår vi en relativt enkel modell för prediktion av död inom 90 dagar efter en höftprotesoperation till följd av artros. Denna modell visade sig ha god prediktiv styrka. Vi erbjuder också en webbkalkylator för användning i klinisk verksamhet (`https://erikbulow.shinyapps.io/thamortpred`).

Contents

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. E Bülow
coder: An R package for code-based item classification.
In manuscript

II. E Bülow, O Rolfson, P Cnudde, C Rogmark, G Garellick, S Nemes.
Comorbidity does not predict long-term mortality after total hip arthroplasty.
Acta Orthopaedica, **88** (July) 2017.

III. E Bülow, P Cnudde, C Rogmark, O Rolfson, S Nemes.
Low predictive power of comorbidity indices identified for mortality after acute arthroplasty surgery undertaken for femoral neck fracture.
The Bone & Joint Journal. 2019;**101**-B(1):104-112.
doi:10.1302/0301-620X.101B1.BJJ-2018-0894.R1

IV. A Garland, E Bülow,* E Lenguerrand, A Blom, JM Wilkinson, A Sayers, O Rolfson, NP Hailer.
Prediction of 90-day mortality after total hip arthroplasty: A simplified and externally validated model based on observational registry data from Sweden, England and Wales
In manuscript

V. E Bülow, O Rolfson, S Nemes.
Comorbidity decreased the restricted mean survival time for patients with total hip arthroplasty: An observational register study of 150,367 patients from the Swedish Hip Arthroplasty Register 1999–2015
In manuscript

*EB and AG contributed equally.

Additional papers related to the field but not part of the thesis

Nemes, Szilard; Greene, Meridith E; **Bülow, Erik**; Rolfson, Ola. Summary statistics for Patient-reported Outcome Measures: the improvement ratio. *European journal for Person Centered Healthcare*, **3**, 3, 334-342, 2015.

Cnudde, Peter; Nemes, Szilard; **Bülow, Erik**; Timperley, John; Malchau, Henrik; Kärrholm, Johan; Garellick, Göran; Rolfson, Ola. Trends in hip replacements between 1999 and 2012 in Sweden. *Journal of Orthopaedic Research*, **36**, 1, 432-442, 2018

Eneqvist, Ted; Nemes, Szilárd; **Bülow, Erik**; Mohaddes, Maziar; Rolfson, Ola. Can patient-reported outcomes predict re-operations after total hip replacement? *International orthopaedics*, **42**, 2, 273-279, 2018, Springer Berlin Heidelberg.

Eneqvist, Ted; **Bülow, Erik**; Nemes, Szilárd; Brisby, Helena; Garellick, Göran; Fritzell, Peter; Rolfson, Ola. Patients with a previous total hip replacement experience less reduction of back pain following lumbar back surgery. *Journal of Orthopaedic Research*, **36**, 9, 2484-2490, 2018.

Cnudde, Peter HJ; Nemes, Szilard; **Bülow, Erik**; Timperley, A John; Whitehouse, Sarah L; Kärrholm, Johan; Rolfson, Ola. Risk of further surgery on the same or opposite side and mortality after primary total hip arthroplasty: A multi-state analysis of 133,654 patients from the Swedish Hip Arthroplasty Register. *Acta orthopaedica*, **89**, 4, 386-393, 2018, Taylor & Francis.

Berg, Urban; **Bülow, Erik**; Sundberg, Martin; Rolfson, Ola. No increase in readmissions or adverse events after implementation of fast-track program in total hip and knee replacement at 8 Swedish hospitals: An observational before-and-after study of 14,148 total joint replacements 20112015. *Acta orthopaedica*, **89**, 5, 522-527, 2018, Taylor & Francis.

Nemes, Szilard; Lind, Dennis; Cnudde, Peter; **Bülow, Erik**; Rolfson, Ola; Rogmark, Cecilia. Relative survival following hemi-and total hip arthroplasty for hip fractures in Sweden. *BMC musculoskeletal disorders*, **19**, 1, 407, 2018, BioMed Central.

Jawad, Z; Nemes, S; **Bülow, E**; Rogmark, C; Cnudde, P. Multi-state analysis of hemi-and total hip arthroplasty for hip fractures in the Swedish populationResults from a Swedish national database study of 38,912 patients. *Injury*, **50**, 2, 272-277, 2019, Elsevier.

Cnudde, Peter; **Bülow, Erik**; Nemes, Szilard; Tyson, Yosef; Mohaddes, Maziar; Rolfson, Ola. Association between patient survival following reoperation after total hip replacement and the reason for reoperation: an analysis of 9,926 patients in the Swedish Hip Arthroplasty Register. *Acta orthopaedica*, **90**, 3, 226-230, 2019, Taylor & Francis.

Ferguson, Rory J; Silman, Alan J; Combescure, Christophe; **Bülow, Erik**; Odin, Daniel; Hannouche, Didier; Glyn-Jones, Siôn; Rolfson, Ola; Lübbeke, Anne. ASA class is associated with early revision and reoperation after total hip arthroplasty: an analysis of the Geneva and Swedish Hip Arthroplasty Registries. *Acta orthopaedica*,**90**, 4, 324-330, 2019, Taylor & Francis.

Wojtowicz, Alex Leigh; Mohaddes, Maziar; Odin, Daniel; **Bülow, Erik**; Nemes, Szilard; Cnudde, Peter. Is Parkinsons disease associated with increased mortality, poorer outcomes scores, and revision risk after THA? Findings from the Swedish Hip Arthroplasty Register. *Clinical Orthopaedics and Related Research*, **477**, 6, 1347-1355, 2019, LWW.

Hansson, Susanne; **Bülow, Erik**; Garland, Anne; Kärrholm, Johan; Rogmark, Cecilia. More hip complications after total hip arthroplasty than after hemiarthroplasty as hip fracture treatment: analysis of 5,815 matched pairs in the Swedish Hip Arthroplasty Register. *Acta orthopaedica*, **91**, 2, 133-138, 2020, Taylor & Francis.

Bülow, Erik; Nemes, Szilard; Rolfson, Ola. Are the First or the Second Hips of Staged Bilateral THAs More Similar to Unilateral Procedures? A Study from the Swedish Hip Arthroplasty Register. *Clinical Orthopaedics and Related Research*, **478**, 6, 1262-1271, 2020, LWW.

Nemes, Szilard; **Bülow, Erik**; Gustavsson, Andreas. A Brief Overview of Restricted Mean Survival Time Estimators and Associated Variances. *stats*, **3**, 2, 2020, 107-119, MDPI.

Eneqvist, Ted; **Bülow, Erik**; Nemes, Szilard; Brisby, Helena; Fritzell, Peter; Rolfson, Ola. Does the order of total hip replacement and lumbar spinal stenosis surgery influence patientreported outcomes: an observational register study. *J Orthop Res*. Published online July 25, 2020:jor.24813. doi:10.1002/jor.24813

Glossary

Acetabulum Concave surface that makes up the pelvic part of the hip joint 1

Adverse event Complication after surgery, often within 30 or 90 days 3

Anatomy Science of form and structure of the body 4

Big data No clear definition but often described in terms of high volume, variety, velocity and veracity 19, 21

Bootstrapping To resample with replacement and to repeat relevant procedures for each sample 16

Calibration Process to measure similarity or dissimilarity between observed and predicted outcomes 16

Causality One event or condition leading to another, often hard (impossible) to establish in observational studies 3

Censoring Unknown (death) status of patients lost to follow-up 11, 18

Charlson Classification of comorbidity with 17 distinct conditions and a weighted index sum 5

Charnley Patient classification for outcome assessment of low-friction hip arthroplasties 5

Classification Grouping of items according to common characteristics 3

Coefficient Multiplicative factor of independent variable(s) in regression analysis 9

Comorbidity Morbidity co-existing with main diagnose 2

Completeness Proportion of relevant patients/procedures reported to the register 8

Concordance index Measure of rank correlation, the ability to assign higher probabilities to true events 17

Confusion matrix Error matrix with combinations of observed and estimated/predicted values 17

Covariate A variable that might be predictive of the outcome 12

Coverage Proportion of health care units (hospitals) affiliated with the register 8

Cox regression Semiparametric survival model assuming proportional hazards 5, 12, 13, 27, 39

Cross-validation To train a model on one partition, to evaluate it on another, and then to repeat 16

Cumulative hazard Accumulated hazard up to a certain point in time 11

Discrimination Ability to distinguish between patients who do, or do not, experience the event of interest (death at a certain time) 16

Double Legacy term for binary64, a float number with double precision used by computers 19

Effective sample size The number of cases with the less probable outcome 15

Elixhauser Classification of comorbidity with 31 distinct conditions 6

Epidemiology Science of spread and control of medical conditions within populations 2

Etiology Underlying cause/origin of disease 3, 4

External validation To assert transportability of a model, by application to a different, yet comparable, population 16, 17

Femur The large bone connecting the pelvis to the knee 1

Float Computer approximation of real numbers 19

Floppy disk Arcane magnetic storage medium made of squared plastic 20

Hazard Instant probability of death or the force of mortality 11, 33

Hemiarthroplasty Prosthesis without acetabulum component 2, 8, 26, 27

Hip joint The joint connecting the pelvic acetabulum to the femur 1

Hip arthroplasty Hip prosthesis v, 1, 8, 9, 11, 13, 15, 23, 25, 35, 38, 39, 42

Infix operator Programming operation similar to a function but with differnt syntyx, for example the arithmetic operators (+, -, / and *) 19

In-hospital Something occurring in a hospital (i.e. e deaths among patients at the hospital) 5

Index disease Disease of main interest 3

Integer Whole number used by computers where the range of available numbers depends on the operating system 19

Internal validation To assert reproducibility of a model, usually with a split-sample, cross-validation or bootstrapping 16

Jackknife Cross validation with only one sample used for validation 16, 30

Linear regression Regression analysis where a weighted sum of independent

covariates are lineary related to the dependent outcome 9

Logistic regression Model relating the probability of a binary outcome to a linear combination of covariates 6, 9, 10, 16, 18, 19, 28, 29, 36, 37

Millenium bug Also known as the year 2000/Y2K problem/bug/glitch; a problem caused by the two digit abbrivation of years such that year 2000 was not distinguised from 1900 20

Morbidity Disease or medical condition 4

Mortality Proportion of deaths within a cohort during a certain period 2

Nomenclature A defined set of names and terms 3

Null hypothesis The (often unrealistic) assumption of no relation/association/effect to be tested against a (more interesting) alternative hypothesis 12

Ockham's razor A philosophy where simplicity/parsimounious is preferred if possible 15

Osteoporosis A bone metabolic disease leading to reduced bone mineral density 1

Out-of-bag Non-sampled data used for internal validation 17

Post-operatively Event happening after surgery 9

Predict To forsee a future event based on baseline variables using a statistical model 14

Pre-operatively Event happening before surgery 9

Primary surgery The first insertion (not a re-operation) of a prosthesis 8, 11

Prosthesis Artificial hip joint 1

R Statistical open source software v, 16, 19, 20, 21, 22, 23, 25, 29, 30, 31, 33, 40

Regression analysis Statistical procedure to estimate a relation between independent and dependent variables 41

Relative risk Ratio of probabilities of an outcome in an exposed versus an unexposed group 12

Re-operation Any additional surgery performed on a hip with privious hip arthroplasty 8

Residual Difference between observed and estimated/predicted outcome 12

Revision Re-operation including replacement or extraction of any part of the prosthesis 8

RxRisk V Classification of comorbidity based on medical codes 7

S Statistical open source software (predecessor of R) 19, 20

Sensitivity Proportion of true positives, observed events predicted as such 17

Sigmoid Mathematical function depicted with s-shaped curve 9

Specificity Proportion of true negatives, observed non-events predicted as such 17

Stratum Online IT-platform for collection, storage and presentation of quality register data 8, 26

Syntactic suger Design elements of a programming language not introducing any new functionality but which improves clearity, concistency or which introduce an alternative programming style 19

Transportability If a model is generalizable to another population 16

Trapezoid rule Numerical technique used to approximate a definite integral 17

Acronyms

AE adverse event 6, 22, 23, 24, 26
AI artificial intelligence 15
AIC Akaike information criteria 15
AIDS/HIV acquired immunodeficiency syndrome/human immunodeficiency virus 5
ANCOVA analysis of covariance 30
API aplication programming interface 24
ASA American Society of Anesthesiologists v, 5, 27, 32, 35, 36, 38, 40
ATC Anatomic Therapeutic Chemical classification system 4, 5, 7, 27
AUC area under the curve v, 17, 18, 27, 31, 32, 33, 34, 35, 36, 40
BIC Bayesian information criteria 15
BMI body mass index 5, 27, 38, 40
BOA Better management of patients with OsteoArthritis 39
CI confidence interval 14, 16, 27, 35, 36, 38, 39
CNS central nervous system v, 32, 40
CPS Comorbidity-poly Pharmacy Score 7, 24
CRAN Central R Archive Network 21, 22, 24
DRG diagnose related group 6
DSL domain specific language 21
EPV events per variable 37, 38
FNF femoral neck fracture v, 1, 2, 9, 23, 27, 31, 33, 40
GDPR European General Data Protection Regulation 26
GEE generalized estimating equation 10, 14, 30, 39
GNU GNU's Not UNIX 20
GPL General Public License 20
HES Hospital Episodes Statistics 9
HGLM hierarchical generalized linear models 10, 38
HR hazard ratio 5, 12, 14, 34, 39
ICD International Classification of Diseases 3, 4, 5, 6, 7, 22, 24, 25, 27, 33, 34, 35, 41
IPW inverse probability weighting 39
LASSO least absolute shrinkage and selection operator v, 16, 29, 36, 37
LISA Longitudinal integrated database for health insurance and labour market studies v, 25, 26
LON League of Nations 4
MAR missing at random 38
MCAR missing completely at random 38
MICE multiple imputation using chained equations 38
ML machine learning 15, 36, 41
MNAR missing not at random 38
NBHW National Board of Health and Welfare 4, 5, 25, 26
NCSP NOMESCO Classification of Surgical Procedures 5
NHS National Health Service 9
NJR National Joint Registry for England, Wales, Northern Ireland, the Isle of Man and the States of Guernsey v, 9, 26, 32
NMR National Musculoskeletal Registry 9
NOMESCO Nordic Medico-Statistical Committee 5
NPR National patient register v, 5, 9, 25, 26, 34
NRI Net reclassification improvement 34
NSE non-standard evaluation 22
OA osteoarthritis v, 1, 2, 23, 27, 29, 30, 31, 32, 33, 36, 40, 41, 42
ODBC open database connectivity 21
OHDSI Observational Health Data Sciences and Informatics 41
OR odds ratio 11, 34, 35, 36, 38
OS operating system 21
PCA principal component analysis 15
PCRE Perl-compatible regular expressions 33
PDL Patient data act 8
PIN personal identity number 7, 8, 15, 24, 26, 35
PJI prosthesis joint infection 42
PROM patient reported outcome measure 9, 42
RAM random access memory 19, 21
RC Centre of registers 8, 26, 43, 44
RCC Regional cancer center 43, 44
RCT randomized clinical trial 35
RMSE root-mean-square error 15
RMST restricted mean survival time v, 13, 14, 23, 29, 30, 39, 40
RMTL restricted mean time lost 14, 32
ROC receiver operating characteristic 17, 18, 27
ROSE random over-sampling examples 36
RWE real world evidence 35
SCB Statistics Sweden 2, 25, 26
SDM shared decision making 16, 23, 39, 40
SFS Swedish code of statues 25
SHAR Swedish Hip Arthroplasty Register v, 5, 8, 9, 23, 25, 26, 27, 32, 43
SJAR Swedish Joint Arthroplasty Register 9
SKAR Swedish Knee Arthroplasty Register 8, 9
SOU official report of the Swedish government 25
SQL Structured Query Language 26
THA total hip arthroplasty v, 1, 2, 6, 8, 23, 26, 27, 29, 30, 31, 32, 36, 40, 41
VGR Region Vastra Gotaland 8, 25
WHO World Health Organisation 3, 4, 5

1 INTRODUCTION

This thesis concerns statistical association and prediction modeling of mortality[1] and comorbidity for patients with hip arthroplasty.

1.1 THE HIP JOINT

The hip joint is the biggest joint in the human body, next to the knee. It is the biggest ball-and socket (spheroid) joint with six degrees of freedom (flexion/extension, internal/external rotation and adduction/abduction), thus with the possibility to move in all directions. It makes us mobile and it provides us with the possibility to escape danger and to hunt for food. The large femur ends with a spherical slippery ball. It moves almost without friction and fits into a hemispherical socket, the acetabulum as part of the pelvis. Mobility is a very central part of human freedom, although we might not think about it if everything works as expected. Most of the time it does, but not for everyone, and not forever.

Osteoarthritis (OA) is a degenerative disease, affecting the elastic hyaline cartilage, which has an extremely low coefficient of friction and which lubricates the joint between the convex femoral head and the concave acetabulum. In 2012, 27 % of all Swedish inhabitants, 45 years and older, were estimated to have OA, with 5.8 % affecting the hip.[2] Lifestyle factors as well as an aging population leads to an increased disease burden.[3] The mean ages at surgery are 67 and 69 years for Swedish males and females and close to 60 % of the patients are female.[4]

The occurrence of a femoral neck fracture (FNF) is a traumatic event, although approximately one third of the cases are predeceased by confirmed osteoporosis weakening the bone by reducing the bone mass and thereby the density. Young individuals might break their bones due to high energy trauma, but the old and frail dominates the cohort.

A broken bone of a young person might heal easily due to a large proportion of elastic collagen. An older bone is more fragile and brittle. Impaired fracture healing is, however, not the main problem; immobilization and comorbidity are. A non-displaced FNF might be treated with internal fixation or by hip arthroplasty, a treatment otherwise most commonly applied to displaced fractures.[5] The mean ages at surgery are 81 and 83 years for Swedish males and females.[4] Approximately three out of four patients are female,[6] but the proportion of males is increasing over time (20 % in year 2000 and 35 % in 2018).[4]

Figure 1.1: Hip prosthesis exposed in the Center for hip surgery at the Wrightington Hospital outside Manchester in the UK. This is where Sir John Charnley developed the low friction hip replacement, a fascinating story described in "The man and the hip" by William Waugh.[7]

1.2 HIP ARTHROPLASTY

Hip arthroplasty (hip replacement/hip prosthesis; figure 1.1) is used as treatment for several diagnoses including tumors, childhood diseases and inflammatory hip diseases. The two most common causes, however, are (primary) OA and FNFs.

Patients with OA are, if operated, treated with total hip arthroplasty (THA), a prosthesis with two main parts, a femoral stem with a caput (head), and an acetabular cup. There were 7,839 patients with primary hip OA, constituting 77 % of all primary hip arthroplasty inserted in Sweden 1999. 13,006 surgeries were performed in 2012 and 14,773 in 2017 (Figure 1.2). There were 4,802 patients with FNF treated with hip arthroplasty in

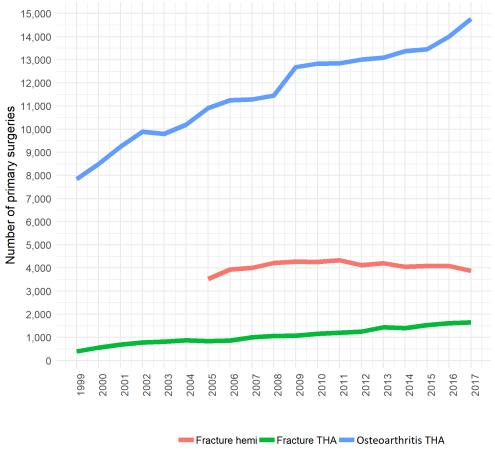

Figure 1.2: Number of primary surgeries performed 1999–2017. Hemiarthroplasties were not recorded in the Swedish Hip Arthroplasty Register (SHAR) before 2005.

2006 and 5,523 in 2017. Those patients were treated with either THA or hemiarthroplasty, essentially a femoral stem with a large caput, but without the acetabular cup.

1.3 MORTALITY

The Cambridge dictionary defines mortality as both the condition of being mortal, as well as the number of deaths within a society during a specified period. In epidemiology, mortality refers to the number of deaths caused by (or at least associated with) a specific condition (disease). This is related, although different, from the proportion of deaths (regardless of cause) among individuals with the condition, who dies during a specified period after an individual index date (such as the onset of a disease). This is termed "survival" (Section 1.12). Both measures (along with incidence and prevalence) are important for example in cancer epidemiology where the condition of interest is itself lethal. In other fields, such as orthopedics, mortality and (the opposite/inverse of) survival are often used interchangeable. We choose to follow this tradition throughout the thesis, although the use of "mortality" could sometimes be re-expressed in terms of survival.[8]

The first organized registration of death begun in northern Italy during a plague pandemic in the 15th century.[9] A death certificate, issued by a physician, or a certified surgeon, was then necessary to regulate the movement of corps and to secure sanitary conditions before burial. The practice spread temporary to France, Switzerland and the Netherlands. A similar practice begun in England; the Bills of mortality, a weekly list of deaths and funeral dates established in 1603 and sold for 4 shillings annually.[10,11]

The birth of epidemiology, however, is attributed to a book by Graunt in 1661.[12,13] In his foreword, he briefly mentions that he oppose polygamy to increase population size. The book is otherwise known for its aggregated death statistics based on the bills of mortality.

It took another 44 years until death registration was systematically introduced in Sweden, handled by the church from 1686. Priests recorded dates of births, baptisms, confirmations, deaths, immigrations, emigrations and disappearances.[14] Those records were decentralized at each parish but a national data aggregation was introduced in 1749.[15] The data collecting process was regionalized at county-level from 1947 and a computerized national population register was introduced in 1967 by the Swedish tax agency. It is available for research through Statistics Sweden (SCB).[16]

Death dates are recorded exactly if known (as for most cases). Approximate dates might be used for individuals who disappear or die unnoticed. People who emigrate and die abroad might be censored (lost to follow-up). Emigration is probably more common among patients with OA compared to patients with FNF, since those are generally younger, healthier and more mobile. Cause of death has been recorded in a separate register since 1952 but were not used in the thesis.[17]

1.4 COMORBIDITY

There is no strict and commonly agreed definition of comorbidity. The term was coined in 1970 by the clinician and epidemiologist

Alvan Feinstein,[18] one of the father figures of clinical epidemiology: "The term comorbidity will refer to any distinct additional clinical entity that has existed or that may occur during the clinical course of a patient who has the index disease under study". He argued that the concept was relevant to distinguish patients with different needs and prognosis, in addition to age, sex and race. He also argued that comorbidity might influence the risk of death more than the index disease itself. He was careful to distinguish between comorbidity and adverse events (complications), where the first is pre-existing and the other occurs after the index disease. Other definitions of comorbidity highlights that it must be independent of the index disease regarding etiology or causality, or less commonly, that it should be a significant factor influencing mortality and resource use in hospitals.[19]

1.5 CODES AND CLASSIFICATIONS

The history of medical coding is a history of international collaboration.[9] It all started by the bills of mortality. Those records were, however, made without a standardized nomenclature of diseases. It was noted in 1839 by William Farr, director of the Registrar-General in England and Wales that: "The advantages of a uniform nomenclature, however imperfect, are so obvious [...]. Each disease has, in many instances, been denoted by three or four terms, and each term has been applied to as many different diseases [... This] should be settled without delay."

A delay of 30 years nevertheless occurred. But then, the Nomenclature of Diseases, presented by the Royal College of Physicians of London, was finally published. Their list was updated and maintained during exactly one hundred years. A similar initiative was taken by the Surgeon General in the USA in the late 19th century, but this activity was soon discontinued. Multiple non-standardized nomenclatures were then used in the USA until 1919, when the Standard Nomenclature of Diseases and Pathological Conditions, Injuries and Poisonings for the United States, was published. This initiative did not last long, however. The Standard Nomenclature of Diseases and Operations was more successful, published 1930–1961. It was succeeded by the Current Medical Terminology used from 1963, and the International Nomenclature of Diseases, a collaborated effort by the International organizations of medical statistics and the World Health Organisation (WHO).

Parallel to the development of detailed nomenclatures, an additional approach was made for statistical classification. For this purpose, groups of conditions were not ordered alphabetically but hierarchically, which made it possible to aggregate data for summary statistics on different levels. Discussions started at the first statistical congress in Brussels 1853. It was the perception by the time that: "a uniform list was impossible because of the different training of doctors and their tendency to call diseases by whatever name they chose".[9]

Florence Nightingale also took part in later discussions and then, in 1893, Jacques Bertillion, chief of statistics in Paris, presented the International List of Causes of Death.

1.5.1 ICD

Bertillon's classification was later approved by the American Public Health Association in 1899 and revised to become the first version of the *International Classification of Diseases (ICD)-1*. It was used internationally 1900–1909 although Bertillion noted that "[European] countries want to be comparable with each other but above all comparable with themselves."[9] A second version, *ICD-2*, was used 1910–1920. It included new medical conditions, as well as a new section concerning stillbirths. The classification was accompanied by an index section; a document of 1,044 typewritten pages. Preparation of *ICD-3* (used 1921–1929) was delayed due to world war I and because Bertillion got seriously ill. *ICD-4* (used 1930–1938) was prepared without him, by a commission including representatives from a newly formed sta-

tistical experts committee within the health section of the League of Nations (LON). Some attempts were made to put more focus on etiology, rather than anatomy. *ICD-5* (used 1939–1948) aimed to be more clinically relevant than its predecessors, although scientific issues were also considered. *ICD-6* (used 1949–1957) was the first version to include morbidity, not only mortality. It was a major revision undertaken by the WHO (as part of the United Nations, the successor of the LON). It was the first ICD version to be adopted by the Swedish National Board of Health and Welfare (NBHW), or more formally its predecessor, the Royal Swedish Medicines Agency, in 1951. Most of the classification was adapted as suggested, but the sections on violence and poisoning, as well as mental disorders, were modified.[20] *ICD-7* (used 1958–1967) was a minor revision compared to ICD-6 but the support organization increased and the first WHO center for Classification of Diseases was established as part of the General register office in England. ICD-7 was the first revision to be used by the cancer register in Sweden, established in 1958. A modified version was published in 1965 including additional subclassification compared to the international standard. One reason for revision was to facilitate automatic computer processing.[21] Sweden was not the only, although one of the more prominent, countries making semi-official modifications to the classification. This practice was acknowledged in *ICD-8* (used 1968–1978), where additional codes were included for diagnostic indexing of clinical records. To reach total consensus was, however, not possible. ICD-8 was therefore also modified before adaptation in Sweden 1969. The medical profession called for an even more fine grained classification, but this had to be compromised to maintain the original purpose of a classification used for data aggregation.[22] *ICD-9* (used 1979–1994) was planned as a minor revision, which became substantial. It was decided to put more focus on medical manifestations rather than on etiology, and to record some conditions twice, once for etiology and once for manifestation. The clinical modification (*ICD-9-CM*) is still used for morbidity in the USA (although no longer for mortality). A more detailed oncological adaptation (*ICD-O*) was also released for use by cancer centers, with additional topographical and morphological coding. ICD-9 was introduced in Sweden 1987.[23] It was decided to make a throughout translation into Swedish, decreasing the use of Latin, which was more prominent in previous versions.

ICD-10 has been used since 1995 (1997 in Sweden). It was once again a major revision due to non-statistical needs. A new alphanumeric code structure was adopted. Codes start with a letter followed by three digits (possibly with a dot between the second and third). Some codes have an additional fifth letter for further sub-classification introduced in different countries. ICD-10 has undergone annual revision since 1997. A modified version (*ICD-10-SE*) is used in Swedish clinical settings since 2011. It contains 33,547 codes whereof 2,800 concern national sub-classification.[24] The clinical adaptation used in the USA, *ICD-10-CM*, contained 72,184 codes in 2020.[25] *ICD-11* is not yet implemented but was released as an online classification tool in June 2018. It will be used from January 2022.[26]

It should be noted that a one-to-one code match is not guaranteed between different versions of ICD, although some cross-walk algorithms exist.[27] It is usually possible to back-translate a newer code to an older version, although some granularity might get lost in the process. To translate an old code to a new version might be more problematic.

The ICD-10 does not contain laterality as part of the individual codes. This might be distinguished by an additional specification of `ZXA00` for right, `ZXA05` for left and `ZXA10` for bilateral conditions. This is less commonly used in practice, however.

1.5.2 ATC

The Anatomic Therapeutic Chemical classification system (ATC) was developed by the WHO Collaborating Centre for Drug Statistics Methodology in 1976. Although an in-

ternational standard, implementations differ between countries, partially due to different vetting processes before national/regional introduction of new medications, as controlled by the Dental and Pharmaceutical Benefits Agency in Sweden. The classification is constantly updated as new compounds are discovered and new drugs are introduced. A Swedish version is updated nightly and provided by the Swedish medical products agency.[28]

1.5.3 NOMESCO

The Nordic Medico-Statistical Committee (NOMESCO) is a delegation with annual meetings and an office in Copenhagen.[29] The NOMESCO Classification of Surgical Procedures (NCSP) was first published in 1996. It was implemented as NCSP-S in Sweden 1997.

1.6 COMORBIDITY DATA

All Swedish hospitals, private and public, are obliged to report patient visits and hospital admissions, to the National patient register (NPR). This register consists of two parts: the inpatient- and the outpatient registries. Somatic diagnoses have been recorded in the inpatient register (the Hospital Discharge Register), since 1964. Psychiatric care was added in 1973 and outpatient visits can be found in the outpatient register since 2001. The diagnose coverage is up to 99 % but varies between different diagnoses.[30] Diagnoses are coded by ICD-10-SE whereas performed medical and surgical procedures are coded by NCSP-S, both since 1997.

ATC codes are recorded in the medical prescription register maintained by the NBHW since 2005.

Some comorbidity data are also captured explicitly by the Swedish Hip Arthroplasty Register (SHAR) (Section 1.9): The American Society of Anesthesiologists (ASA) Physical Status classification is evaluated by an anesthesiologist on a scale of I–VI before surgery: (I) healthy patient, (II) mild systemic disease, (III) severe systemic disease, (IV) severe systemic disease that is a constant threat to life, (V) a moribund person who is not expected to survive without the operation, and (VI; not used by SHAR) a declared brain-dead person whose organs are being removed for donor purposes. Occurrence of dementia is recorded as none, probable or obvious. Obesity (body mass index (BMI) above 30 according to WHO) could be estimated from height and weight, as either supplied by the patient, or as measured at the time of the hospital visit. Occurrence of bilateral hip problems is captured by the Charnley class.

1.7 COMORBIDITY INDICES

There are too many medical codes to be studied individually. It is therefore common to categorize codes as meaningful conditions, such as diabetes, cancer or drug abuse.[31]

1.7.1 CHARLSON

Charlson et al.[32] developed a comorbidity index to predict in-hospital deaths and one-year mortality for 559 patients hospitalized in New York 1984. The cohort was screened for medical history by the time of hospital admission.

The classification entailed 19 categories, although leukemia and lymphoma are often grouped with malignancy, and acquired immunodeficiency syndrome/human immunodeficiency virus (AIDS/HIV) is often omitted since this condition is too rarely observed (it was more prevalent in the 1980:s). Diabetes, cancer and liver disease are included twice with sub-categories based on disease severity.[33]

The unweighted sum of all comorbidities was associated with mortality. To simply count the number of comorbidities would imply, however, that all conditions were considered to have an equal impact on mortality. This was considered unrealistic wherefore a weighted index was suggested. Cox regression (Section 1.12.3) was applied to estimate hazard ratios (HRs) (Section 1.12.1) for important comorbidities. Large enough values were rounded to integers and summed to an index. A modification of the index, a

Combined Age-Charlson comorbidity index (CA-CCI) was suggested for long-term mortality by adding one extra point for each additional ten years of age for patients 40 years and older. This modification has not been widely used, however. It is more common to include age as an additional covariate in multiple regression analysis.

The maximum Charlson score was 37, although scores above 8 have rarely been studied, since those are uncommon in most cohorts.[33] Thus, many studies truncate the Charlson comorbidity index at a lower point.

A self-administrated version of the index was later evaluated on 170 patients by comparing the recalled conditions to medical charts.[34] The Spearman correlation comparing the two measures was moderate (0.63), and lower for patients with less formal education. This either reflects that patients were unaware of their diagnoses, or that their medical records were inaccurate.

Multiple adaptations have been suggested to translate the originally heuristic descriptions for each disease, into formalized code extraction algorithms based on administrative data.[35]

Deyo et al.[36] were first to publish a coding algorithm using ICD-9-CM in 1992. Their coding algorithm was applied to a cohort of 27,111 patients with lumbar spine surgery. Association between the derived index and a number of outcomes, including mortality, were evaluated by logistic regression.

Even though Deyo et al. were the first to publish their adaption in 1992, Romano et al.[37] might have been the first to develop a similar method (published in 1993). They showed that: "the correspondence between the Charlson comorbidity index and ICD-9-CM is not intuitively obvious." They based their classification on the same list of comorbidities as Deyo et al. but they identified additional codes for each category. They advised to avoid the use of previously suggested index weights for surgically treated patients, since those were developed on a too small and too narrowly defined cohort.

D'Hoore et al. suggested a coding algorithm for ICD-9 (in addition to ICD-9-CM) in 1993.[38,39]

In 1996, Ghali et al. suggested new index weights to use with existing classifications.[40] They studied 13,117 patients with coronary artery bypass surgery and used logistic regression with in-hospital deaths as outcome. They found that only a subset of the originally proposed conditions was needed in their model: recent myocardial infection, cardiovascular disease, peripheral vascular disease and congestive heart failure.

Ghali acted as senior author for a series of papers developing an adaptation for ICD-10.[41–43] A new version based on ICD-9-CM was also suggested based on back-translation. The same group of researchers proposed their own index weights in 2011.[44] More than 25 years had passed since the original development of the Charlson index, and new treatments, altering the relation between comorbidity and mortality, had been introduced. Only 12 of the original comorbidities had stayed relevant.

In 2010, Armitage et al. suggested to consider 14 conditions, and to not use any index weights, but to simply count the number of comorbidities.[45] They applied their model on a cohort of 238,999 patients with elective THA.

Two Swedish researchers, Nele Brusselaers and Jesper Lagergren introduced a back-translated version to ICD-8 and ICD-9 in 2017.[46]

1.7.2 ELIXHAUSER

Elixhauser et al. proposed an alternative classification including 31 conditions, based on ICD-9-CM, in 1998.[19] They studied 1,779,167 patients, a considerably larger sample than the 559 patients studied by Charlson et al.[32] It was their explicit aim to include some comorbidities not used by Charlson, such as mental disorders, drug and alcohol abuse, obesity, coagulopathy, weight loss and fluid and electrolyte disorder. Potential comorbidities were distinguished from adverse events (AEs) by excluding conditions from the same diagnose related group (DRG) as the primary condition for each patient. Conditions known as common complications after treat-

ment were also excluded, such as pneumonia, pleural effusion, urinary tract infection, cardiac arrest, cardiogenic shock and respiratory failure. No weights were assigned to individual comorbidities and it was recommended not to use the classification with any standardized index, but to consider all conditions as separate variables in any regression model. This advice is reasonable for large cohorts, studying common events. To model 31 covariates in a small sample, or to estimate coefficients for rare events, is more difficult. It is therefore common to use an unweighted sum of the identified conditions as an aggregated score.

Quan et al. adapted the ICD-9-CM classification to ICD-10 in 2005.[43] Elixhauser et al. have also regularly revised their own algorithms. Later versions are based on ICD-10 but without cardiac arrhythmia.[47]

A set of index weights was suggested by van Walraven et al. in 2009.[48] Their version was developed to predict in-hospital deaths for 345,795 patients in a Canadian hospital. Some conditions were not associated with mortality and therefore excluded, some were positively associated with death, and some conditions had a protective effect on mortality (likely due to confounding with the reason for hospitalization).

Thompson et al. performed a similar study in 2015 with 228,365 patients in the USA.[49] They derived two new sets of index weights, one with and one without cardiac arrhythmia (complicated and uncomplicated hypertension combined).

1.7.3 OTHER COMORBIDITY SCORES

The RxRisk score is based on medical prescription data.[50] The original version included 39 medical conditions, but later version with 42,[51] 45,[52] and 50[53] conditions have been considered as well. RxRisk V was developed for 126,075 military veterans (dominantly men) in the USA. The index has been used for patients with hip arthroplasty in Australia.[54,55] A version based on 46 conditions coded by ATC was also developed for Australian veterans.[56]

In addition to ICD- or ATC-based scoring systems, there are alternatives including data from multiple sources. The Comorbidity-poly Pharmacy Score (CPS) is a relatively simple score suggested for trauma patients; a count of all pre-injury comorbid conditions and medications.[57] Comparisons have showed that using this index is comparable to the Charlson index, wherefore the need of additional data might be questioned. Another comprehensive score includes 34 variables measured by inpatient diagnoses (ICD-9-CM) and drug prescription (ATC).[58]

1.8 PERSONAL IDENTITY NUMBER

All Swedish inhabitants are assigned a personal identity number (PIN), either at birth or at immigration.[59] The system was introduced in 1947, the same year as the county-wise population registers (Section 1.3). The number had nine digits, although a tenth was added in 1967, the same year as the computerization of the census register, for both new and existing PINs. The system is governed by the Swedish tax agency and was the first of its kind in the world. The first six digits are the date of birth given by two digits for year, two for month and two for day of month. Unknown birth dates might be approximated. If too many people have the same (approximate) date of birth, another date might be chosen. This is more common for some dates than others, especially the first of January and the first of July, which are used as proxies for many immigrants with unknown birth dates.

The seventh and eighth digits indicates county of birth for inhabitants born 1947–1990 in Sweden, or the county of residence by the first of January 1947 for people born earlier (in Sweden or not). People born later (in Sweden or not) receives a random number. Immigrants born outside Sweden 1947–1989 had numbers between 93 and 99. Those numbers could also be used if too many births occurred on the same day in the same county. Digit number nine is odd for male and even for females. The last digit is a control number based on the Luhm Algorithm (US patent 2950048).

Some PINs might get re-used by immigrants with a birth date without an available PIN. This is done after an incubation period after the death of the previous PIN-owner.

1.9 SHAR

The Swedish Hip Arthroplasty Register (SHAR) is a national quality register. As such, it constitute an automated and structured collection of personal data that has been set up specifically for the purpose of systematically and continuously developing, and ensuring the quality of care (the Patient data act (PDL) 7:1). There are approximately 100 officially recognized national quality registers in Sweden, covering different phases of care (latency, acute, investigation, planning, intervention, follow-up and rehabilitation). SHAR covers interventions and follow-up regarding hip arthroplasty. It is the second oldest quality register in Sweden, preceded only by the Swedish Knee Arthroplasty Register (SKAR). It is also the oldest national hip arthroplasty register in the world.[60] In 2019, the register comprised 470,000 primary hip arthroplasties and 85,000 re-operations for 370,000 patients.[60]

Peter Herberts was responsible for arthroplasty surgery at the Sahlgrenska hospital in Gothenburg. In 1976, he initiated a national register of re-operations after THA. It started as a research project for 18 months. The effort was well appreciated and the need to study re-operations, especially revisions (Re-operation including replacement or extraction of any part of the prosthesis), was well acknowledged. The first of January 1979, The National Register for Total Hip Arthroplasty saw the light of day (Figure 1.3), still as a research project for the first ten years.[61,62]

Aggregated data for primary surgery were reported annually from each participating hospital. Only re-operations were recorded in detail for each patient identified by their PIN. Primary surgery has been recorded for each patient since 1992 and detailed prosthesis data since 1999. A web platform was released the same year, allowing participating hospitals to report and access their own data through an on-line interface. THA has been recorded from the start, and hemiarthroplasty since 2005. In 2017, the database was migrated to its current IT-platform, Stratum, maintained and develop by the Centre of registers (RC) in Region Vastra Gotaland (VGR).

All private and public hospitals performing hip arthroplasty surgery in Sweden participate in the register, yielding 100 % coverage. The completeness of primary surgery was above 98 % for THA and 96 % for hemiarthroplasty in 2016.[63] The register contains several modules with different units of interest (Table 1.1). The data base is linked to the national population register and therefore includes death dates for patients who are no longer alive. Each re-operation is recorded as either revision (some component replaced or extracted), or as any other type of open surgery performed to the hip. Each hip can have multiple re-operations. An accompanying component database is used to store details of each prosthesis model such as dimensions, materials, producers and more.

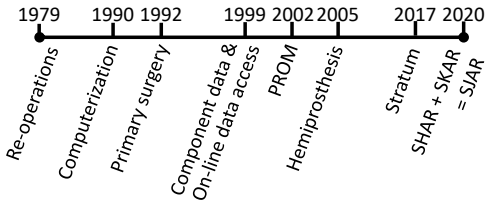

Figure 1.3: Timeline with important dates of The Swedish Hip Arthroplasty Register (SHAR).[60] (SKAR = The Swedish Knee Arthroplasty Register. SJAR = The Swedish Joint Arthroplasty Register. Stratum = on-line IT-platform).

Table 1.1: Modules of the Swedish Hip Arthroplasty Register. (PROM = patient reported outcome measures)

Table	unit	started
Primary surgery	Hips	1992
Re-operations	Re-operations	1979
Component data	Components	1999
Environment data	Hospital by year	1969*
PROM**	Patients by date	2002/2008***

* Few registrations in early years.
** Pre-operative, 1-, 6-, and 10-year postoperative.
*** Some hospitals since 2002; national coverage since 2008.

This is of interest to manufactures for post-market surveillance. Environment data is recorded annually, including data on operating facilities that are not changed between surgeries. Patient reported outcome measure (PROM) are centered around each patient and has been collected nationally since 2008.[64] Patients with elective surgery respond pre-operatively, as well as one, six and ten years post-operatively. Patients with FNF participate only post-operatively.

SHAR and the Swedish Knee Arthroplasty Register (SKAR) formally merged in 2020 to become the Swedish Joint Arthroplasty Register (SJAR). We still use the old name in this thesis since most of the work was performed prior to this merge.

1.10 NJR

The National Joint Registry for England, Wales, Northern Ireland, the Isle of Man and the States of Guernsey (NJR) was established in 2002 and has published annual reports since 2004. The registry holds more than 2.8 million records for five joint replacement procedures: hips, knees, ankles, shoulders and elbows. More than one million records concern hip arthroplasty. It is the largest arthroplasty register in the world. Increased collaboration is planned between NJR and other orthopedic registries in the UK to form the National Musculoskeletal Registry (NMR). The registry is part of the National Health Service (NHS) and is led by a steering committee.

Reporting to the register is mandatory for all NHS trusts and foundation trusts within NHS England, as well as for all NHS Wales hospitals.[65] The register coverage is thus 100 % for such hospitals, although privately founded hip arthroplasty is not included. Patient participation in the register is based on informed consent. Consent rates varies slightly between years and regions but were 92.3 % in both England and Wales in 2018.[66] This would constitute the completeness of the register within the covered hospitals.

Research data from the register is provided through a data access portal after permission by a research committee. Provided data sets incorporate some pre-specified linkage by individual NHS-numbers, or by name, age, sex and address. This includes the Hospital Episodes Statistics (HES) registry (comparable to the Swedish NPR), as well as mortality data linked from the Office of national statistics.[66]

1.11 REGRESSION ANALYSIS

Assume that Y is an outcome observed as $y = y_1, \dots, y_n$ for patients $i = 1, \dots, n$ with additional k-dimensional baseline covariate vectors $X_{i\cdot} = (1, X_{i1}, \dots, X_{ik})$. The goal of regression analysis is to relate $X = [X_1 \ \dots \ X_n]'$ to Y, involving some variable coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$ (where v' is the transpose of v), such that $g(Y) = f(X\beta + \varepsilon)$ for some functions $f, g : \mathbb{R} \to \mathbb{R}$ where $\mathbb{R}$ is the set of real numbers, and where ε is a random noise vector $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ and $\varepsilon_i \sim N(0, \sigma^2)$ with N representing the normal/Gaussian distribution with some unknown variance σ^2. The simplest form concerns linear regression with $Y \in \mathbb{R}$ and $f = g = I$, the identity function: $Y = X\beta + \varepsilon$. Assume, however, that $Y \in \{0, 1\}$. A linear relation between Y and $X\beta$ is then unreasonable, although a logistic transformation, $f(z) = 1/(1 - e^z)$, might imply a sigmoid relation between $z = X\beta + \varepsilon$ and $g(Y) = P(Y = 1) = p$. This is logistic regression, usually denoted by $p = [1 - \exp(-(\beta X + \varepsilon)]^{-1}$. The fact that $\exp(\varepsilon)$ is included as a multiplicative factor is different from linear regression, although commonly neglected in the medical literature. Logistic regression is often used for short-term mortality where $Y = 1 =$ death.

The logistic function is the inverse of the logit function, the natural logarithm of the odds of $Y = 1$ (Section 1.11.2), thus $\text{logit}(p) = \ln[p/(1 - p)] = X\beta + \varepsilon$. This is one example of generalized linear regression where $X\beta$ might be additionally transformed by elementary functions, polynomials, or splines. Further generalizations include generalized additive models, regularized regression (Section 1.14), boosted regression, and random/mixed effects models

(Section 1.11.1), as well as various combinations of those, such as fractional polynomials and splines (piecewise polynomial functions connected at certain coordinates/"knots").

Coefficients from linear regression are collapsible, meaning that their implied association, as measured by their magnitude and direction, does not change in relation to other variables in the same multivariable regression model. This is rarely true for coefficients of logistic regression with implicit dependency on the baseline levels of all other variables (the background/baseline cohort). Hence, if certain levels of a categorical variable are collapsed, this might change both the direction and magnitude of other variables, since they all relate to the background, which is no longer the same.[67]

Another generalization is piece-wise linear regression (segmented- or broken-stick regression/interrupted time series), where $z = X\beta + \varepsilon$ is partitioned into L segments $\bigcup_{l=1}^{L} \xi_l$ where $\xi_l = (s_{l-1}, s_l]$ for some break points $\{s_0, \dots, s_L\}$ with $s_0 = \min(z)$ and $s_L = \max(z)$. Individual slopes are fitted within each segment ξ_l and knots/breakpoints are chosen so that all segments are connected by one-degree splines. Optimal knots can be identified by numerical methods to maximize the likelihood of the model, given the observed data.

1.11.1 CORRELATED DATA

Traditional regression techniques assume independency among samples. For correlated data, the estimated effects can be either marginal or conditional. The difference is important, and the relevant framework should be chosen based on the question of study.

In marginal effects models, the coefficients are estimated by their average effects over intra-dependent clusters. This might be performed by generalized estimating equation (GEE) based on a quasi-likelihood estimation procedure with differential equations and numerical iterative methods. The covariance structure between samples is central, and is often modeled as a robust covariance matrix using a "sandwich estimator" (matrix notation BMB where B and M represents the container bread and the surrounded meat). It has been found in empirical studies, however, that GEE is rather insensitive to the exact matrix assumption. A "working correlation" must nevertheless be supplied by the modeler. This might be rather subjective and a simple identity matrix might suffice in absence of more intricate assumptions.[68]

Alternatively, the cluster effects could be explicitly modeled, although not necessarily estimated, by hierarchical generalized linear models (HGLM), including fixed or random intercepts, and/or possibly (but less commonly) random slopes for each cluster. A fixed effect is explicitly modeled by a dummy variable in the statistical model. Random effects are considered unknown but with known distribution (usually normal). A mixed effects model contains both fixed and random effects. HGLM implies conditional estimates, where a unit change of a covariate will have the estimated effect of the coefficient among individuals conditioned on the remaining fixed effects.

Marginal effects are also called "population-averaged models". This is a simplification since marginal models are equivalent to conditional models ignoring some known or unknown cluster effects.[69] It has therefore been argued that conditional modeling should be considered the norm.[70] GEE might, however, be preferred to HGLM for computational reasons or to avoid additional distributional assumptions of the random effects.

In linear regression, the marginal and conditional effects are the same. In log-linear models such as Poisson regression, all parameters except the intercept, will also coincide. In logistic regression with random cluster effects $r \sim N(0, \sigma^2)$ and a conditional coefficient vector β, the marginal equivalent is approximately $\beta^{\mathbf{m}} \approx (1 + 0.35\sigma^2)^{-1/2}\beta$.[69]

1.11.2 ODDS AND ODDS RATIOS

The coefficients of logistic regression are often of less relevance. Their exponentiated

form is usually of more interest since those correspond to odds ratios (ORs), comparing the odds of the outcome in groups with and without certain conditions, or with a unit change in continuous variables. Hence, $X\beta$ is regressed to the natural logarithm of the odds of an event: $O_A = p_A/(1-p_A)$ where $p_A = P(Y = 1|A)$ for some condition A. The exponentiated variable coefficients are ORs, ratios of two such odds, for patients with mutually exclusive conditions A and B (where B might equal A^C, the complement of A): OR $= O_A/O_B$. Odds ratios are different from risk ratios (relative risks): RR $= p_A/p_B$, although commonly approximated as such.[71] Relative risks are often easier to interpret and the approximation is possible for rare events and ORs relatively close to 1. A common prerequisite to assume OR $\approx$ RR is for $\hat{p} < 0.1$ combined with OR $\in [0.5, 2.5]$.[72,73] According to the central limit theorem, the odds will be normally distributed as the sample size $n \to \infty$. With small sample sizes, the estimated odds become biased and overestimated.[74] This is equally important for low effective sample sizes, thus for rare events data.

A note of caution is that different terminology is sometimes used in what is called "(clinical) epidemiology" and "mathematical epidemiology". The "case fatality rate" for example is used in (clinical) epidemiology, defined as the *proportion* of deaths among patients with a certain condition. This "rate" is thus a mathematical *probability*.[75]

1.12 SURVIVAL ANALYSIS

The first recognized study on mortality known to the western culture was performed by Graunt (Section 1.3).[12] It was a revolutionary study of its time, performed by a businessman interested in life tables and summary statistics. The theoretical foundation of survival analysis has been developed since then.

In this thesis, we measure survival time, T, as the number of days from primary surgery with hip arthroplasty until death. The cumulative survival function, $S(t) = P(T > t)$, models the probability to survive at least t days.

From now on, we will ignore the error term, ε, to simplify notation. A naive estimate of S (on a population level) is $\hat{S} = 1 - d/n$, where d is the number of deaths before t and n is the total number of patients. Life tables constructed this way have been used since the days of Graunt and by actuarial science thereafter. This estimator assumes that we know the actual death dates for all patients. This was somehow true for Graunt, since he only studied individuals who died in retrospect. For prospective studies, it is almost never the case that all *actual* survival times are known. It is common to include patients with only partially known follow-up times, who are still alive at the end of the study. Such censoring makes the naive estimator naïve. Kaplan and Meier developed a non-parametric theory for individuals followed until either death or censoring (loss to follow-up).[76] Hence, the study of *observed survival time*, from 0 to $T_o = \min(T, T_c)$ where T_c is the time of possible censoring, and T_o is the measure of observed time to whatever comes first, as indicated by a companying binary indicator variable $\delta = 1$ if $T_o = T$ and 0 otherwise.

1.12.1 HAZARDS

To study T_o and δ is challenging considering variable adjustments. An alternative approach is to study the hazards, the probability of death within a very short (infinitesimal) time dt after t: $\lambda(t)dt = P(t < T < t + dt|T_o \geq t)$. We thus state that $T_o \geq t$, hence we only consider patients with a follow-up time (until death or censoring) of at least t. Those patients are still at risk at time t, and their quantity, n_t, is used as the nominator when calculating the proportion of patients surviving the short time interval dt. We can thus estimate the hazard empirically by: $\hat{\lambda}(t) = d_t/n_t$ where d_t is the observed number of deaths between t and $t+dt$ among the n_t at risk.

We are often interested in the cumulative hazard Λ, the sum of hazards estimated at observed failure times $t_{(1)} < t_{(2)} < ... <$

$t_{(d)} < t$, given by $\Lambda(t) = \sum_{t_{(i)}<t} \lambda(t_{(i)}) \approx \sum_{i\,:\,t_{(i)}<t} (d_i/n_i) = \hat{\Lambda}(t)$, where d_i is the number of deaths between $t_{(i)}$ and $t_{(i+1)}$, and n_i is the number at risk at time $t_{(i)}$.

Hazards and cumulative hazards might be interesting in their own rights, but our main interest is the proportion of patients who survive until time t, not the rate of patients who dies during a theoretically short time interval dt. Fortunately, those concepts are related: $S(t) = \exp[-\Lambda(t)]$. The maximum likelihood estimator of S is $\hat{S}(t) = \prod_{i\,:\,t_{(i)}<t}(1 - d_i/n_i)$, the essentially unbiased product-limit/Kaplan–Meier estimator for the survival probability among patients still at risk.[76]

1.12.2 PROPORTIONAL HAZARDS

Estimation of cohort survival or hazards is one part of modeling survival. Another is the ability to test differences or equivalences between patients with different conditions. A null hypothesis for two groups of patients would be $H_0 : S_1(t) = S_0(t)$, where S_0 and S_1 denote the survival of each group. Thus, if H_0 is true, there would be no survival differences between those groups.

The corresponding default alternative hypothesis $H_1 : S_1(t) \neq S_0(t)$ is generally too broad, however. Alternative approaches have been debated. One solution is based on rank tests, the Lehmann alternatives.[77] Assume that S_1 is specified by a function $g(x) = x^{\varphi}$ of S_0. Thus $S_1(t) = S_0^{\varphi}(t)$, or equivalently $\lambda_1(t) = \varphi\lambda_0(t)$ for some proportionality constant φ. The practice to compare two groups could be generalized to include covariates by assuming $\varphi = \exp(X\beta)$. Then, the HR indicates the relative increase or decrease in hazard comparing two groups, as differentiated by $X_{\cdot j}$ with associated β_j: $\text{HR} = \lambda_1/\lambda_0 = \varphi\lambda_0/\lambda_0 = \exp(\beta_j X_{\cdot j})(\lambda_0/\lambda_0) = \exp(\beta_j X_{\cdot j})$. There are three scenarios: (1) $\beta_j < 0$ (HR < 1) means that patient i with $x_{ij} = 1$ have a lower hazard than patients with $x_{\cdot j} = 0$, and therefore better survival. $X_{\cdot j}$ thus has a protective effect; (2) $\beta_j = 0$ (HR = 1; as coherent with H_0) means that both groups are similar with no excess hazard for any group and; (3) $\beta_j > 0$ (HR > 1) means that patient i with $x_{ij} = 1$ have larger hazard and lower survival compared to patients with $x_{\cdot j} = 0$; thus $X_{\cdot j}$ has an adverse effect. Similar reasoning applies to a unit change of a continous variable $X_{\cdot j}$.

1.12.3 COX REGRESSION

Sir David Cox suggested to regress the hazard to covariates on the form $\lambda(t) = \lambda_0(t)\exp(X\beta)$.[78] He called it proportional hazard regression, often referred to as Cox regression. The effect of a change in the covariates is associated with the instantaneous probability of death for patients who survived up to at least time t. The instantaneous death rate at any given time during the follow-up is HR times higher (or lower) among one group compared to another, if assuming proportional hazards. Note that HR is a relative *rate*, $\text{HR} = \exp(\beta_j)$, and not a relative risk/probability, which would be $[1 - S_0(t)\exp(\beta_j)]/[1 - S_0(t)]$, where $S_0(t)$ is the baseline survivor function at time t (the probability to survive at least until t among the "controls"). HR and the relative risk have similar directions but can differ in magnitude.[71] An approximation of HRs as relative risks might be acceptable for outcomes with rare events and low HRs.

Hence, modeling β this way assumes proportional hazards, hence that $X\beta$ is constant over time. If not, $\hat{\beta}$ as the average effect, might be irrelevant, or even misleading, if it no longer corresponds to any actual value of β at any observed point in time. David Schoenfeld took his PhD two years after the Cox regression model was proposed. Six years later, in 1980, he suggested a Chi-square (χ^2) test to evaluate the goodness-of-fit for the proportional hazards assumption.[79] Two years later, he also proposed a graphical method later known as Schoenfeld residuals.[80] Such residuals are asymptotically independent. Hence, if the proportional hazards assumption holds, expected values for the residuals are 0 and plotting the residuals versus time will depict random noise scattered around 0. If this is not the case, the proportional hazards as-

sumption might not hold, and Cox regression might be inadequate.

1.12.4 EXTENDED COX REGRESSION

If proportional hazards cannot be assumed, hence if $X\beta$ is time dependent, alternatives to the regular Cox regression model must be considered. There are two alternatives: time-dependent covariates, $X = X(t)$, or time-varying coefficients $\beta = \beta(t)$. The age of a patient for example will inevitably change and is therefore a time-varying covariate (the age at surgery, however, will stay constant). Time-dependent covariates are either internal or external/exogenous.[81,82] An external covariate is not directly related to the failure mechanism, such as age. Age is deterministic based on date of birth and the time passed since then. It is possible to calculate hypothetical ages even for patients that are no longer alive. Internal time-dependent covariates are harder to predict. They depend on the individuals being studied. Blood pressure or smoking status are classical examples.

Alternatively, the coefficients might vary with time, $\beta = \beta(t)$. Even if the set of comorbidities are fixed before surgery, the effect of those comorbidities on mortality might change. This could happen for example if new treatments make the presence of a certain comorbidity less severe over time. We will assume that this time-dependent change can be captured by a function h, such that $h = h(\beta, t)$. This might not be true for individual patients but could work reasonably well when aggregated to the population level. We can thus rewrite the hazard as $\lambda = \lambda_0 \exp[Xh(\beta, t)]$. Assume for example that the effect of pre-surgery drug abuse diminishes (wash out) with time if the patient stops using drugs. Thus, assume that $h(\beta_j, t) = \beta_j t^{-1}$ which represents an interaction effect between β_j and a function h of t. Hence, $\lambda = \lambda_0 \exp[Xh(\beta_j, t)] = \lambda_0 \exp(X\beta_j t^{-1})$.

Another scenario is to assume constant effects during limited time intervals. This could be relevant for some comorbidities, for example some cancer diagnosis, with initially high mortality. Patients surviving more than five years, however, might be considered "statistically cured", thus with a lower remaining risk of death. It is then reasonable to assume that hazard ratios for two groups are proportional within each interval, although different in different intervals. We can define h as for example $h(\beta_j, t) = \beta_j \cdot [I(t < 5) + I(t \geq 5)/10]$ where $I(t < 5) = 1$ if $t < 5$ and 0 otherwise (vice versa for $I(t \geq 5)$). This is a Heaviside step function, assuming that the association between cancer and mortality will decrease to only 10 % after year five. The number of intervals is arbitrary. Interaction modeling with discrete time intervals is similar to stratification since no general time- or covariate effects are assumed.

It is also common to represent the occurrence of death as a counting process.[82,83] This is mostly a data management tweak to transform the data for easier calculation. It is done by including each patient once per interval, if he or she is still at risk (alive). A cancer patient who dies after 7 years will then be included once for the first five-year period, and once again for a second period from year 5 to 7. In practice, cut points are set at each observed time of death, $\{t_{(1)}, \ldots t_{(d)}\}$.

1.12.5 RMST

A basic summary of cohort survival is the median survival time. Apart from censoring, this value can be estimated as soon as half of the cohort has died. It is a popular measure in medical research. The mean on the other hand: $\mu = E[T] = \int_0^\infty S(t)dt \approx (1/n)\sum_{i=1}^n t_i = \bar{t} = \hat{\mu}$, can only be estimated if all t_i:s are known. This is unrealistic even with a good population register, since many patients with hip arthroplasty survive decades after surgery. We would need to wait until they all die before calculating $\bar{t}$. This estimate would thus be obsolete even before calculated. The restricted mean survival time (RMST) is a shortcut, where we only consider the mean survival time up to a time point τ,* instead of ∞ in theory or

*Also known as t-year mean survival time[84], restricted mean event time[85] or restricted mean lifetime.[86]

$t_{(d)}$ in practice; $\mu_\tau = \text{RMST}_\tau = \int_0^\tau S(t)dt \approx \int_0^\tau \hat{S}(t)dt = \hat{\mu}_\tau$.[87] If, for example, $\tau = 10$, we only need survival data for the first 10 years after surgery. μ_{10} would never exceed 10, and it would most likely fall strictly below. The restricted mean time lost (RMTL) is $\tau - \mu_\tau$, the remaining gap between RMST and τ. Both measures are asymptotically normal: $\text{RMST}_\tau, \text{RMTL}_\tau \sim N(\cdot, \sigma^2)$ with variance $\sigma^2 = 2\int_0^\tau S(t)dt - [\int_0^\tau tS(t)dt]^2$,[88] estimated by $\text{Var}(\hat{\mu}_\tau) = \frac{n}{n-1}\left[\sum_{i'=1}^{n}\left[\sum_{i=i}^{n}(t_{(i+1)} - t_{(i)}\hat{S}(t_{(i)})\right]^2 \frac{1}{n_{i'}(n_{i'}-1)}\right]$ where $t_{(i)}$ is the i^{th} smallest of the sorted event times (possibly adjusted to separate ties) and n_i is the size of the risk set prior to $t_{(i)}$.[89]

μ_τ is easily estimated as the area under the survival curve $S \approx \hat{S}$, where $\hat{S}$ can be calculated by a method of choice, commonly by the Kaplan–Meier estimator.[76] The area is then estimated by numerical integration, i.e. by the trapezoid rule. Survival for two groups, S_0 and S_1, can be subtracted and compared without the need of the proportional hazards assumption:[84] $D = \hat{\mu}_{1\tau} - \hat{\mu}_{0\tau} = \int_0^\tau [\hat{S}_1(t) - \hat{S}_0(t)]dt$ with $\text{Var}(D) = \text{Var}(\hat{\mu}_{1\tau}) + \text{Var}(\hat{\mu}_{0\tau})$.[90] Hypothesis tests and confidence intervals (CIs) can be based on the normality assumption, although the standard Wald test might be sub-optimal for small samples and large censoring proportions.[91]

One way to adjust for covariate effects is by pseudo-observations regressed by GEE. First, each t_i (including censored values where $\delta = 0$) is replaced by pseudo-observation: $\hat{\mu}_{\tau,i} = n\hat{\mu}_\tau - (n-1)\hat{\mu}_\tau^{-i} = \int_0^\tau [n\hat{S}(t) - (n-1)\hat{S}(t)^{-i}]dt$ for each patient i, where $\hat{\mu}_\tau^{-i}$ and $S(t)^{-i}$ are estimates of μ_τ and $S(t)$ for all patients except the i^{th}.[92] The pseudo-observations thus represents the contribution of each individual to the overall aggregate (μ_τ) and transforms the possibly censored data into a data set without censoring.[93] This is useful in many settings, additional to RMST, considering survival analysis with censored data.[94]

Estimations of the pseudo-values largely consider the same data, implying dependency among observations. GEE is therefore preferred to regress those on potential covariates and to model the marginal effects of each covariate (Section 1.11.1). If assuming a working correlation of 1 (the scalar version of an identity matrix),[92] the β-vector could be estimated by partial differential equations solving $0 = \sum_i \frac{\partial}{\partial\beta}(\beta' X_i)'(\mu_{\tau,i}^{(\gamma)} - \beta' X_i)$ where $\mu_\tau^{(\gamma)}$ represents the RMST among patients in group γ (i.e. patients with a certain comorbidity score).[95] Such β is asymptotically normally distributed.[94]

1.13 PREDICTION

There are essentially two aims of statistical modeling: to explain (describe/understand) or to predict.[96–98] Study II–IV were focused on prediction, although some degree of understanding is always necessary. Prediction modeling is the term used in statistics, but terms like prognostic modeling, prediction rules, risk score, forecasting or foreseeing are also used in the medical field.[99] Understanding a process usually involves estimation of association and effect sizes, such as β-coefficients or HRs. Overly complicated models and methods are not desirable since those might lack a natural interpretation. For predictions on the other hand, we might accept the use of "black boxes" (opaque models). It is often less important exactly why predictions work, as long as they do. It is therefore sometimes recommended to include as much data and as many variables as possible when building a predictive model. However, complex models using administrative data might not be feasible if this data is not accessible in a clinical setting where future predictions are supposed to be made. This differentiate the clinical setting from weather forecasting or financial model building, where super computers and specialists work full-time.[100]

On a conceptual level, the goal of prediction is to find a model M, which could predict the relevant outcome Y from covariates X, present at the time of prediction. We can conceptualize M as a function f, esti-

mated by $\hat{f}$ using a training data set, with patients for whom $X = x_d$, and $Y = y_d$ with d denoting the derivation/training cohort. The possibility of $\hat{f}$ to approximate f is then evaluated based on observed covariates $X = x_v$ where v indicates an independet validation sample. Observed outcomes $Y = y_v$ are compared to predicted $f(X = x_v) = \hat{Y}_v$. The better those values resemble each other, the better approximation of $\hat{f}$ to f, and the higher predictive power we have. The comparison is made by a loss function l such as the root-mean-square error (RMSE): $l(y_v, \hat{Y}_v) = \sqrt{(1/n_v)\sum_{i=1}^{n_v}(y_{vi} - \hat{Y}_{vi})^2}$ where n_v is the number of patients in the validation cohort. This is slightly different from explanatory modeling where $\hat{f}$ relies on predefined theories and where the focus is on the functional form of f. Also, the model fit is usually assessed by the same sample as used to estimate $\hat{f}$.

For prediction modeling, we do not need to accurately estimate each β_j.[101] We only need $\hat{Y} = \hat{f}(x\hat{\beta})$. This is the reason why a pre-calculated comorbidity score such as Charlson or Elixhauser might work. This is somehow similar to statistical techniques such as principal component analysis (PCA), aimed to reduce the dimensionality of the model.

The model M could be based on complex machine learning (ML) algorithms or artificial intelligence (AI). We focus on traditional statistical regression modeling, however, since those models are generally easier to interpret and since earlier results have shown limited or no benefits for more complex models in settings similar to ours. Prediction modeling has been applied to several post-operative conditions after hip arthroplasty. Mortality is most commonly studied within a relatively short period after surgery, sometimes limited to in-hospital deaths in countries without centralized PINs.[102–105] Unfortunately, not all published prediction models relies on sample sizes large enough,[106] or use the optimal statistical technics, to achieve unbiased and truly useful models.[107]

1.14 VARIABLE SELECTION

It is common that patient data used for prediction modeling include only a few strong predictors and several weaker ones.[108]

It is then desirable to identify those important variables, not only to simplify the model (Ockham's Razor), but also to avoid the risk of over-fitting and to exclude variables that are truly irrelevant for generalizations outside the training set. There are several ways to perform variable selection. One is to only consider potential predictors with a known (or hypothesized) relation to the outcome. This is not always possible, wherefore statistical procedures might be considered. Traditional methods such as univariable screening is an easy first step, although not generally recommended. Here, the analyst uses separate univariable models, regressing each potential predictor to the outcome. A hypothesis test is performed with $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ and all variables with a p-value above a predefined threshold are excluded before any multivariable regression. Such threshold is usually higher than the traditional 0.05, for example 0.15 or 0.2, since the purpose is not to draw conclusions of individual effects but only to screen out variables that are truly irrelevant.

A more elaborate version is to fit multiple multivariable models iteratively, including and excluding variables based on intermediate results. This process is known as stepwise regression, including forward selection or backward/bidirectional elimination.[109] Nested models are compared by some criteria, for example the Akaike information criteria (AIC) or the Bayesian information criteria (BIC) $a_c - 2\ln\hat{L}$ where $a_{AIC} = 2k, c_{BIC} = k\ln n$, k is the number of parameters estimated by the model (including dummy variables and polynomial coefficients), n is the sample size, or better the effective sample size (the number of cases with the least probable outcome).[110] $\hat{L}$ is the maximum likelihood conditioned on the observed data. A model with a low AIC-/BIC-value is preferred to models with higher values.

Univariable screening and stepwise regression are both criticized for the risk of over-fitting the model to the data. An alternative approach is regularized/penalized regression to shrink the estimated coefficients at the time of parameter fitting:[111] $\beta = \text{argmin}_{\beta \in \mathbb{R}^{k+1}} \|D(y, X\beta)\|_2^2 + \lambda \|\beta\|_r^r$, where $\mathbb{R}^{k+1}$ is the $k+1$-dimensional real-valued vector space containing the model parameters (including the intercept). D is here the deviance, a common loss function in logistic regression with $D(y_i, \beta X_i) = 0$ for $y_i = \beta X_i$ and $D > 0$ otherwise. $\|z\|_2 = \sqrt{\sum_{i=1}^{n} z_i^2}$ the sum of squares based on the Euclidian- or L_2-norm, $\|z\|_1 = \sum_{i=1}^{n} |z_i|$, the Manhattan- or L_1-norm, $|z_i|$ the absolute value of z_i ($|z_i| = z_i$ if $z_i \geq 0$ or $-z_i$ if $z_i < 0$). $\lambda \in \mathbb{R}^+$ is a positive tuning parameter used as a penalty/regularization term, often chosen to minimize the loss function based on v-fold cross validation (Section 1.15).

Ridge regression, with $r = 2$, shrinks the parameters closer to each other, thus limiting the effects of individual parameters. Least absolute shrinkage and selection operator (LASSO)-regression, with $r = 1$, excludes seemingly irrelevant factors by faster shrinkage to 0. Thus, LASSO is a method for variable selection, while ridge regression can be combined with LASSO in "elastic net" regression to achieve the same goal. Iterative numeric methods are used for parameter estimation. A method has been implemented in the popular `glmnet` R-package, which is computationally more efficient than the traditional numerical method by Newton–Raphson.[112]

Even though the penalty term used in LASSO will itself lower the risk of overfitting, an additional safety net is to simultaneously vary the underlying training data by at least 100 bootstrap replicates (Section 1.15).[113–117]

1.15 MODEL VALIDATION

Model validation is important for prediction modeling, although often neglected.[118–120] There are four important aspects when evaluating a prediction model:[110,121] (1) discrimination, how good is the model to distinguish between patients who do, or do not, experience the event of interest (rank-correlation); (2) calibration, how similar or dissimilar are the observed and predicted values; (3) transportability, is the model generalizable to another population; and (4) clinical usefulness, is the prediction model useful as a shared decision making tool in clinical practice?

There are three ways to evaluate 1-3: (1) internal validation asserting reproducibility, either to split the data to one part for training, and one for evaluation, or preferable, to use techniques such as cross-validation or bootstrapping; (2) temporal validation, to train the model on data from one period and to evaluate it with data from a later period; and (3) external validation asserting transportability: to train and validate the model on different, yet comparable, populations.

The focus of this thesis was internal and external validation using cross-validation, bootstrapping, and an external cohort.

For v-fold cross-validation, the sample with n patients is partitioned into v groups of approximately equal size. The model is trained on $v - 1$ of the partitions and evaluated by a loss function using data from the v^{th} partition. This is repeated v times and the results are averaged. The whole procedure might be repeated several times with new v-fold partitions. 10×10-fold cross validation is a popular setting. If $v = n$ (the sample size), only one item is left out for validation each time. This is known as Jackknife resampling.

For bootstrapping, we repeatedly draw new samples from the initial sample. The sample size is retained using sampling with replacement. Hence, the same patient can be included more than once. A point estimate of p, a parameter of interest, is given by: $\hat{p} = (1/B) \sum_{l=1}^{B} \hat{p}_l$ where B is the number of re-samples (often multiples of 100 or 1,000) and $\hat{p}_l$ is the parameter estimate from re-sample l. Measures of uncertainty, such as an empirical CI, is estimated by relevant quantiles of $\hat{p}_l$, usually considering the range after excluding the smallest and largest 2.5 %

of all $\hat{p}_l$:s. Bootstrap validation is usually preferred to the other methods due to stable estimates with low bias.[122,123] To increase B will yield better estimates, but the method is computer intensive and potentially time consuming.

Both cross-validation and bootstrapping might be hampered by rare events data, where one partition or re-sample might lack events. Stratified versions have been proposed to retain the initial proportions.

1.15.1 DISCRIMINATION

The discriminative ability of a prediction model is measured by rank similarity between observed and predicted outcomes in a sample other than the training set, either in out-of-bag samples (Non-sampled data used for internal validation), or in a different cohort for external validation. Assume that the observed outcome is binary (dead or alive), while the predicted outcome is a probability of such event. Certain thresholds could then be applied to suggest that a patient will die if the probability of death exceeds 0.5 (or any other value) given his or her covariates. Then, a simple classification of predictions versus outcomes would consider four distinct groups of patients: dead patients predicted to (1) die or (2) survive, as well as surviving patients predicted to (3) survive or (4) die.

All patients must fall in one of those categories. A better model will result in larger proportions in category 1 and 3, indicating sensitivity and specificity. All patients must also fall in *only* one of those categories. Thus, an increase in sensitivity might decrease specificity (although rearrangements using all four categories make this relation less direct). The relation between those categories might be presented as a 2×2 confusion matrix, and several measures of discriminative ability might be derived from it (true/false positive/negative rates, recall, fall-out, probability of detection/false alarm et cetera), each on the $[0, 1]$ scale.

Instead of a fixed threshold, we might illustrate the relation between sensitivity and specificity for all possible thresholds by a receiver operating characteristic (ROC) curve where the x-axis marks the false positive rate ($1-$ specificity), and the y-axis the sensitivity/recall.[124] Such curves are necessarily monotone (non-decreasing). A model without any discriminative ability will be presented as a diagonal straight line from origo to the upper right corner of the quadrant. A model with better discriminative ability will yield a concave curve closer to the upper left corner.

The ROC curve as a whole is a useful measure of the relation between sensitivity and specificity, since the give-and-take between the two is not necessarily symmetrical.[125] It is nevertheless common to seek for a summary measure to convey all information as dense as possible; the area under the curve (AUC). The worst possible model (with a ROC-curve mimicking the diagnoal line) yields AUC $= 0.5$, thus the area of half the unit square. Better models will cover larger areas, yielding a maximal value of AUC $= 1$. Obviously, the area alone cannot distinguish models with high sensitivity but low specificity from models with low sensitivity but high specificity for certain thresholds. If one of those measures is more important than the other, different weighting schemes could apply. It is often recommended to present the full ROC curve, and not only the summarized AUC value, although the actual benefits have been debated.[126] The area is calculated by integration or estimated by numerical methods such as the trapezoid rule.

AUC is often used as a concordance index (C), interpreted as the probability that given two patients, one survivor and one who dies, the prediction model will assign a higher probability of death to the latter.[127] C-indices from the medical field are often interpreted as moderate if $C \in [0.5, 0.7)$, good if $C \in [0.7, 0.8)$, and excellent if $C \in [0.8, 0.9)$. $C \in [0.9, 1.0]$ might be considered suspicious and attributed to over-fitting in settings with observational data and many unmeasured covariates.[128] Such C-values are more common in physics with data from carefully controlled experiments. The AUC-value has

become almost synonymous with a concordance index for survival data, at least with logistic regression modeling in-hospital deaths or short-term mortality, where AUC coincide with Harrell's concordance index c, and with the Wilcoxon–Mann–Whitney–U-statistic.

Some additional complexity arise for survival data with censoring. Several generalizations of sensitivity and specificity exist in this case.[129] Kaplan–Meier estimates are generally insensitive to censoring, but not for estimating sensitivity and specificity. Both measures might decrease simultaneously in the presence of censoring, implying a non-monotonic ROC-curve with sudden drops. Heagerty et al. repaired such curves using estimates from near-by points (a kernel function with a smoothing parameter based on nearest-neighbors).[129] A semiparametric model with a bi-variate distribution considering covariates X and time t, was used for ROC-curves of cumulative survival accounting for censoring. Five years later, the authors modified their proposal to an incidence time-dependent ROC-curve.[130] Hence, considering death at time $t + dt$ for an infinitesimal time dt. This required an updated, dynamic, risk set of patients at time t (compared to previous models using a static risk set from time 0). Thus, each patient was initially considered alive (a "control"), but then transferred to a "case" at the moment of death. This way, we can evaluate the discriminative ability of the model for instant death at time t by a value AUC(t) estimated by numerical integration or weighted averaging. Then, integrating once more over the dimension of time, yields a time averaged AUC up to time τ: $C^\tau = \int_0^\tau \text{AUC}(t) \cdot w^\tau(t)dt$ where $w^\tau(t)$ are regularization weights such that $\int_0^\tau w^\tau(t)dt = 1$. Heagerty et al. described C^τ as the probability that the predictions for a random pair of subjects are concordant with their outcomes, given that the smaller event time occurs in $(0, \tau)$.[130] Adaptations for non-parametric survival models have been suggested as well.[131] It has been proven that the time-dependent AUC is a proper concordance index for estimation of mortality at fixed time points.[132]

1.15.2 CALIBRATION

Calibration is another important aspect of model validation, indicating proximity between observed and predicted values. Discrimination might influence calibration but the relation is modified by case mix, wherefore both measures should be assessed in tandem.[133] Unfortunately, this is often neglected; only one third (25 of 78) of medical prediction studies assessed calibration according to a systematic review in 2014.[119]

An over-all heuristic for a binary regression model is "calibration-at-large", comparing the observed and estimated event rates. Calibration is otherwise most intuitive for linear regression, comparing the observed y and the predicted $\hat{y}$. The residuals, $\varepsilon = y - \hat{y}$, should be as small as possible, preferably independently and randomly distributed with a normal distribution $N(0, \sigma^2)$ for some variance σ^2. There are conceptual similarities with logistic regression, although the observed values are always 0 or 1, and therefore not directly comparable to the predicted probabilities $\hat{p}_i \approx P(Y = 1|X_{i\cdot}) \in [0, 1]$. Instead, predicted probabilities might be ranked and partitioned into deciles. The sum of observed events within each decile $l \in \{1, \dots, 10\}$ is $o_{1l} = \sum_i y_i$ which might be compared with the sum of estimated/expected probabilities $e_{1l} = \sum_i \hat{p}_i$ by $X^2 = \sum_{l=1}^{10}(o_{1l} - e_{ll})^2/e_{1l}$. This is a classic approach, later improved by Hosmer and Lemeshow,[134] who introduced the corresponding $e_{0l} = \sum_i(1 - \hat{p}_i)$ and $o_{0l} = \sum_i(1 - y_i)$ and who proposed two summary measures: $C^* = \sum_{k=0}^{1}\sum_{l=1}^{10}(o_{kl} - e_{kl})^2/e_{kl}$ and H^*, a similar measure based on fixed partitions of p rather than observed deciles of $\hat{p}$. $C^*, H^* \sim \chi^2_{g-1}$ where g is the number of groups (10 for deciles, which could be generalized). Austin and Steyerberg provides a brief summary of several adaptations and modifications (analytical and graphical) based on those statistics.[133] A simple hypothesis would consider $H_o : C^* = 0$ versus $H_1 : C^* \neq 0$. We could also plot $o_1 = (o_{1,1}, \dots, o_{1,10})$ versus $e_1 = (e_{1,1}, \dots, e_{1,10})$ and hope for a straight line assessed by linear

regression* evaluated by its estimated slope and intercept. A good model would have a zero intercept and a unit slope. Deviating lines would indicate miss-calibration. To fit such a straight line is common, but it is only an over-all assessment which is not informative considering alternative forms if the relation between o_1 and e_1 is non-linear. It is actually a measure of discrimination rather than calibration.[135] An alternative regression model for external validation was proposed by Finazzi et al.[136] They suggested to compare a flexible calibration curve based on a parametric model with fractional polynomials including terms up to a certain degree chosen by forward selection. Proximity between such line and the theoretically preferred straight bisector would graphically indicate good calibration. The graphical approach was later accompanied by an analytical test based on cumulative distributions,[137] and the framework was later extended to internal validation as well (goodness-of-fit).[138]

Logistic regression models applied to rare events data often under-estimate the probabilities in external validation data sets. Thus, even if the discriminative ability persists, the calibration curve often falls below the optimal bisector. Re-calibration of the intercept and an over-all slope has been suggested as a relatively simple solution, whereas model revision might be necessary for larger discrepancies.[139] Such re-calibration might also be necessary to adjust for population differneces between populations used for model derivation and validation.[140]

1.16 STATISTICAL SOFTWARE

One (of many) definitions of big data considers volume: to analyze data larger than the computers random access memory (RAM). Intermediate processing steps for our comorbidity data (Section 1.6) reached this limit. We used R as a computer environment throughout the thesis and some background might be of relevance for Study I, in which we developed an R package to estimate comorbidity from large data sets.

*Logistic regression considering y and $\hat{y}$.

1.16.1 S AND S-PLUS

Before R, there was S. S was initially developed by statisticians, most famously John Chambers, at Bell Labs (currently part of Nokia and previously of AT&T) in 1975–1976.[141] Fortran, the first high level programming language developed in 1954-1957,[142] was previously used for statistical analysis within the company. S was less declarative, did not require pre-compilation, and was used interactively through a UNIX prompt. It included graphical procedures to enhance visualization. UNIX, as well as the C programming language had been previously developed at the Bell Labs. It was therefore natural to make S part of the same eco-system. The name S itself also elude to C. Functions and vectors were fundamental parts of S and numerical vectors of length one were used instead of scalars. A dimension attribute was used to emulate more complicated matrices, arrays, and general (possibly nested) list objects. There were no distinctions between integers†, floats‡ and doubles,§ which were all considered "numeric" (although whole numbers suffixed by `L` were passed as integers to underlying layers). The design was meant to be simple. Functions were used for all sorts of data manipulation. Infix operators¶ were added only as syntactic suger.‖ Functions could have an arbitrary number of arguments and those could also be left unspecified by the user if default values were already provided by the developer. Vector-subscription was very flexible, both compared to mathematical notation, as well as to other computer languages. It was easy to import subroutines from Fortran, and later C. Version 2 included routines for miss-

†Whole number used by computers where the range of available numbers depends on the operating system.

‡Computer approximation of real numbers.

§Legacy term for binary64, a float number with double precision used by computers.

¶Programming operation similar to a function but with differnt syntyx, for example the arithmetic operators (+, -, / and *).

‖Design elements of a programming language not introducing any new functionality but which improves clearity, concistency or which introduce an alternative programming style.

ing data, loops and character handling. Factor vectors were later introduced as numerical vectors with labels. Those details might seem minor, but they were novel at the time, and some of them still are, compared to other computing environments used for data analysis.

Prior to 1984, S was distributed ad hoc as open source software from the developers themselves, to research facilities and universities. From 1984, the software got licensed and officially adopted by the AT&T sales organization. Major changes occurred until 1988 when the New S was released. New S made user defined functions first class objects, and later contained debug- and trace functionality. A new object-oriented approach was also introduced. Objects (including data structures and functions) could then be modified and re-allocated. This is slightly different from some other object-oriented languages, which also allow in-place modifications of existing objects. The class system also made object dispatch possible and more abstract methods were easily applied to objects of different classes. A minor detail was the introduction of "<–" as assignment operator (due to a custom key on the Execuport keyboard used to develop S).[143] Another notation detail, "~", was introduced in 1993 for formula objects.

S-plus was a commercial product sold by Statistical Sciences from 1988. It enhanced S as one of many possible implementations of the language. In 1993, however, Bell Labs sold an exclusive license of S to Statistical Sciences. The open source language was still developed at Bell Labs, but the only way to get a commercial license was via S-plus. After several acquisitions and merges, both S and S-plus got owned by Insightful in 2004. Tibco bought Insightful in 2008, one year after their acquisition of Spotfire, wherein S-plus was soon integrated. Version 8.2 of Tibco Spotfire S+ had its latest update in 2012.[144] Those details might be interesting trivia since both Spotfire (founded in 1997 by Christopher Ahlberg)[145] and the office of SHAR, are located close to Linnégatan in Gothenburg.

1.16.2 R AND R-PACKAGES

R was created in 1992 with its first binary beta release in 1993 (the same year as the exclusive license of S to S-plus), by *R*oss Ihaka and *R*obert Gentleman at the university of Auckland, New Zealand. Thus, the history of S ends close to SHAR, while the history of R began as far away as ever possible. In 1995, R was released as an open source software under the GNU's Not UNIX (GNU) General Public License (GPL). The R-language is sometimes distinguished from this "GNU-R" as one of several (possible and actual) implementations.* The language specification is not well-defined, however, making the GNU-implementation almost an integral part of the language itself.[146] The implementation of R got very similar to S and most user-written computer scripts could be executed in both environments. Initial development was made by e-mail exchange and by sending floppy disks. A group called the R Development core team was founded in 1997 to formalize the development.

R version 1.0.0 was released 2000-02-29.[147] The date was chosen carefully. 2000-01-01 was a dangerous date in software development considering the Millenium bug.† February 29 was its less known sibling, a leap day in a year divisible by 400, the second of its kind since 1582, when the Gregorian calendar established the current leap rules.

The R foundation was established in 2003 to hold copyright and to provide a reference point for further R development.[148] The first international user conference was held in Vienna the following year. R was initially used in academia, but several commercial companies have later made large contributions to the R community. The R consortium was therefore founded in 2015 to secure the future of R and to support the R foundation, as well as different user initiatives. The R foundation, supported by the R consor-

*Tibco Enterprise Runtime for R, the successor of Tibco Spotfire S+, being one of them.

†Also known as the year 2000/Y2K problem/bug/glitch; a problem caused by the two digit abbrivation of years such that year 2000 was not distinguised from 1900

tium, ensures that the R Development core team develops and maintains GNU-R. The popularity of R, however, could be largely attributed to the large number of available user-developed extensions, R packages. The Central R Archive Network (CRAN) was suggested by Kurt Hornik in 1996. He and other volunteers have since then maintained a library containing more than 15,000 packages, growing exponentially.[149]

There are several class systems implemented in R, allowing methods dispatch and generic functions for ad hoc polymorphism. R3 was the first implementation and is still the most popular.

Although R itself is a programming language accessed by command-line tools, several integrated development environments have been developed over the years. The most successful, provided by RStudio since 2011, has become almost synonymous with R itself. RStudio is a certified B corporation since December 2019. As such, it holds responsibility not only to its shareholders, but also to users and other stake holders. Another software from RStudio is the `tidyverse` package suite (initialy `hadleyverse` after its creator Hadley Wickham). It is a unified framework with multiple packages for data management and graphics. It has become almost its own paradigm parallel to "base R", with its own domain specific language (DSL) and design philosophy. One signature feature of `tidyverse` is the linear flow of object modifications using "verbs" that are chained, or piped, by the operator `%>%`.[150] The concept is known from UNIX and relies on functional composition, $\circ$, such that two functions f and g are combined as: $(f \circ g)(x) = f[g(x)]$. Thus, the outcome from one function is automatically passed as input to another function without the need of repeated explicit intermediate object assignments or deeply nested function calls.

1.16.3 PERFORMANCE

In R, objects, including data sets, are loaded into the RAM of the computer. This implies a relatively hard threshold for working with big data, although some redirection to disk storage is possible. R data sets themselves, might be considerably smaller than the RAM but re-sampling techniques such as bootstrapping, or different data management steps might require substantial intermediate data duplication. Other computer applications running simultaneously might also compete for the same memory and some limitations might be imposed by the computer's operating system (OS) with 32 or 64 bits memory storage.[151]

A long-standing feature of R was to make deep copies of all modified objects prior to any modification (copy-on-modification), thus requiring at least twice the available memory. This procedure has been relaxed since R version 3.1 released in 2014. Multiple pointers (object names) can now refer to the same underlying object (pointing to the same physical memory address), allowing shallow copies where only the pointers are modified. Objects that are no longer needed (from intermediate computational steps or with all user accessible pointers explicitly removed), are automatically deleted by a garbage collector to increase the available memory.[146] An alternative approach is *reference semantics* where individual data columns are changed in place, while leaving unchanged columns as is. Reference semantic is common in other object-oriented programming languages, but not in R. It was implemented through C in the `data.table` package by Matt Dowle in 2006 with its own DSL.[152] A third alternative is to use memory mapping where data on disk is mapped to the RAM and treated as such.[153] A fourth option is to use R with an open database connectivity (ODBC) driver to perform some calculations remotely in a database, before loading the pre-modified data into memory.

R was not built to maximize computational performance but rather to provide an intuitive statistical user-interface combined with a dynamic, lazy (avoiding premature and unnecessary evaluation), functional, object-oriented language.[154] The language is quite unique in its flexibility including computation "on the language" and

non-standard evaluation (NSE). R was initially single-threaded but now includes several implementations for multithreaded and distributed computations. Iterative procedures (for-loops) were long-time bottlenecks in R, partially because objects assignments are made with dynamic type declarations, implying repeated methods dispatch for every iteration. Instead, vectorized operations (internally implemented by for-loops in C) were recommended and is still the norm, although the performance of for-loops has significantly improved in later versions of R. High-performance R-packages often relies on more efficient compiled Fortran/C/C++ code under the hood. This has always been possible, although cumbersome. This process got simplified by the introduction of the `Rcpp` package by Dirk Eddelbuettel in 2008.[155]

1.16.4 CLASSIFICATION OF COMORBIDITY

Dr. Mary Charlson has been involved in the development of a web-calculator that "[p]redicts 10-year survival in patients with multiple conditions".[156] Elixhauser et al. developed their own "Elixhauser Comorbidity Software" (SAS macros).[47]

Several R-packages have also been developed to estimate both the Charlson and the Elixhauser comorbidity indices based on medical coding. At least three of those packages are available through CRAN: `icd` by Jack O. Wasey et al,[157] `comorbidity` by Alessandro Gasparini et al,[158] and `medicalrisk` by Patrick McCormick and Thomas Joseph.[159] At least two other packages are freely and publicly available through GitHub: `comorbidities.icd10` by Max Gordon[160] and `icdcoder` by Wade Cooper.[161]

`medicalrisk` can be used with ICD-9-CM codes but is not up to date with the latest version of ICD-10. `comorbidities.icd10` and `icdcoder` are no longer actively developed or maintained. Both `comorbidities.icd10` and `icd` were previously very slow to use due to inefficient data structures and methods. `icd` has drastically improved since the initiation of the thesis. Current versions rely on efficient matrix algebra. `comorbidity` is a relatively new package which was not available at the beginning of the thesis project. As by now, both `icd` and `comorbidity` are good packages for calculating predefined Charlson- and Elixhauser comorbidity indices. All packages, however, rely on internal and static implementations of certain coding versions and index weights. No package implements algorithms for RxRisk V, or can be used with diagnosis-specific coding schemes for AEs et cetera.

2 AIM

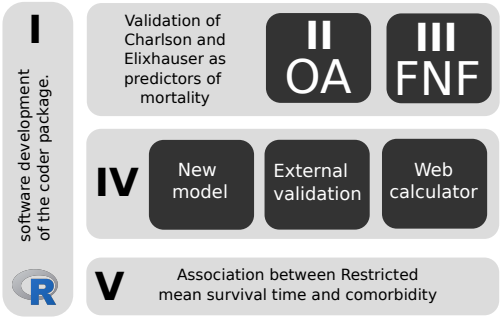

Figure 2.1: Schematic representation of the five studies. (OA = osteoarthritis. FNF = Femoral neck fracture.)

A long-term vision for SHAR has been to develop a tool for shared decision making for patients with hip arthroplasty. Hip arthroplasty should only be performed if the benefits outweighs the risks. To describe such risks, based on individual patient data, might help potential patients and their surgeons to make an informed decision whether to operate or not. The aim of this thesis was to provide some building blocks towards this goal.

Study I was a prerequisite to support Study II–V. Study II and III were twin studies to evaluate current recommendations for prediction modeling of mortality after hip arthroplasty based on comorbidity. An improved model was developed in Study IV and a better use of existing comorbidity indices was assessed in Study V.

2.1 STUDY I

We aimed to develop an efficient R-package for categorization of coded data based on generic code schemes and current best practice for R package development. The immediate rationale for such functionality was to categorize pre-operative patient comorbidities based on medical code data, but to also allow for easy extensions to post-surgery AEs et cetera. We aimed for an intuitive user interface with accessible documentation.

2.2 STUDY II

We aimed to investigate the long-term discriminatory abilities of the Elixhauser and Charlson comorbidity indices for patients with THA due to OA. We hypothesized that such indices would have limited predictive power, despite common recommendations to always include such measures in prediction models of mortality.

2.3 STUDY III

We aimed to investigate the same aspects and hypothesis from Study II for patients with hip arthroplasty due to FNF.

2.4 STUDY IV

We aimed to develop and validate a new prediction model for 90-day mortality after cemented THA due to OA. We hypothesized that such model would include some comorbidity, but not necessarily in the form of a pre-specified general-purpose comorbidity index. A secondary aim was to provide a web calculator to aid clinical usage and shared decision making prior to a decision whether to operate or not.

2.5 STUDY V

We aimed to visualize the associative relation between the Elixhauser comorbidity index and post-operative mortality after THA due to OA. We hypothesized that patients, on average, would have longer restricted mean survival time (RMST) if they had fewer pre-operative comorbidities.

3 PATIENTS AND METHODS

Study I concerned software development without patient data. Study II–V relied on observational prospective register data from Swedish and British health care, linked by PINs. Study II–IV concerned prediction modeling, Study II–III with the same validation techniques, and Study IV with essentially three parts: (1) model derivation and internal validation; (2) external validation and; (3) building a web calculator. Study V was based on inferential statistics and associative measures.

3.1 STUDY I

We used `data.table` as back-end for the `coder` package and decided to have no other required dependencies, making the package update-cycle less vulnerable to external dependencies and developers. Although `data.table` was used internally, it was also decided to adopt the design philosophy of `tidyverse` for the aplication programming interface (API).

We used S3 to implement a new class called `classcodes`, holding classification schemes to categorize individual codes into broader conditions (such as comorbidity). We provided a function `as.classcodes()` for the user to implement his or her own classification schemes.

We included default comorbidity `classcodes` for Charlson, Elixhauser, Rx Risk V and CPS. Additional `classcodes` for AEs after hip or knee arthroplasty were also provided. Each `classcodes` object contained several named conditions, each of them associated with possibly multiple versions of regular expressions identifying the relevant codes. For example, the first Charlson[32] condition "Myocardial infarction", later interpreted as "Acute myocardial infarction and old myocardial infarction"[36] was codified by ICD-9-CM as "410.x, 412.x". We interpreted this as all codes starting with 410 and 412 (with "x" representing a wildcard acting as a placeholder for any additional characters). This was formulated as a regular expression (ignoring the intermediate dot) as "`^41[02]`" where "`^`" denotes the beginning of a character string, "`41`" a literal continuation and "`[02]`" is either 0 or 2. Corresponding codes for ICD-10 are I21.x, I22.x and I25.2,[43] formalized as "`^I2([12]|52)`" where "`|`" act as a logical "or" within the parentheses. This is a simple example, although some conditions might require more intricate coding, for example "Ischemic heart disease: hypertension" according to RxRisk V:[56]

```
C(0(7A(A(0[1-689]|[1-9][0-9])|
B0[0-3]|G01)|8(C[A-Z][0-9]2|
DB0[01])|9(BB(0[2-9]|10)|
D(B0[1-4]|X0[13])))|10BX03)
```

A `summary()` method* based on the `decoder` package (a related CRAN-package by the author)[162] provides a more pedagogical over-view of relevant codes. A function `visualize()` was provided for a graphical representation with similar aim. The `classcodes` objects also contain different weighting schemes to calculate combined index values. An additional hierarchical attribute structure was imposed for the Elixhauser `classcodes` where "solid tumors" are subordinate to "metastatic cancer". A patient with both conditions will still be classified as such but a possible (weighted) index value will only account for metastatic cancer. The same is true for "diabetes uncomplicated" as subordinate of "diabetes complicated".[19] Another optional attribute concerns "conditions", as used for example by the default AE `classcodes`, where an AE might be conditioned on the main diagnose or the index hospitalization only.

The `classcodes` object is one of three required objects used for categorization by the `coder` package. The other two must be supplied by the user: (1) unit (patient) data with unique element identification (PIN or simi-

*A "method" acts like a sub-function called by a "generic" function in R. Hence, if the user calls a function `foo` with argument `baz` as `foo(baz)`, the call will be dispatched to method `foo.bar(baz)` if `baz` is of S3-class `bar`.

lar) and a possible reference date for an index event (hip arthroplasty) and; (2) an additional code data set with corresponding identification keys, as well as relevant (diagnoses) codes and optional dates (of recorded comorbidity). For Study II–V, (1) was taken from SHAR and (2) from the NPR with ICD-10-codes and dates of corresponding hospital visits and admissions.

There are three important steps to categorize the data: (1) codification of units (patients) based on the additional coding data, as implemented by the `codify()` function; (2) classification of the resulting data using the `classcodes` object, as implemented by the `classify()` function and; (3) to aggregate such outcome by the proposed index, as implemented by the `index()` function. Those steps could either be performed sequentially, chained as `codify() %>% classify() %>% index()` or directly by the main function of the package, `categorize()`, combining the individual functions under the hood.

3.1.1 DEVELOPMENT

Different versions of R and RStudio were used during the process together with dedicated support packages for package development (`devtools`, `usethis`, `testthat`, `roxygen2`, `pkgdown` et cetera). The `profvis` package was used for interactive visualization for code-profiling and optimization. Git was used for version control and the package was developed as open source software publicly available through GitHub.* A website with documentation was published with vignettes and a reference manual.† A suite of unit tests was developed to ensure functionality and stable development. Continuous integration tests were deployed to ensure compatibility with Windows, Ubuntu, Red Hat Linux, Mac OS and Solaris. Recommendations and best practice from the rOpenSci project were followed when possible.[163]

*`https://github.com/eribul/coder` (accessed 2020-08-02)

†`https://eribul.github.io/coder/` (accessed 2020-08-02)

3.2 ETHICS AND LEGAL ASPECTS

None of the studies involved any patients directly, nor any medical journals or biological samples. Study II–V, however, used extensive data sets with personal data of sensitive nature. It was thus important to respect and secure the integrity of all patients indirectly involved. Inclusion in the NPR and the Longitudinal integrated database for health insurance and labour market studies (LISA) is mandated by national law (6 §. Swedish code of statues (SFS) 2001:707 for NPR and various sources for different parts of LISA, combined according to appendix 3 of the official report of the Swedish government (SOU) 2003:13). Patients have no legal obligation to participate in quality registers such as SHAR, however. They can choose to opt-out, or to withdraw earlier implicit consent at any time without providing any explanation (7 chap. 2 §. SFS 2008:355).

Retrieval of Swedish registry data is subject to the principle of public access to official records, first introduced in 1677.[15] The current incarnation, the press act from 1949 (TF 2:1), grant every citizen (and others for most parts), the right to access any public document. Personal records concerning health status are, however, also subject to the personal protection data act (OSL 21:1), which requires explicit approval from ethical authorities prior to any data sharing motivated by research (OSL 21:7 and 3 §. EPL). Such approval was granted by regional review boards prior to 2019, and by the Swedish Ethical Review Authority since then (24 §. EPL). Ethical approval for the Swedish data was granted by the Regional Ethical Review Board in Gothenburg (reference number 271-14). This was a necessary but not sufficient requirement. Each legal entity responsibility for a specific register must also approve data sharing based on their own risk assessment considering patient data protection and secrecy. Such assessments were thus performed by the NBHW for the NPR and the medical prescription register, and by SCB for LISA. VGR is the legal entity responsible for SHAR.

Data retrieval was performed prior to the implementation of the European General Data Protection Regulation (GDPR), but data was stored and treated according to current laws and regulations. We made sure that no patients could be identified on an individual level and that patients and other interested parties were kept informed about the research (through SHAR's online project database at www.shpr.se).

Study IV included additional data from England and Wales for external model validation. This part of the study was performed by the co-authors in England. Hence, this data was treated separately and never left the UK. Permission was granted by the NJR (with reference number RSC2017/21). Informed patient consent was not mandatory according to the UK law for pseudonymized data.

Quality register data (as formally unverified by the treating physician) cannot be used for treatment decisions directly.[164] We therefore provided a stand-alone web calculator in Study IV, where patients and physicians can choose to re-enter some data for a predicted probability of 90-day mortality. This tool is technically independent/separated from the quality register and Stratum. The model is static and is not automatically updated as new patients appear in the registers.

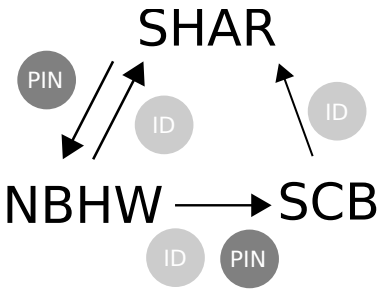

Figure 3.1: Schematic representation of data linkage. PIN = Personal Identity Number. ID = Anonymized patient ID. SHAR = The Swedish Hip Arthroplasty Register. NBHW = The National Board of Health and Welfare (Socialstyrelsen). SCB = Statistics Sweden (Statistiska Centralbyrån).

3.3 PATIENT DATA

We used two versions of a large linkage data base involving SHAR, NPR, LISA, the medical prescription register and additional registries used for related research, although not explicitly for this thesis. The first version included patients with THA 1992–2012 and hemiarthroplasty 2005–2012. This data set was already available before the implementation of Study II.[165] A new data base was built prior to Study III–V. The process was similar as described by Cnudde et al.[165] but for an extended time frame: (1) all patients from SHAR, operated with primary THA 1992–February 2018 and hemiarthroplasty 2005–February 2018, were identified by their PINs and laterality (left or right hip); (2) this data set was submitted to the NBHW who identified comorbidity data from the NPR for the patients operated 1999–2015; (3) NBHW returned the matched data to SHAR with PIN replaced by an anonymous id; (4) NBHW also sent PINs with corresponding id:s to SCB and; (5) SCB linked those PINs to LISA, removed the PINs and returned the data to SHAR (Figure 3.1). Hence the updated linkage data base included patients operated 1992–February 2018 but only with comorbidity for patients operated 1999–2015. We built a data base using the Structured Query Language (SQL) and the SQLite software[166] indexed by the anonymized id, laterality and date of surgery, re-operation and hospital visits. We stored the data on a secure server maintained by the University of Gothenburg, accessed only through a secure network with trusted computers at RC in Gothenburg.

Pre-operative comorbidity were estimated for both the first and second version of the linkage data base. We used look-back periods of 1, 2 and 5 years for each operated hip, comparing dates of out-patient hospital visits and in-hospital discharge, to dates of primary surgery. Hence, only hospital visits with out-patient visits or in-patient discharges, prior to surgery, were included to avoid that AEs were falsely classified as pre-existing comorbidity. Only the one-year look-back period was later used, however, since longer periods provided no improve-

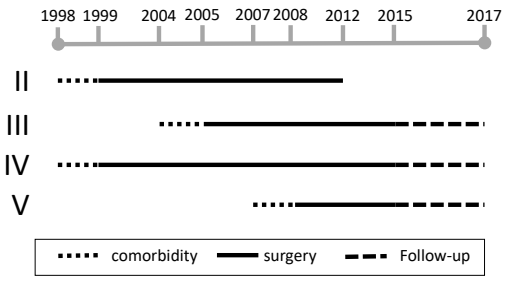

Figure 3.2: Study periods for Study II-V.

ment considering discriminatory ability (supplementary Figure 4 in Study II). The `coder` package (Study I) was used to identify individual comorbidities according to Charlson and Elixhauser based on ICD-10.[43] The number of identified comorbidities according to Elixhauser was summed as an unweighted index score and two different weighing schemes were used for Charlson. In Study II–III, we refer to those as the "original", as proposed by Charlson et al. in 1987[32] and the "updated", as proposed by Quan et al. in 2011.[44] Patients with no recorded hospital visits within the look-back period were assumed to have no comorbidity. Rx Risk V was calculated in a similar way with ATC-codes from the medical prescription register. Inclusion and exclusion criteria differed for each study (Table 3.1). Hemiarthroplasty has been reported to SHAR since 2005, wherefore patients with FNFs (Study III) were studied from this year only (Figure 3.2). BMI and ASA have been recorded with sufficient completeness in SHAR since 2008, which was therefore the first year included in Study V.

Survival times were calculated for each patient from the day of primary surgery, to the day of either death or censoring, whatever came first. Administrative censoring was applied to all patients still alive at 2012-12-31 (Study II) and 2017-12-31* (Study III–V).

*This date could as well have been set to February 2018.

3.4 STUDY II

We used the first version of the linkage data base,[165] and included the first THA for each patient operated 1999–2012 due to OA. We excluded patients who died on the day of surgery. We used univariable Cox-regression to regress the Elixhauser as well as the "original" and "updated" Charlson comorbidity indices to mortality.

Schoenfeld residuals were used to evaluate if the proportional hazards assumption was met. It was not in its entirely, but we did have proportional hazards within shorter periods starting by year 0, 5 and 8. Those time points were used as cut points for stratification used by the extended Cox model with time dependent covariates. The same cut points were used for all models except for a base model with age and sex. We fitted 103 Cox regression models to the data: (1) 48 univariable models, one for each comorbidity condition identified by either Charlson or Elixhauser; (2) the same 48 models adjusted for age and sex; (3) 3 univariable models for each comorbidity index; (4) the same 3 models adjusted for age and sex and; (5) a base model with age and sex only.

Estimated coefficients from each model were used to predict survival for each patient of the sample. This was done repeatedly for each time point when at least one death had occurred. Observed and predicted values were compared to calculate sensitivity and specificity for each model at each time point. A ROC curve was calculated for each of those, and the AUC was estimated by numerical integration. Numerical integration was applied once more over a second dimension, time. We used 100 bootstrap replicates to estimate 95 % CIs for each AUC.

3.5 STUDY III

Study III was methodologically similar to Study II, applied to patients with FNF from the updated linkage data base. Those patients were treated with either THA or hemiarthroplasty 2005–2015. The estimated cut-points to receive proportional hazards within sub-intervals were different from Study II:

Table 3.1: Patients and outcomes in Study II–V. OA = osteoarthritis. FNF = Femural neck fracture. n = sample size. ECI = Elixhauser comorbidity index. CCI = Charlson comorbidity index. AUC = Area Under the (Reciever Operating Characteristics) Curve. THA = Total Hip Arthroplasty. RMST = Restricted mean survival time

Study	Diag.	Years	Prosthesis	n	Comorbidity	Mortality	Outcome
II	OA	1999–2012	THA	120,836	ECI /CCI	up to 14 years	AUC
III	FNF	2005–2015	THA + hemi	43,224	ECI/CCI	up to 7 years	AUC
IV	OA	2008–2015	cemented THA	53,099 + 125,428*	new model	90 days	AUC
V	OA	1999–2015	THA	150,367	ECI	up to 10 years	RMST

* Two cohorts were included in Study IV with patients from SHAR and NJR.

160 days, as well as 1, 2 and 5 years after surgery. We also applied the van Walraven weights of the Elixhauser comorbidity index for a sensitivity analysis.[48]

3.6 STUDY IV

There are two versions of Study IV. The first was included in a thesis by Anne Garland.[167] Data from the first linkage data base were used with explorative multivariable logistic regression and comorbidities acknowledged by Charlson, Elixhauser and Rx Risk V. For the second version, in this thesis, we used data from the second linkage data base. We used different statistical methods and performed an external validation with registry data from England and Wales. Rx Risk V was no longer used, however, since the required data were not available for the external validation cohort from England and Wales.

3.6.1 LITERATURE REVIEW

We performed an initial ad hoc literature review based on Google Scholar,[168] personal recommendations and relevant references, to identify models similar to the one we aimed to develop. We then used 2Dsearch[169] to build a search query for a more systematic review.*

We used no time or language constraints and used this query in PubMed,[170] where we identified 143 articles, whereof one was a duplicate. We performed an article screening with Rayyan.[171] 131 articles were excluded based on titles, and an additional 6 after reading the abstracts. 5 articles remained, whereof 3 were already considered during the ad hoc review process. Research on cadavers, canines, and patients with diagnoses other than elective hip disorders (including hip fractures) were excluded during the screening process, as were prediction models with outcomes other than mortality. In addition, we used Open Grey with the same query, as well as with a shorter simpler query (`hip AND prediction`).[172] We found 9 items, whereof none was considered relevant. The same queries were used with Epistemonikos.[173] No items were found using the longer query, but 33 systematic reviews based on the shorter. 30 were excluded after title screening, and the remaining 3 after reading the abstracts.

*`(femur OR hip OR bone OR bones OR osteoarthritis OR rheumatoid OR rheumatism) AND (prognostic OR prediction OR forecast OR forecasting OR forecasts OR predict OR predicted OR predicting OR predictions OR predicts OR projections) AND (death OR mortality OR deaths OR survival OR survive) AND (replacement OR replaced OR replacements OR arthroplasty OR "Arthroplasty, Replacement" OR prosthesis OR implant OR prostheses OR prosthetic OR prosthetics) AND (calibration OR calibrate OR calibrated OR validation OR validate OR validated OR validating OR validity OR precision OR accuracy OR accurate OR precise OR "external validation" OR evaluation OR verification OR internal) AND (build OR building OR construct OR create OR develop OR establish OR train OR derive OR derivation) AND (models OR modelling OR "statistical model" OR modeling) AND (comorbidity OR comorbid OR co-morbid OR co-occurring OR diagnoses OR comorbidity OR Elixhauser OR Charlson OR "comorbidity index" OR "comorbidity score" OR "Rx Risk V" OR ASA)`

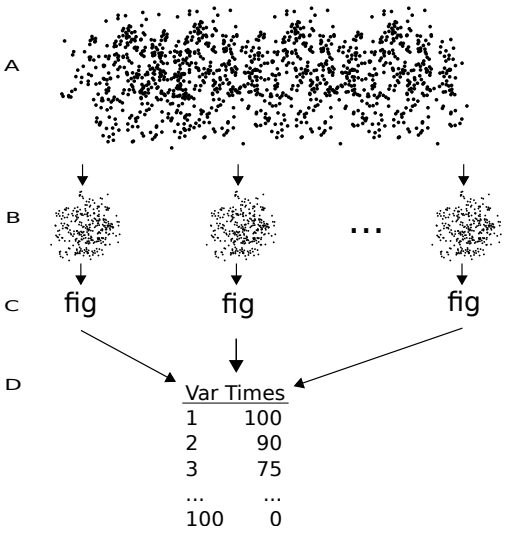

Figure 3.3: Outer steps of the bootstrap ranking procedure. All included patients (A) were re-sampled with replacements 100 times (B). For each bootstrap sample, the inner process (C; as depicted in separate figure) was applied. The final output was a table with all potential predictors ranked by the number of times they got selected (D). Potential predictors selected at least once were included in the main model, and each predictor selected at least 33 out of the 100 times, were included in a reduced model as well.

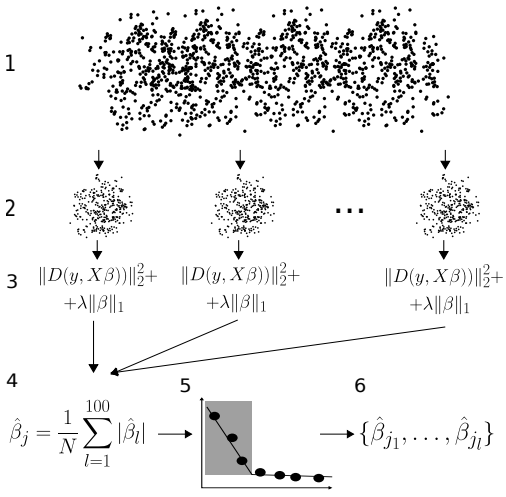

Figure 3.4: Inner steps of the bootstrap ranking procedure. For each bootstrap sample (1), 100 new bootstrap samples were created (2). Logistic LASSO-regression was applied within each sample (3). Absolute values of the estimated coefficient values (whereof some shrunk to 0 by the LASSO), were averaged as a measure of variable importance (4). Variables with their estimated variable importance above an estimated breakpoint from linear piece-wise regression (5) were kept as potential predictors (6).

In summary, we found that several prediction models had been developed for short-term mortality in patients operated with THA. Most of them were based on small sample sizes, included predictors that were unavailable in a clinical setting, only considered in-hospital deaths, were inadequately validated, or had no national population coverage leading to biased samples. Only few model descriptions included nomograms, formulas or accompanying tools to aid clinical usage. The relevant models were further described and referenced in the study manuscript.

3.6.2 STATISTICAL METHODS

The study underwent several iterations with different methods for variable selection. We tried bootstrap aggregating (bagging) with step-wise logistic regression, Bayesian modeling averaging[174] and Bolasso, a combination of bootstrapping and the LASSO.[117] The final bootstrap ranking procedure was similar to Bolasso,[114] but with additional steps of variable ranking and piece-wise linear regression.[115] Further details are provided in the manuscript and visualized in Figure 3.3 and 3.4. R-scripts for the exact implementation are provided through an on-line repository.*

3.7 STUDY V

We studied association between preoperative comorbidity and the RMST. Details are provided in the manuscript and by the R-scripts deposited online.†

A preliminary version of the paper considered both the Charlson and Elixhauser comorbidity indices. The results were similar and Elixhauser was previously shown beneficial for patients with OA (Study II). There-

*https://doi.org/10.5281/zenodo.3732852 (accessed 2020-06-02)

†https://doi.org/10.5281/zenodo.3458031 (accessed 2020-06-24)

fore, only the Elixhauser index was kept in the final version of the manuscript.

The use of RMST was introduced already in 1949.[87] A point estimate is easily estimated as the area under a Kaplan–Meier curve by most statistical software. Combining RMST with regression is less common, however, although an increased interest has been observed during recent years.[91] The `survRM2` R-package only includes functionality for covariate adjustments based on analysis of covariance (ANCOVA).[175] There is an R-package `psuedo` for regression based on pseudo-values for censored data, which we were not able to use due to high computational burden and assumed suboptimal internal procedures.[176] The `prodlim` package,[177] however, offered a fast implementation, which we combined with a Jackknife procedure to estimate pseudo-values. There are several R-packages suited for GEEs. We used `geepack`[178] with a working variance of 1.

The RMST is a time-specific value evaluated for some τ as the number of days since THA due to OA. We repeatedly estimated μ_τ for $\tau = 1, \ldots, 3650$, thus for every day during approximately ten years after surgery, to estimate a RMST curve with a time scale on the y-axis instead of the traditional proportion presented for standard survival curves.[179]

4 RESULTS

4.1 STUDY I

The `coder` package has been released as open source software under the MIT license, granting permission for private and commercial use, modifications and distribution (Figure 4.1). There are 8 default code schemes

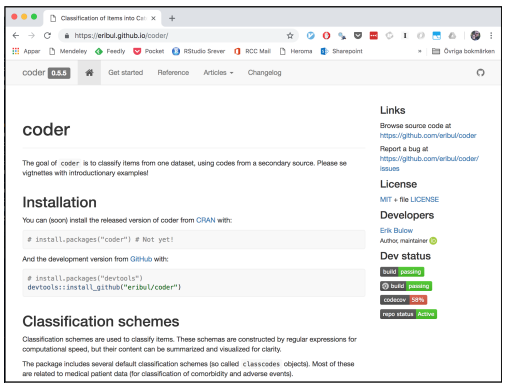

Figure 4.1: Screenshot of package documentation website (eribul.github.io/coder)

(`classcodes` objects) included in the package (Table 4.1). They comprise a total of 117 individual conditions, each formulated by up to eight different versions of regular expressions (Figure 4.2).

Using the package reduced computation time from approximately 18 hours using iterative and non-optimized procedures in base R, to around 30 seconds, for data management in Study II. Initial benchmarking showed that the package was around 400–600 times faster than comparable packages. This is no longer true, however, for newer versions of the `icd` and `comorbidity` R-packages, with significantly improved performance. The `coder` package is unique in terms of flexibility and regarding the variety of default classification schemes.

4.2 STUDY II

We found that neither the Charlson, nor the Elixhauser comorbidity indices were sufficient to accurately predict mortality after THA due to OA. The Elixhauser index was

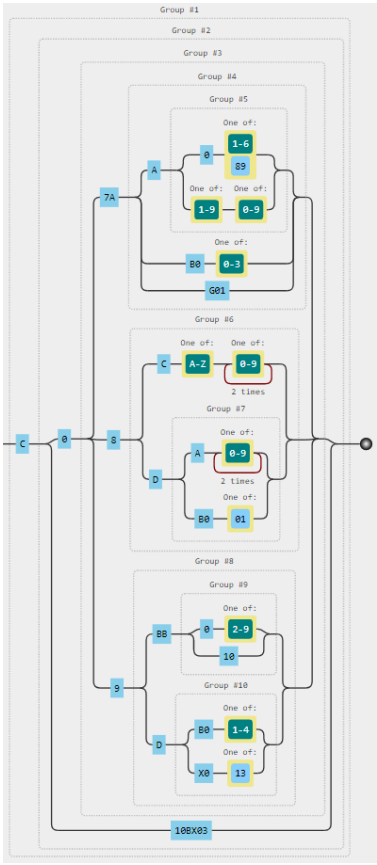

Figure 4.2: Example output from `visualize()` with regular expression for Ischemic heart disease with hypertension based on ATC codes in the RxRisk V.[56]

the best candidate, however, with AUC 0.61, 0.60, 0.59, 0.58 and 0.56 for deaths within 90 days, as well as 1, 5, 8 and 14 years. The simple baseline model with age and sex performed better (AUC around 0.74 regardless of period), and further improvements were seen for a multivariable model with age, sex and the Elixhauser comorbidity index combined (AUC close to 0.76 for deaths within 5 years).

4.3 STUDY III

The Charlson comorbidity index was superior to Elixhauser for patients with FNF; the

Table 4.1: Default classcode schemes with corresponding code matching and different versions of regular expressions (regex) and weighted index sums (indices). CPS = Comorbidity-poly pharmacy score. AE = adverse event.

classcodes	regex	indices
charlson	icd10, icd9cm_deyo, icd9cm_enhanced, icd10_rcs, icd8_brusselaers, icd9_brusselaers	index_charlson, index_deyo_ramano, index_dhoore, index_ghali, index_quan_original, index_quan_updated
cps	icd10	index_only_ordinary
elixhauser	icd10, icd10_short, icd9cm, icd9cm_ahrqweb, icd9cm_enhanced	index_sum_all, index_sum_all_ahrq, index_walraven, index_sid29, index_sid30, index_ahrq_mort, index_ahrq_readm
ex_carbrands		
hip_ae	icd10, kva, icd10_fracture	
hip_ae_hailer	icd10, kva	
knee_ae	icd10, kva	
rxriskv	pratt, caughey, garland	index_pratt, index_sum_all

estimated AUCs were 0.60, 0.59 and 0.57 for deaths within 90 days, as well as 1 and 5 years. The "updated"[44] Charlson index performed slightly better than the "original"[32] version, but differences were small and clinically irrelevant. The base model with age and sex was better than all of the comorbidity indices, but not as good as for patients with OA (AUC around 0.65 for all periods). Not even the multivariable model combining age, sex and the Charlson comorbidity index reached the desired AUC of 0.7 (0.69 up to 2 years after surgery).

4.4 STUDY IV

An improved prediction model for 90-day mortality for patients with cemented THA due to OA combined age, sex, ASA class and the presence of cancer, central nervous system (CNS) disease, kidney disease and obesity. The model had good discriminatory ability, both internally and externally, with AUC 0.78 and 0.75 for patients in the SHAR and the NJR respectively. The model was also well calibrated for predicted probabilities up to 5 %. The web calculator is available in both English and Swedish (Figure 4.3).

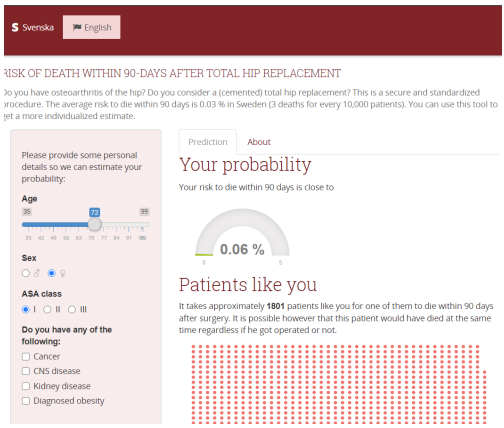

Figure 4.3: Screenshot of web calculator for the developed prediction model from Study IV. (https://erikbulow.shinyapps.io/thamortpred/)

4.5 STUDY V

Patients without comorbidity had a RMTL of only about 3 hours within the first 90 days, compared to 26 hours for patients with at least four comorbidities. RMTL at ten years increased to 0.93 and 3.3 years respectively, a difference with clinical relevance. Regression modeling with and without adjustment for age and sex led to similar results. The effect of age and sex increased by longer follow-up. Being male had almost twice the effect on RMTL as the difference between zero and one comorbidity.

5 DISCUSSION

5.1 STUDY I

We designed and built a unifying framework for code categorization based on regular expressions. This method was more computationally efficient compared to traditional procedures comparing each individual code by a look-up-table. We used the default regex engine in R, which in turn relies on Perl-compatible regular expressions (PCRE) implemented in C. The R adaptation was insufficiently implemented with a quadratic time algorithm, $O(n^2)$, up to R version 3.5.2. A faster implementation with a linear time algorithm, $O(n)$ is provided by Google as the `RE2` package written in C++ and wrapped by the `re2r` package for R. This engine might still be more efficient than PCRE, but the difference is less relevant from R version 3.6.[180]

5.2 STUDY II–III

The results were as hypothesized; that prespecified comorbidity indices had limited predictive power, in spite of common recommendations to always include such measures in prediction of mortality. Our results confirmed a study by Armitage et al.[45] and were coherent with some previous concerns regarding the usefulness of existing comorbidity indices.[181]

Observed AUC values for patients with FNF were lower compared to patients with OA. This indicated that mortality for those patients were harder to predict based on preexisting comorbidity, age and sex. This indicates that FNF is in itself a severe condition, altering the remaining life trajectory.

The AUCs were higher for Elixhauser compared to Charlson for patients with OA (Study II), but the opposite was observed for patients with FNF (Study III). This might be partially explained by the methods used to develop the Elixhauser index. Medical conditions that were found among patients in the original cohort, but without association to length of stay, hospital charges or in-hospital deaths, were not considered relevant and therefore excluded. OA was one of those conditions. This was in our favor for Study II, since we could then use the classification without modification (the same condition should not be classified as both the index disease and comorbidity). Dementia was also excluded for the same reason. This is a disadvantage for Study III, however, since dementia is an important comorbidity for patients with FNFs. Patients discharged to other institutions after the hospitalization were also excluded from the cohort used to develop the index. Those patients likely resembles those with FNF, who might therefore have been less represented in the data set used by Elixhauser et al.[19]

The follow-up period was stratified with cut-points at 5 and 8 years for Study II and at 160 days as well as 1, 2 and 5 years for Study III, to model the period specific hazards over time. It might be questioned, however, whether predictions of mortality would be clinically relevant for such long-time horizons after surgery. Post-operative life trajectories (introduction of additional comorbidities and other events) starts to influence the remaining survival more and more, as time pass by.

5.2.1 METHODOLOGICAL CONSIDERATIONS

The Elixhauser, and especially the Charlson index, exist in many versions.[46] We were only able to use adaptations based on ICD-10 and we choose to compare two versions of the Charlson index, as well as the unweighted Elixhauser score. Considering those versions, Elixhauser performed better in Study II and Charlson performed better in Study III. It seems reasonable to conclude, however, that the versions we used should be representative for their respective comorbidity index. The index weights also exist in many versions. This might seem less motivated from a theoretical perspective, since all weights are necessarily cohort-specific and might not be applied to patients who are different from the training data used for each in-

dividual model. Pre-defined weights might nevertheless be useful in a clinical setting for patients who are similar to the training cohort, since new weights could not be estimated based on single patients alone. In a research setting such as ours, we do have access to large amounts of data, wherefore most weights could be re-calibrated. This is a reasonable alternative to consider before developing a new predictive model from scratch. Thus to update, revise or re-calibrate an existing model, before abandon it.[108]

We used a one-year look-back period for comorbidity recorded in the NPR. This is contrary to some suggestions to use all available data. We could have used a longer look-back period based on the ICD-10 for a marginal gain in predictive performance. Theoretically, we could also include additional data based on ICD-8 and ICD-9 to back-trace comorbidities since 1968.[46] We did not have access to such data, however, but we did perform a sensitivity analysis using some alternative look-back periods (supplementary material for Study II).

5.2.2 STATISTICAL CONSIDERATIONS

A summarized comorbidity index might be sufficient if the weights are estimated correctly. Hence, no individual comorbidity might be needed as long as a weighted sum of all relevant conditions is provided.[101] Thus, knowledge of $Z = X\beta$ is enough to predict Y if $Y = f(X\beta) = f(Z)$. Unfortunately, the Charlson index score for patient i is not $z_i = \sum_j \hat{\beta}_j x_{ij}$ but $z_i^{(c)} = \sum_j r(\exp(\hat{\beta}_j x_{ij}))$, where $r(w) = 0$ for $w \leq 1.2$, otherwise a function rounding real numbers to integers. Hence, the estimated $\hat{\beta}$-coefficients are used to estimate HRs.* Those HRs are then ignored (set to zero) for point estimates smaller than 1.2. Larger HRs are rounded to their nearerst integer and summed. Thus, HRs are summed and not multiplied, since Charlson et al. assummed $\exp(\sum_l a_l) \approx \sum_l \exp(a_l)$, instead of $\exp(\sum_l a_l) = \prod_l \exp(a_l)$. We can illustrate the differences for a scoring system with three conditions (presence of three comorbidities). First, assume that $\hat{\beta} = (-0.15, 0, 0.15)'$. A patient with all conditions will thus have an index value of $z_i = \sum_j \hat{\beta}_j x_{ij} = -0.15 \cdot 1 + 0 \cdot 1 + 0.15 \cdot 1 = 0$ (assuming a zero intercept, $\beta_0 = 0$). Hence the total effect would cancel out. Similarly, since $\exp(.15) \approx 1.16 < 1.2 \Rightarrow z_i^{(c)} = 0$. But if $\hat{\beta} = (-0.6, 0, 0.6)'$ we still have $z_i = 0$ but $z_i^{(c)} = \sum_j r(\exp(\hat{\beta}_j x_{ij})) = r(\exp(-0.6)) + r(\exp(0)) + r(\exp(0.6)) = r(.55) + r(1) + r(1.8) = 0 + 0 + 2 = 2$. Hence, the protective effect would be ignored, and the index sum would indicate increased mortality. Admittedly, the original Charlson score did not include any conditions with $\hat{\beta}_j \leq 0$, but other versions and comorbidity indices do.[48,56] The mistake has been noted several times,[40,182] and it was also acknowledged by the editors of the Journal of Clinical Epidemiology (the successor of the Journal of Chronically Diseases, which published the original paper).[184] Charlson et al, however, have not confirmed or corrected their mistake but have instead stated that "[t]he simplicity of calculating the Charlson comorbidity index and its interpretability has likely propelled its widespread use".[185]

The AUC was used as a concordance index as the sole entity for model validation. It was estimated using sound methods for censored survival data, but it has been criticized for its inability to sometimes detect meaningful differences in discriminative ability when adding new predictors to a model.[186] This is especially true for rare events data where large ORs are needed before any visible change of AUC might be detected. It is therefore possible that the added value of comorbidity to the base model with age and sex, is underestimated. The Net reclassification improvement (NRI) is an alternative measure, which is more sensitive. It lacks a natural interpretation, however, and the value might lead to confusion if not carefully scrutinized.[110] Another alternative assessment tool is the coefficient of determination, $R^2 \in [0, 1]$, as often used for internal goodness-of-fit validation in inference stud-

*Assumed to approximate relative risks by Charlson et al.[32]

ies. It is an estimate of correlation, the explained variability, between observed and expected values in linear regression. $R^2 = 0$ means no correlation, hence no predictive ability. $R^2 = 1$ implies a deterministic relationship, which might seem suspicious in most cases. Values between 0.2 and 0.3 are common in medical prognostic settings (predictions of the future), although higher values might be expected for diagnostics (predictions of a current feature).[110] There are several generalizations of pseudo-R^2 to generalized linear models (Nagelkerke's, McFadden, Cox-Snell and others). Unfortunately, there is no consensus of which version to use. The Brier Score is another measure, combining discrimination and calibration. Its scale varies with the base incidence rate, however, wherefore comparisons between different models and cohorts are difficult. A rescaled version from 0 to 1, the "index of prediction accuracy", has been proposed.[187] The Brier Score is popular but quite insensitive in settings with very rare outcomes (such as ours). The Lorenz-curve and the Gini index (coefficient) are two additional methods used in economy, but rarely in medicine.[110]

The lack of discriminative ability also indicates poor calibration and a lack of clinical usefulnes, wherefore such measures were not explicitly assessed.

5.2.3 STRENGTHS AND LIMITATIONS

Observational register studies have the potential to capture real world evidence (RWE) to a larger extent than randomized clinical trials (RCTs), although causal relationships are harder to find. The use of Swedish PINs and registers with high quality data and national coverage was beneficial. The data used, however, were collected for other primary purposes than research. This is one of the criticism for ICD codes in general. They were originally developed by, and for, the statistical community to allow for data aggregation and follow-up over time. It has become more of an administrative tool, however, with constant revisions, as well as false incentives for reporting some conditions more than others.[9,188]

Only the first operated hip was included for patients with bilateral hip arthroplasty. This is a common practice in the orthopedic literature. We have later showed, however, that the second hip better resembles unilateral hips, wherefore those were included in Study IV–V.[189]

We choose to differentiate discriminative ability for different age groups in Study III. Cut points for those age groups (70 and 90 years) were based on observed data. Such data-driven approaches are often criticized, and more clinically relevant age groups might very well be preferred.

A limitation applicable to all empirical studies (II-V) is the lack of multiplicity correction for p-values and CIs, where the unadjusted $\alpha = 0.05$ might be too liberal. This, however, is a common setting in the medical literature.

5.3 STUDY IV

We derived a model with some predictors of interest. A relevant question is which predictors were the most important? This has no clear answer in a prediction setting with variable selection combined with ensemble methods. Variable inclusion was based on predictive power but potential exclusion of one variable might lead to another variable taking its place.[98] If we nevertheless consider the proposed model, the magnitudes (absolute values) of the estimated ORs could act as ad hoc measures of variable importance. This reveals a large effect for ASA class III (patients with severe systematic disease), compared to patients with ASA class I (normal healthy patients) as baseline. Comparing two patients, one with ASA class I and one with ASA class III, it seems reasonable to assign a higher probability of death to the latter. This would correspond to good discriminative ability of the model (increasing the AUC value) if the patient with ASA class III dies, while the patient with ASA class I survives. The wide CI for the estimated OR for ASA class III, however, indicated large uncertainty and sub-optimal calibration. Male sex might seem as a less important predic-

tor, with only moderate OR compared to the other predictors. This factor level is common, however, relevant for almost 40 % of the patients, implying a relatively narrow CI. The effect of kidney disease was larger in magnitude but was only relevant for 1 % of the patients in the Swedish cohort. It was thus a useful predictor for this small portion of patients, but it contributes less to the overall relevance of the model. The age-effect might be transformed to a more meaningful time scale, such as age in decades instead of individual years. Alternatively, the normalization (to mean 0 and variance 1), used prior to variable selection might be retained. It is otherwise hard to compare variable importance for this continuous variable to the factor variables. The CI was very narrow, however, since age is a relevant predictor for all patients in the cohort. It should be noted that ORs are relative measures which must be interpreted in relation to the baseline cohort due to the non-collapsibility feature of logistic regression.

As discussed in the manuscript, obesity was included in the model although its estimated OR was not statistically significantly different from 1. A model without this predictor would yield worse performance, however. The non-linear relationship among predictors in logistic regression is non-intuitive, due to its multiplicative error term affecting the interpretation of coefficients in relation to non-included variables as well. Possible correlation (non-modeled interaction effects) might also affect the association between predictors and outcome. Obesity is a known risk factor for some diagnosis of cancer, as well as a causal effect increasing the ASA class. This is one reason why the magnitude of the ORs is often an ill-conceived measure of variable importance.[190–192] A better comparison of variable importance would consider the difference in AUC for models with and without each predictor. Unfortunately, the AUC, as a measure of rank correlation, is rather insensitive in this setting. The permutation test is a possible alternative,[192] where the variable importance of predictor $X_{\cdot j}$ is estimated by comparing model M based on the observed $x_{\cdot j} = (x_{1j}, \ldots, x_{nj})'$ and a scrambled/permutated perm($x_{\cdot j}$) where x_{ij} is no longer assigned to patient i but to another patient at random. Thus, any possible association between perm($x_{\cdot j}$) and y would be purely random. Other methods compare pseudo R^2-values for models with or without the predictor of interest,[190] or uses dominance analysis with similar methods extended to all nested sub models.[191] To formally assess variable importance for logistic regression is, however, unusual, although much more common for linear regression or for ML techniques such as random forest (a method combining multiple decision trees).

5.3.1 METHODOLOGICAL CONSIDERATIONS

We preferred a model with logistic regression due to ease of interpretation and a wish to present the result as transparent as possible. It is possible that other modeling techniques, such as generalized boosting models,[193] would improve the discriminative ability.[194] A systematic review, however, found no benefits of ML compared to logistic regression.[195] Tree-models have also performed worse than main effects regression models in some validation settings.[114] However, very few patients die within 90 days of THA due to OA. Thus, the outcome is rare, and the data unbalanced/imbalanced. This is problematic for logistic regression, which is often not recommended for modeled probabilities $p \notin [0.2, 0.8]$, for which $\hat{p}$ might be too small, thus under-estimating the true p.[196] We therefore tried, or at least considered, several methods to re-balance the data prior to further analysis. This included traditional under- and oversampling, random over-sampling examples (ROSE),[197] class-imbalanced subsampling LASSO logistic regression,[198] ensemble methods with boosting, bagging and hybrid-based approaches,[199,200] as well as stratified sampling for each bootstrap- and cross-validation step. Such methods are popular with classification and linear regression, where distorted intercepts (due to different sampling schemes) might be compensated

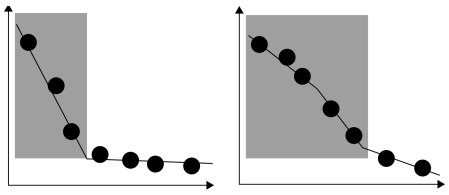

Figure 5.1: Schematic representation of variable selection based on estimated coefficient values larger than a single break-point (left), versus the second of possibly multiple break points (right).

for, by adjustments to the over-all incidence rate.[201] This is more difficult for, a possibly miss-specified, logistic regression model due to non-collapsibility. We therefore decided to use logistic regression as is.

We used traditional LASSO regression, but there are several versions which might be considered. Especially the group LASSO could be a relevant method in our case, to either include or exclude all levels of multilevel factors simultaneously. Adaptive LASSO, introducing individual penalty terms for each parameter, has showed promising results in settings with high dimensional data, but might be less relevant for lower dimensional data sets such as ours.[202] Firth regression has also shown some benefits over LASSO considering data with few events per variable (EPV).[203] LASSO however, is still considered the norm for variable selection using penalized regression in the medical field. We used piece-wise linear regression to find a breakpoint based on variable importance measured as the mean of the absolute values of the estimated coefficients. Only variables with coefficients above this break-point were further considered. This method was proposed by Guo et al.[115] and is fairly intuitive in a classical "Pareto setting" where a few vital variables contribute to the majority of the predictive power (Figure 5.1 left panel). Alternative settings could apply where an individual breakpoint is less obvious or where several consecutive breakpoints might apply (Figure 5.1 right panel). To include all variables (relying on the LASSO penalty term alone) could lead to theoretically better models, preferred outside clinical settings where the trade-of between model parsimony and accuracy is less important. Another alternative is to replace the current piecewise regression (splines with one knot and one degree polynomials) with multiple linear, or even non-linear, segments and breakpoints, possibly identified using Bayesian methods.[204,205] A further generalization might consider alternative ranking procedures in addition to the absolute values of the mean estimated coefficients.[116]

We choose to assess calibration graphically by a flexible calibration curve based on a parametric model with fractional polynomials.[136] Several alternatives have been suggested based on non-parametric smoothers/loess-functions, for example the integrated calibration index, described as the "weighted difference between observed and predicted probabilities, in which observations are weighted by the empirical density function of the predicted probabilities".[206] This method is not restricted to logistic regression and is therefore a good alternative if comparing models derived by different statistical techniques such as random forest or boosted regression.

5.3.2 STRENGTHS AND LIMITATIONS

Most predictive models suggested for clinical use are never validated;[207] only 25 % (32 of 127) according to a systematic review from 2015.[208] Our external validation with patient data from the largest arthroplasty register in the world,[66] is thus a strength of the study. Even with our rare outcome, we still had more than 650 deaths in the external validation cohort. This was 3.7 times as many as in the Swedish derivation cohort, yielding a reasonably large effective sample size, greater than the minimal 100 cases as sometimes recommended as a rule-of-thumb for external validation.[140] We also consider the transparent reporting of the final model, as well as the provided web-calculator, as strengths of the study.

The EPV was low for individual variables, however. It is recommended to only include potential predictor variables (includ-

ing dummy variables and polynomial coefficients) with at least 10–20 events.[123,209] More conservative recommendations require an EPV of 50.[110] We included all variables except in the presence of full separation, where no patients with a particular condition died. This might lead to biased estimates of regression coefficients and ORs, an unfortunate limitation of the study.[74]

We were not able to include any interaction terms in the final model, although all second and third-order interactions were considered, and discarded, in an earlier attempt. We simply did not have enough observed deaths to model such interaction terms adequately.

The ASA class was modeled as an unordered categorical variable for convenience, although a proportional odds model assuming an ordinal scale, might have been more accurate.[210]

The complete case analysis might be another limitation. Patients with missing data on BMI, ASA class, educational level and type of hospital were excluded from the model derivation cohort. This is a common but often criticized practice in medicine. It decreases the sample size and increases the risk of bias if data is missing not at random (MNAR). The mechanism for missing data is unknown but we might speculate that BMI is less well-recorded for patients with lower values, since those might seem less relevant to record. Patients with high values on the other hand, might be more reluctant to state their true weight if asked to, wherefore their values might get underestimated if self-reported. To exclude patients with missing BMI is thus unfortunate but more or less unavoidable. Educational level is more commonly missing for patients educated abroad, and possibly also for older patients with their education dating back to the pre-computer register era. We might hypothesize that such data are more commonly missing for lower educational levels (since higher education was relatively less common in the past). Hence, both BMI and educational level might be MNAR. ASA class and type of hospital might be missing at random (MAR), thus with missingness correlated with other variables, but not with the missing values themselves. They might even be missing completely at random (MCAR), thus without any conceivable patterns. Such variables might be imputed, either "vertically" by an estimate from the observed variable values (its mean, median or mode), or "horizontally" conditioning on non-missing variables from the same patient, such as using a regression model to find the most probable value. The latter is often preferred outside prediction modeling with the goal to estimate variable coefficients as accurate as possible. Multiple imputation using chained equations (MICE) is the preferred approach to also capture the additional uncertainty introduced by those non-observed values and to impute several missing variables for each patient. Simpler methods might be preferred in prediction settings, however, where causal assumptions are less relevant and where the computational burden is already high.[211] A potential use of a separate missing data indicator for categorical variables is more controversial. It introduces bias for the estimated model coefficients, but the practice has been advocated for models with a pure prediction purpose,[212] although such advice is also ill-perceived by others.[213]

Possible heterogeneity among surgeons, hospitals or counties might be of relevance to the model, but was ignored in the study. HGLM might have been used within the modeling process, but administrative and organizational effects might be hard to explain in clinical practice for patients considering hip arthroplasty.

CIs for the estimated coefficient values did not incorporate any uncertainty imposed by the variable selection procedure. The width of those CIs thus underestimate the true uncertainty of the model. It is rare to provide accurate CIs under those circumstances, although a method combining bootstrapping and the delta-method has been proposed.[202] It would also be desirable to provide prediction intervals for individual patients, but similar challenges apply, as well as that observations are binary (dead or not;

$Y \in \{0,1\}$), whereas probabilities are estimated on the interval scale from zero to one, $\hat{p} \in [0,1]$.

We did not formally evaluate the clinical usefulness of the prediction model as a tool for shared decision making. We believe however that the web-calculator is useful, mostly to confirm that the risk of short-term mortality is minimal for most patients. Most individuals are likely aware that high age, systematic disease (ASA class III) and severe comorbidities are risk factors of death, but to what extent might be less known.

We were not able to draw any causal conclusions regarding the potential effect of comorbidity on mortality after hip arthroplasty, since we had no observed intervention and no control group. To form such control group from observational data seems difficult, however. One possibility worth exploring might be to include data from the registry of Better management of patients with OsteoArthritis (BOA), combined with some propensity score method.

5.4 STUDY V

Individual patients' predictions were the focus of the thesis, although some assessment of aggregated population level dynamics is also useful. It is possible to predict RMST on a population level as well,[85] but this would be less useful for individual patients.

5.4.1 METHODOLOGICAL CONSIDERATIONS

We used linear regression with an identity link function and GEE, based on pseudo-observations as a semi-parametric method for covariate adjustment. Alternative link functions, such as the log-linear, were not considered but might be equally relevant[85,94,214]

An alternative modeling approach might consider a flexible parametric model, yielding smaller variance estimates and increased statistical power.[91] Cox regression is another alternative. This would require proportional hazards, however, which contradicts one of the common rationales for using RMST. It is nevertheless possible using the Breslow estimator for the cumulative baseline hazard.[215,216]

Another parameter-free alternative to GEE with pseudo-observations, is a Kaplan–Meier curve with inverse probability weighting (IPW) based on propensity scores. This is a method used for causal inference with intentional treatment groups. To estimate "propensity scores" for groups based on the Elixhauser comorbidity index, might seem far-fetched, however.[89,95]

We illustrated uncertainty of the RMST curves by pointwise 95 % CIs ($\mu_\tau \pm 1.96\sigma/\sqrt{n}$). An alternative approach with confidence bands might also apply.[179]

5.4.2 STRENGTHS AND LIMITATIONS

We consider the intuitive results of RMST as a strength of the study. Especially so with the use of pseudo-observations and GEE, which provided a natural time scale also for the effect of covariate adjustment. Traditional survival modeling based on HRs would be less interpretable, and potentially misleading without correct assumptions regarding the observed non-proportionality, as recognized in Study II–III.[217]

The applied methods assumed random censoring, which we assumed without formal assessment.* Administrative censoring is more common for patients with hip arthroplasty in later years. It is thus possible that time trends, for example caused by increased comorbidity coding, would be correlated with the censoring process.[214] A double-robust alternative allowing for non-random censoring patterns has been suggested.[86] It is, however, best suited for comparisons of intentional treatment groups, and is thus less applicable to our observed Elixhauser comorbidity. It is a strength of our linkage data, however, that we were not constrained to in-hospital deaths. If so, there might have been correlation between Elixhauser and censoring, if we assume that patients with more comorbidity are more likely to die in hospitals, whereas healthier patients might as well die of unrelated causes in the society.[85]

*This is relevant also for Study II–III.

6 CONCLUSIONS

We developed an R-package *coder* as a generic tool for data classification based on external code data (Study I). The package is released as open source software with online documentation (https://eribul.github.io/coder/).

It was not possible to accurately predict mortality from neither the Charlson, nor the Elixhauser comorbidity indices for patients with neither OA (Study II), nor FNF (Study III). A simple model with age and sex was a better alternative in both cases. However, if any comorbidity index should be used for such predictions, we recommend to use Elixhauser for OA and Charlson for FNF.

We found an alternative model to predict 90-day mortality after cemented THA due to OA (Study IV). The parsimonious main effects model considering age, sex, BMI, ASA class and the presence of cancer, CNS disease, kidney disease and obesity had an AUC statistically significantly above 0.7 due to internal and external validation. We hope that the supplementing web-calculator (https://erikbulow.shinyapps.io/thamortpred/) will aid shared decision making in clinical practice.

Although the Elixhauser comorbidity index was not sufficient to accurately predict mortality, it was associated with RMST after THA due to OA (Study V).

7 FUTURE PERSPECTIVE

Introducing new software, such as the `coder` package (Study I) is a long-term commitment. Software dependencies are often complex and new updates might break existing functionality. New versions of `coder` might therefore be prompted regardless of potentially new feature requests. New updates based on ICD-11 might also be relevant. A further improvement inspired by later versions of the `icd` package is also possible. `icd` implements an alternative classification approach using sparse matrix multiplications and linear algebra.[218] It would be relatively easy to automatically translate our regular expressions into such sparse matrices as well, using the related `decoder` package for intermediate code translation. This might further increase the computational efficiency of the package. If so, the regular expressions would still be used for a compact representation, although not for direct code comparisons for each individual code.

We developed a prediction model trained on patients operated with cemented THA due to OA 2008–2015 (Study IV). The underlying population survival changes over time, however. This, combined with possible case mix shifts, modified indications for surgery, and changes made to surgical, as well as administrative procedures, might lead to a potential calibration drift over time. It might therefore be necessary to re-calibrate the model parameters every couple of years.[219]

7.1 MACHINE LEARNING

Regression analysis, as used in this thesis, is a popular technique, even compared to modern ML algorithms. This is not only due to historical reasons. (Semi)parametrical formulations aid interpretation and generalizations as often required in the medical field. ML, such as classification, tree models (including random forest), support-vector machines, neural nets and deep learning, can perform equally well (or better), but their outcomes are harder to interpret and the required data sets are larger, since many more degrees of freedom are spent on finding relational forms and tuning parameters without a pre-specified parametric model.[110] To use larger data sets, perhaps from international collaboration, or to include images or other types of data, could be of interest in a future setting incorporating ML. A unifying framework for international collaboration of prediction modeling based on observational patient data has been initiated through Observational Health Data Sciences and Informatics (OHDSI).[220] A standardized data structure complemented by relevant open source software makes it easier to derive and validate models in different countries.[221] To combine data from different countries and sources is difficult, however, not only due to technical issues. This was brought into public attention during the initial phase of the Covid-19 pandemic, when mortality data from different countries were hard to compare.

Modeling rare/extreme events (imbalanced/unbalanced data) is challenging. Several compensatory methods have been proposed and it would be of interest to explore those in more details. XGBoost, Catboost and related methods are popular for prediction modeling and might be of relevance.

ML models are relatively common in the orthopedic literature, although the most successful applications concern medical imaging[222] such as fracture classification based on X-rays,[223] or kinetic skeletal tracking.[224] The results of such classifications are immediately recognized and comparable by human assessment, wherefore less interpretable models might be accepted. This is different from predictions of future events.

7.2 ALTERNATIVE OUTCOMES

Death is probably the most studied end point in medicine. It is the final outcome and it is often well recorded. We studied all-cause mortality since this information is most reliable. It should be possible, however, to more

clearly distinguish only the relevant causes of death. Accidental deaths due to traffic accidents or violent crimes for example might be excluded since those are unlikely related to pre-surgery comorbidity before hip arthroplasty. This is possible through data linkage with the cause of death register.[17] This could also strengthen the indication of a hypothetical causal effect in Study IV.

Other important outcomes include reoperation or revision after surgery, as well as PROMs. Those outcomes are also widely studied in the orthopedic literature and we plan to extend the modeling from Study IV to prediction of prosthesis joint infection (PJI) and dislocation after elective THA. There is similar interest to include fracture patients as well, and to build a comprehensive web calculator combining those outcomes.

7.3 ADDITIONAL PATIENT GROUPS

We studied patients with hip arthroplasty. To predict mortality for patients with OA treated with physiotherapy might be less relevant. There is no reason to believe that those patients, without any severe disease or any invasive treatment, should have a mortality rate different from the general population. Patients with hip fractures treated with internal fixation seems like a more relevant cohort, however. To predict their mortality is as relevant as for those patients treated with hip arthroplasty and such procedures are recorded in the Swedish Fracture Register.

7.4 CLINICAL USEFULNESS

An important but often neglected discussion concerns the clinical usefulness of a proposed prediction model. A model estimating probabilities of an event (death) might be most useful if this probability exceeds a threshold leading to action (a decision to operate or not).[110,225] A default threshold for $\hat{p} > 0.5$ could be assumed, thus to not operate if the risk of death exceeds 50 %. Such decision becomes irrelevant for rare but fatal events, however. Instead, the overall 90-day probability of death, 0.3 %, might be chosen as an alternative threshold, implying a decision not to operate if the probability of death exceeds this limit. This might seem overly conservative, however, and such judgment is subjective. It is more relevant to make an initiated trade-of between risk and benefit, such that the risk and severity of death is balanced against the possible benefits of the operation. Such act of balance might include factors external to the statistical model, for example the individual risk propensity for each patient. This has not been considered as part of the thesis but might constitute a relevant continuation.[226]

ACKNOWLEDGEMENT

First of all, I would like to thank my main supervisor *Szilard Nemes*! I remember the first time we met, when you had just joined our calculus class in 2005. There and then I was actually able to help you (with some multidimensional integration)! Since then, the table has turned completely, and I am greatly grateful for all what you have done for me over the years! This includes working together (as cleaners at the Swedish exhibition center/Gothia Towers and as statisticians at the regional Oncological Center, the Regional cancer center (RCC), the RC and the SHAR), as well as spending time together with friends and family! Köszönöm szépen!

Also my co-supervisors deserves a big thank you! *Göran Garellick* went over fire to secure my PhD position. There would have been no thesis without your superior determination! I have actually no idea how *Ola Rolfson* found the time and energy to manage my PhD studies among his other thousand projects. But he did! And not only did you teach me how to do research, but also, literally, how to steer a sailing boat without colliding with an air-craft carrier! The enthusiasm of *Nils Hailer* has been almost overwhelming! You did not only welcome me to work with your prediction model, but have already involved me in further research!

The thesis would not have been written without close collaboration among the co-authors. Gratitude goes to *Peter Cnudde*, also a fellow PhD student, travel partner and source of inspiration and never-ending enthusiasm! Thank you *Cecilia Rogmark* for bringing the fracture perspective and for caring about the old and frail! Thanks to *Anne Garland* for letting me share your prediction study, to *Erik Lenguerrand* for making it happen, and to *Adrian Sayers, Ashley Blom* and *Mark J Wilkinson* for letting us stand on your shoulders! I am also grateful for being invited as co-author and collaborator for the impressive work performed by my PhD peers: *Alex Wojtowicz, Rory Ferguson, Georgios Chatziagorou, Johan Simonsson, Per Johlbäck, Susanne Hansson, Ted Eneqvist, Urban Berg* and *Z Jawad*.

I am also thankful for my registry colleges version 1! My first encounter with *Kajsa Eriksson* involved a question of whether I was able to "rain dance". Admittedly, this made me a little perplex, but after that, we and Karin $\sum_{k \in K} k$, where $K = \{Davidsson, Lindberg, Pettersson\}$ conquered the world together (little did we know about world-wide pandemics and travel-bans). In the beginning of my thesis, I also had the honor to be room-mate with *Johan Kärrholm*, *Hans Lindahl*, *Henrik Malchau*, *Maziar Mohaddes* and *Daniel Odin*. I was certainly a mouse among those elephants in the room!

Also my registry colleges version 2 have made my time at SHAR a sheer pleasure! I have enjoyed the statistical and software-related collaboration with *Emma Nauclér*, *Jonatan Nåtman* and *Pär Werner*. What I appreciate the most with my fellow PhD student *Johanna Vinblad* is your shared appreciation of hyphens, en-dashes and m-dashes. Finally *Sandra Olausson*, who I highly admire for your impressive endurance regarding "fika". In light of the recent global development, I just wish I would have taken the chance more often!

I would not have been able to enroll as a PhD student without extensive support and encouragement from my previous and present, formal and in-formal bosses at RCC and RC: *Erik Holmberg, Katrín-Ásta Gunnarsdóttír, Thomas Björk-Eriksson, Peter Gidlund* and *Ulrika Frithiofsson*.

I would also like to thank my statistical colleagues for inspiration, collaboration and friendship over the years: *Anna Gennell, Caddie Zhou, Chenyang Zhang, Claudia Adok, Christian Staf, Jan Ekelund, Leyla Nunez, Linda Akrami, Ludwig Andersson, Madeleine Helmersson, Mervete Miftaraj, Mikael Holtenman, Peter Wessman, Rebecka Bertilsson* and *Stefan Franzén*.

I would also direct a direct thank to *Ramin Namitabar* for responsive help and elucidation regarding available IT-resources,

as well as an in-direct in-formal thank to *Marcus Marin* for some additional help under the radar. I am also very grateful to *Cina Holmér* who provided map, compass and porterage to navigate the administrative jungle at the university. I am also very thankful to all other colleagues at RC and RCC!

I would also take the opportunity to thank all my teachers over the years, especially: *Maria W, Britt Gars, Minni König, Katarina Lidfors and Yuri Omelchenko*. You inspired me to never stop learning!

Last, but far from least, I am also ∞ thankful for all family support! My younger brother *Simon Bülow* was directly involved in the thesis by painting the wonderful cover image. My youngest brother *Johannes Bülow* was indirectly involved by teaching me the secrets of emergency medicine and ambulatory care! My father *Hans Bülow* inspired me to read, write and publish! My mother *Margit Bülow* not only took me to the library, but also to the hospital, where you showed me your passion for health care, the importance of involving and respecting the patients, and the clinical potential in quality registers. I know you were in fact star-struck when you realized I worked in the same building as some of your heroes. I just wish I would have gotten the chance to introduce you in person!

I am likewise thankful to my extended Moldavian family. Multumesc mult! for providing distraction, wine, and for reminding me to enjoy all aspects of life!

Unfortunately, there is no way I could thank my wonderful wife *Marina* and my son *Gabriel* (Figure 7.1) enough! Not only for software related input, but also for keeping me (relatively) sane during this process. I owe you tremendously and you know that I have some serious depts to pay. I am really looking forward to spend more time with you as soon as this thesis is finalized!

Finally, my grandfather, *Lennart Johansson* (Figure 7.1), was the one who made this research personal. You asked at your hospital bed: "Do I end up in your statistics now?" You did not, although you broke your femur on the 6th of June 2018.* You were 91 years old with a Charlson comorbidity index of seven. If included in "my statistics", you would have been recorded as an "event" within six weeks. We did not know that on the day of national celebration. We cannot know the future, not even with the best registry in the world. But we can try to do our best! And that's what you always did!†

Figure 7.1: Grandpa Lennart with broken femur at the Royal River Hospital in June 2018.

*Treated by internal fixation.

†P.S. A special thank of course goes to the wholly ESO-spirit and all members of NMN. That's all!

REFERENCES

1 Lystad RP, Brown BT. 'Death is certain, the time is not': Mortality and survival in Game of Thrones. *Injury Epidemiology 2018 5:1* 2018; **5**: 44.

2 Turkiewicz A, Petersson IF, Björk J *et al.* Current and future impact of osteoarthritis on health care: A population-based study with projections to year 2032. *Osteoarthritis and Cartilage* 2014; **22**: 1826–32.

3 Inacio MCS, Paxton EW, Graves SE, Namba RS, Nemes S. Projected increase in total knee arthroplasty in the United States – an alternative projection model. *Osteoarthritis and Cartilage* 2017; **25**: 1797–803.

4 Kärrholm J, Rogmark C, Nauclér E, Vinblad J, Mohaddes M, Rolfson O. Svenska höftprotesregistret årsrapport 2018. 2019.

5 Hansson S, Bülow E, Garland A, Kärrholm J, Rogmark C. More hip complications after total hip arthroplasty than after hemiarthroplasty as hip fracture treatment: Analysis of 5,815 matched pairs in the Swedish Hip Arthroplasty Register. *Acta Orthopaedica* 2020; **91**: 133–8.

6 Jennison T, Yarlagadda R. Hip fractures. *Surgery (Oxford)* 2020; **38**: 70–3.

7 Waugh W. John charnley: The man and the hip. Place of publication not identified: Springer, 1990.

8 Ellis L, Woods LM, Estève J, Eloranta S, Coleman MP, Rachet B. Cancer incidence, survival and mortality: Explaining the concepts. *International Journal of Cancer* 2014; **135**: 1774–82.

9 Moriyama IM, Loy RM, Robb-Smith AHT. History of the statistical classification of diseases and causes of death (2011). Washington: National Center for Health Statistics, 2011 http://www.cdc.gov/nchs/data/misc/classification_diseases2011.pdf.

10 Morabia A. Epidemiology's 350th anniversary: 1662-2012. *Epidemiology (Cambridge, Mass)* 2013; **24**: 179–83.

11 Boyce N. Bills of Mortality: Tracking disease in early modern London. *The Lancet* 2020; **395**: 1186–7.

12 Graunt J. Natural and political observations mentioned in a following index, and made upon the bills of mortality. London, 1661.

13 Graunt J. Foundations of vital statistics. In: Newman JR, ed.. London: George Allen & Unwin, 1960. https://archive.org/details/TheWorldOfMathematicsVolume3/page/n13.

14 Skatteverket. Folkbokföringens historia. https://www.skatteverket.se/privat/folkbokforing/attvarafolkbokford/folkbokforingenshistoria.4.18e1b10334ebe8bc80003006.html (accessed Aug 2, 2020).

15 Stjernschantz Forsberg J. Registerforskning – Etiken bakom juridiken. Solna: Karolinska Institutet, 2013.

16 Ludvigsson JF, Almqvist C, Bonamy AKE *et al.* Registers of the Swedish total population and their use in medical research. *European Journal of Epidemiology* 2016;: 1–12.

17 Brooke HL, Talbäck M, Hörnblad J *et al.* The Swedish cause of death register. *European Journal of Epidemiology* 2017; **32**: 765–73.

18 Feinstein AR. The pre-therapeutic classification of comorbidity in chronic disease. *Journal of Chronic Diseases* 1970; **23**: 455–68.

19 Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Medical care* 1998; **36**: 8–27.

20 Höjer JA, Soop E, Hultgren G. Statistisk klassifikation av sjukdomar, skador och dödsorsaker. Kungliga Medicinalstyrelsen, 1951 https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/dokument-webb/klassifikationer-och-koder/icd6-ks54-statistik-klassifikation-sjukdomar.pdf (accessed Aug 2, 2020).

21 Engel A, Soop E. Klassifikation av sjukdomar del 1. Kungliga Medecinalstyrelse, 1965 https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/dokument-webb/klassifikationer-och-koder/icd7-ks64-klassifikation-sjukdomar.pdf (accessed Aug 2, 2020).

22 Sjöström Å, Westerholm B. Klassifikation av sjukdomar m m 1968: Systematisk förteckning. [ICD8]. Medecinalstyrelsen, 1968.

23 Socialstyrelsen. Klassifikation av sjukdomar 1987: Systematisk förteckning. Liber Information, 1987 https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/dokument-webb/klassifikationer-och-koder/icd9-ks87-inledning-1987.pdf (accessed Aug 2, 2020).

24 Socialstyrelsen. Klassifikationen ICD-10. Socialstyrelsen. https://www.socialstyrelsen.se/utveckla-verksamhet/e-halsa/klassificering-och-koder/icd-10/ (accessed Aug 2, 2020).

25 Centers for disease control and prevention (CDC). ICD-10-CM 2020. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2020/ (accessed Aug 2, 2020).

26 World health organization (WHO). International Classification of Diseases, 11th Revision (ICD-11). WHO. http://www.who.int/classifications/icd/en/ (accessed Aug 2, 2020).

27 icd.codes. ICD-10-CM to ICD-9-CM. https://icd.codes/convert/icd10-to-icd9-cm (accessed Aug 2, 2020).

28 Swedish medical products agency. Nationellt substansregister för läkemedel (NSL) - Läkemedelsverket. https://nsl.mpa.se/ (accessed Aug 2, 2020).

29 Nowbase. Nordic Welfare dataBASE. http://nowbase.org/ (accessed Aug 2, 2020).

30 Ludvigsson JF, Andersson E, Ekbom A *et al.* External review and validation of the Swedish national inpatient register. *BMC Public Health* 2011; **11**: 450.

31 Yurkovich M, Avina-Zubieta JA, Thomas J, Gorenchtein M, Lacaille D. A systematic review identifies valid comorbidity indices derived from administrative health data. *Journal of Clinical Epidemiology* 2015; **68**: 3–14.

32 Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases* 1987; **40**: 373–83.

33 Roffman CE, Buchanan J, Allison GT. Charlson comorbidities index. Journal of Physiotherapy. 2016; **62**: 171.

34 Katz JN, Chang LC, Sangha O, Fossel AH, Bates DW. Can comorbidity be measured by questionnaire rather than medical record review? *Medical Care* 1996; **34**. https://journals.lww.com/lww-medicalcare/Fulltext/1996/01000/Can_Comorbidity_Be_Measured_By_Questionnaire.6.aspx.

35 Sharabiani MTA, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. *Medical Care* 2012; **50**: 1109–18.

36 Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *Journal of Clinical Epidemiology* 1992; **45**: 613–9.

37 Romano PS, Roos LL, Jollis JG. Presentation adapting a clinical comorbidity index for use with ICD-9-CM administrative data: Differing perspectives. *Journal of Clinical Epidemiology* 1993; **46**: 1075–9.

38 D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in outcome assessment: The Charlson comorbidity index. *Methods of information in medicine* 1993; **32**: 382–7.

39 D'Hoore W, Bouckaert A, Tilquin C. Practical considerations on the use of the charlson comorbidity index with administrative data bases. *Journal of Clinical Epidemiology* 1996; **49**: 1429–33.

40 Ghali WA, Hall RE, Rosen AK, Ash AS, Moskowitz MA. Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data. *Journal of Clinical Epidemiology* 1996; **49**: 273–8.

41 Halfon P, Eggli Y, van Melle G, Chevalier J, Wasserfallen J-B, Burnand B. Measuring potentially avoidable hospital readmissions. *Journal of Clinical Epidemiology* 2002; **55**: 573–87.

42 Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *Journal of Clinical Epidemiology* 2004; **57**: 1288–94.

43 Quan H, Sundararajan V, Halfon P *et al.* Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical care* 2005; **43**: 1130–9.

44 Quan H, Li B, Couris CM *et al.* Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *American Journal of Epidemiology* 2011; **173**: 676–82.

45 Armitage JN, van der Meulen JH. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *British Journal of Surgery* 2010; **97**: 772–81.

46 Brusselaers N, Lagergren J. The Charlson comorbidity index in registry-based research. *Methods of Information in Medicine* 2017; **56**: 401–6.

47 Healthcare Cost and Utilization Project (HCUP). HCUP Elixhauser Comorbidity Software. Rockville: Agency for Healthcare Research and Quality, 2017 `www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp` (accessed Aug 2, 2020).

48 Walraven C van, Austin PC, Jennings A *et al.* A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med* 2009; **47**: 626–33.

49 Thompson NR, Fan Y, Dalton JE *et al.* A new Elixhauser-based comorbidity summary measure to predict in-hospital mortality. *Med Care* 2015; **53**: 374–9.

50 Fishman PA, Goodman MJ, Hornbrook MC, Meenan RT, Bachman DJ, Rosetti MCO. Risk adjustment using automated ambulatory pharmacy data: The RxRisk model. *Medical care* 2003; **41**: 84–99.

51 Caughey GE, Roughead EE, Vitry AI, McDermott RA, Shakib S, Gilbert AL. Comorbidity in the elderly with diabetes: Identification of areas of potential treatment conflicts. *Diabetes Research and Clinical Practice* 2010; **87**: 385–93.

52 Sloan KL, Sales AE, Liu C-F *et al.* Construction and characteristics of the RxRisk-V: A VA-adapted pharmacy-based case-mix instrument. *Medical care* 2003; **41**: 761–74.

53 Johnson ML, El-Serag HB, Tran TT, Hartman C, Richardson P, Abraham NS. Adapting the Rx-Risk-V for mortality prediction in outpatient populations. *Medical care* 2006; **44**: 793–7.

54 Inacio MCS, Pratt NL, Roughead EE, Graves SE. Using medications for prediction of revision after total joint arthroplasty. *The Journal of arthroplasty* 2015; **30**: 2061–70.

55 Inacio M, Pratt N, Roughead E, Graves S. Evaluation of three co-morbidity measures to predict mortality in patients undergoing total joint arthroplasty. *Osteoarthritis and cartilage* 2016; **24**: 1718–26.

56 Pratt NL, Kerr M, Barratt JD *et al.* The validity of the Rx-Risk comorbidity index using medicines mapped to the anatomical therapeutic chemical (ATC) classification system. *BMJ Open* 2018; **8**. DOI:10.1136/bmjopen-2017-021122.

57 Stawicki SP, Kalra S, Jones C *et al.* Comorbidity polypharmacy score and its clinical utility: A pragmatic practitioner's perspective. *Journal of emergencies, trauma, and shock* 2015; **8**: 224–31.

58 Corrao G, Rea F, Martino MD *et al.* Developing and validating a novel multisource comorbidity score from administrative data: A large population-based cohort study from Italy. *BMJ Open* 2017; **7**: e019503.

59 Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekbom A. The Swedish personal identity number: Possibilities and pitfalls in healthcare and medical research. *European Journal of Epidemiology* 2009; **24**: 659–67.

60 Rolfson O. 40 years with the Swedish Hip Arthroplasty Register. In: Acta Orthopaedica, Ezine edition. Gothenburg, Sweden: Taylor & Francis, 2019: 1–4.

61 Herberts P. Registrens historia: Svenska registren blev internationell förebild. *Ortopediskt Magasin* 2014; **2**: 12–4.

62 Kärrholm J. The swedish hip arthroplasty register (www.shpr.se). *Acta Orthopaedica* 2010; **81**: 3–4.

63 Kärrholm J, Mohaddes M, Odin D, Vinblad J, Rogmark C, Rolfson O. Svenska höftprotesregistret årsrapport 2017. 2018 `https://doi.org/10.18158/ryA0-C4pW`.

64 Rolfson O, Rothwell A, Sedrakyan A *et al.* Use of patient-reported outcomes in the context of different levels of data. *The Journal of Bone and Joint Surgery* 2011; **93**: 66–71.

65 National Joint Registry (NJR). Data-Completeness-and-quality. `https://reports.njrcentre.org.uk/Data-Completeness-and-quality` (accessed Aug 2, 2020).

66 The NJR Editorial Board. NJR 16th Annual Report 2019.Pdf. NJR `https://reports.njrcentre.org.uk/downloads` (accessed Aug 2, 2020).

67 Guo J, Geng Z. Collapsibility of Logistic Regression Coefficients. *Journal of the Royal Statistical Society Series B (Methodological)* 1995; **57**: 263–7.

68 Zorn CJW. Generalized Estimating Equation Models for Correlated Data: A Review with Applications. *American Journal of Political Science* 2001; **45**: 470–90.

69 Muff S, Held L, Keller LF. Marginal or conditional regression models for correlated non-normal data? *Methods in Ecology and Evolution* 2016; **7**: 1514–24.

70 Lee Y, Neider JA. Conditional and Marginal Models: Another View. *Statistical Science* 2004; **19**: 219–28.

71 Sutradhar R, Austin PC. Relative rates not relative risks: Addressing a widespread misinterpretation of hazard ratios. *Annals of Epidemiology* 2018; **28**: 54–7.

72 Viera AJ. Odds Ratios and Risk Ratios: What's the difference and why does it matter? *Southern Medical Journal* 2008; **101**: 730–4.

73 Bangdiwala SI. At odds with ratios. *International Journal of Injury Control and Safety Promotion* 2010; **17**: 73–6.

74 Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology* 2009; **9**: 56.

75 Bjørnstad ON. Epidemics: Models and Data using R. New York: Springer, 2018.

76 Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**: 457–81.

77 Lehman EL. The power of rank tests. *The Annals of Mathematical Statistics* 1953; **24**: 23–43.

78 Cox DR. Models and life-tables regression. *Journal of the Royal Statistical Society, Series B* 1972; **34**: 187–220.

79 Schoenfeld D. Chi-squared goodness-of-fit tests for the pro-

portional hazards regression model. *Biometrika* 1980; **67**: 145–53.

80 Schoenfeld D. Partial residuals for the proportionnal hazards regression model. *Biometrika* 1982; **69**: 239–41.

81 Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health* 1999; **20**: 145–57.

82 Thomas L, Reyes E. Tutorial: Survival estimation for cox regression models with time-varying coefficients using SAS and R. *Journal of Statistical Software* 2014; **61**: 1–23.

83 Therneau T, Crowson C, Atkinson E. Using time dependent covariates and time dependent coefficients in the Cox model. 2018.

84 Tian L, Fu H, Ruberg SJ, Uno H, Wei L-J. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics* 2018; **74**: 694–702.

85 Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* 2014; **15**: 222–33.

86 Zhang M, Schaubel DE. Double-Robust Semiparametric Estimator for Differences in Restricted Mean Lifetimes in Observational Studies. *Biometrics* 2012; **68**: 999–1009.

87 Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene* 1949; **47**: 188–9.

88 Royston P, Parmar MKB. Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology* 2013; **13**. DOI:10.1186/1471-2288-13-152.

89 Hasegawa T, Misawa S, Nakagawa S *et al.* Restricted mean survival time as a summary measure of time-to-event outcome. *Pharmaceutical Statistics* 2020; **2020**. DOI:10.1002/pst.2004.

90 Lueza B, Rotolo F, Bonastre J, Pignon J-P, Michiels S. Bias and precision of methods for estimating the difference in restricted mean survival time from an individual patient data meta-analysis. *BMC Med Res Methodol* 2016; **16**: 37.

91 Nemes S, Bülow E, Gustavsson A. A Brief Overview of Restricted Mean Survival Time Estimators and Associated Variances. *Stats* 2020; **3**: 107–19.

92 Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. 2004: 335–50.

93 Nygård Johansen M, LundbyeChristensen S, Thorlund Parner E. Regression models using parametric pseudoobservations. *Statistics in Medicine* 2020; published online June 10. DOI:10.1002/sim.8586.

94 Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 2010; **19**: 71–99.

95 Conner SC, Sullivan LM, Benjamin EJ, LaValley MP, Galea S, Trinquart L. Adjusted restricted mean survival times in observational studies. *Statistics in Medicine* 2019; **2020**: 3832–60.

96 Breiman L. Statistical modeling: The two cultures. *Statistical Science* 2001; **16**: 199–215.

97 Shmueli G. To explain or to predict? *Statistical Science* 2011; **25**: 289–310.

98 Efron B. Prediction, Estimation, and Attribution. *Journal of the American Statistical Association* 2020; **115**: 636–55.

99 Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? *BMJ* 2009; **338**: b375–5.

100 Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009; **338**. http://www.bmj.com/content/338/bmj.b604.abstract.

101 Austin SR, Wong Y-n, Uzzo RG, Beck JR, Egleston BL. Why summary comorbidity measures such as the Charlson comorbidity index and Elixhauser score work. *Medical Care* 2015; **53**: 65–72.

102 Hunt LP, Ben-Shlomo Y, Clark EM *et al.* 90-day mortality after 409 096 total hip replacements for osteoarthritis, from the National Joint Registry for England and Wales: A retrospective analysis. *The Lancet* 2013; **382**: 1097–104.

103 Bozic KJ, Ong K, Lau E *et al.* Estimating risk in medicare patients with THA: An electronic risk calculator for periprosthetic joint infection and mortality. *Clinical Orthopaedics and Related Researchő* 2013; **471**: 574–83.

104 Harris AH, Kuo AC, Bowe T, Gupta S, Nordin D, Giori NJ. Prediction models for 30-Day mortality and complications after total knee and hip arthroplasties for veteran health administration patients with osteoarthritis. *The Journal of Arthroplasty* 2018; **33**: 1539–45.

105 Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clinical Orthopaedics and Related Research* 2019; **477**: 452–60.

106 Li S, Zhang X. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. *Neural Comput & Applic* 2020; **32**: 1971–9.

107 Steyerberg EW, Uno H, Ioannidis JPA *et al.* Poor performance of clinical prediction models: The harm of commonly applied methods. *Journal of Clinical Epidemiology* 2018; **98**: 133–43.

108 Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ (Clinical research ed)* 2009; **338**: b606.

109 Pavlou M, Ambler G, Seaman SR *et al.* How to develop a more accurate risk prediction model when there are few events. *BMJ (Clinical research ed)* 2015; **351**: h3868.

110 Steyerberg EW. Clinical prediction models. Statistics for biology and health. 2nd edition, 2nd edn. Cham, Switzerland: Springer, 2019.

111 Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1999; **48**: 313–29.

112 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 2010; **33**: 1–22.

113 Zellner D, Keller F, Zellner GE. Variable selection in logistic regression models. *Communications in Statistics - Simulation and Computation* 2004; **33**: 787–805.

114 Austin PC, Tu JV. Bootstrap methods for developing predictive models. *The American Statistician* 2004; **58**: 131–7.

115 Guo P, Zeng F, Hu X *et al.* Improved variable selection algorithm using a LASSO-Type penalty, with an application to assessing hepatitis B infection relevant factors in community residents. *PLOS ONE* 2015; **10**: e0134151.

116 Baranowski R, Chen Y, Fryzlewicz P. Ranking-based variable selection for high-dimensional data. *Statistica Sinica* 2020. DOI:10.5705/ss.202017.0139.

117 Bach FR, R. F. Bolasso. Helsinki, Finland: ACM Press, 2008: 33–40.

118 Kattan MW. Judging new markers by their ability to improve predictive accuracy. *JNCI Journal of the National Cancer Institute* 2003; **95**: 634–5.

119 Collins GS, Groot JAD, Dutton S *et al.* External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology* 2014; **14**: 1–11.

120 Steyerberg EW, Vickers AJ, Cook NR *et al.* Assessing the performance of prediction models: A framework for tradi-

tional and novel measures. *Epidemiology* 2010; **21**: 128–38.

121 Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: Validating a prognostic model. *BMJ* 2009; **338**. http://www.bmj.com/content/338/bmj.b605.abstract.

122 Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JF. Internal validation of predictive models. *Journal of Clinical Epidemiology* 2001; **54**: 774–81.

123 Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical methods in medical research* 2017; **26**: 796–808.

124 Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters* 2006; **27**: 861–74.

125 Lobo JM, Jiménez-valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 2008; **17**: 145–51.

126 Verbakel JY, Steyerberg EW, Uno H *et al.* ROC curves for clinical prediction models part 1: ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Journal of Clinical Epidemiology* 2020; published online July 23. DOI:10.1016/j.jclinepi.2020.01.028.

127 Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**: 361–87.

128 Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; **115**: 928–35.

129 Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**: 337–44.

130 Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; **61**: 92–105.

131 Saha-Chaudhuri P, Heagerty PJ. Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics* 2013; **14**: 42–59.

132 Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics* 2019; **20**: 347–57.

133 Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine* 2014; **33**: 517–35.

134 Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology* 1982; **115**: 92–106.

135 Wang J. Calibration slope versus discrimination slope: Shoes on the wrong feet. *Journal of Clinical Epidemiology* 2020; published online June 4. DOI:10.1016/j.jclinepi.2020.06.002.

136 Finazzi S, Poole D, Luciani D, Cogo PE, Bertolini G. Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLoS ONE* 2011; **6**. DOI:10.1371/journal.pone.0016110.

137 Nattino G, Finazzi S, Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Statistics in Medicine* 2014; **33**: 2390–407.

138 Nattino G, Finazzi S, Bertolini G. A new test and graphical tool to assess the goodness of fit of logistic regression models. *Statistics in Medicine* 2016; **35**: 709–20.

139 Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Statistics in Medicine* 2004; **23**: 2567–86.

140 Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology* 2005; **58**: 475–83.

141 Becker RA. A brief history of S. 1994. http://www.math.uwaterloo.ca/~rwoldfor/software/R-code/historyOfS.pdf (accessed Aug 2, 2020).

142 Wikipedia. Fortran. Wikipedia. 2020; published online July 14. https://en.wikipedia.org/w/index.php?title=Fortran&oldid=967626170 (accessed Aug 2, 2020).

143 Chambers J. Assignments with the = Operator. 2001; published online Dec 16. https://developer.r-project.org/equalAssign.html (accessed Aug 2, 2020).

144 Rosell F. List of hotfixes for TIBCO Spotfire S+. Tibco community. 2019; published online Aug 26. https://community.tibco.com/wiki/list-hotfixes-tibco-spotfire-s (accessed Aug 2, 2020).

145 Wikipedia. Christopher Ahlberg. Wikipedia. 2020; published online June 20. https://en.wikipedia.org/w/index.php?title=Christopher_Ahlberg&oldid=963566443 (accessed Aug 2, 2020).

146 Wickham H. R packages. O'Relly, 2015.

147 Dalgard P. R-1.0.0 is released. 2000; published online Feb 29. https://stat.ethz.ch/pipermail/r-announce/2000/000127.html (accessed Aug 2, 2020).

148 The R foundation. The R Foundation. https://www.r-project.org/foundation/ (accessed Aug 2, 2020).

149 CRAN. The Comprehensive R Archive Network. https://cran.r-project.org/ (accessed Aug 2, 2020).

150 Wickham H, Grolemund G. R for data science : Import, tidy, transform, visualize, and model data. O'Reilly Media, 2017.

151 R Core Team. R: Memory Limits in R. https://stat.ethz.ch/R-manual/R-devel/library/base/html/Memory-limits.html (accessed Aug 2, 2020).

152 Dowle M, Srinivasan A. Data.Table. https://rdatatable.gitlab.io/data.table/ (accessed Aug 2, 2020).

153 Zeng Y, Breheny P. The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R. 2018; published online March 11. http://arxiv.org/abs/1701.05936 (accessed June 8, 2020).

154 Morondat F, Hill B, Osvald L, Vitek J. Evaluating the Design of the R Language. Springer, 2012. DOI:10.1007/978-3-642-31057-7_6.

155 Eddelbuettel D, Balamuta JJ. Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician* 2018; **72**: 28–36.

156 Charlson M. Charlson Comorbidity Index (CCI). MDCalc. https://www.mdcalc.com/charlson-comorbidity-index-cci (accessed Aug 2, 2020).

157 Wasey JO, Lang M, R Core Team *et al.* Comorbidity Calculations and Tools for ICD-9 and ICD-10 Codes. https://jackwasey.github.io/icd/ (accessed Aug 2, 2020).

158 Gasparini A. Comorbidity: An R package for computing comorbidity scores. *Journal of Open Source Software* 2018; **3**: 648.

159 McCormick P, Joseph T. Medicalrisk: Medical Risk and Comorbidity Tools for ICD-9-CM Data. 2020 https://CRAN.R-project.org/package=medicalrisk (accessed Aug 2, 2020).

160 Gordon M. Comorbidities.Icd10. 2019 https://github.com/gforge/comorbidities.icd10 (accessed Aug 2, 2020).

161 Cooper W. Wtcooper/icdcoder. 2019 https://github.com/wtcooper/icdcoder (accessed Aug 2, 2020).

162 Bülow E. Decoder: R-package to decode coded variables to plain text and the back. https://cancercentrum.bitbucket.io/decoder/ (accessed Aug 2, 2020).

163 rOpenSci software review editorial. rOpenSci Packages:

Development, Maintenance, and Peer Review. https://devguide.ropensci.org/ (accessed Aug 2, 2020).

164 Nymark M. Laglighetsprövning av realtidsregister inom cancervården. Enköping, 2017.

165 Cnudde P, Rolfson O, Nemes S *et al.* Linking Swedish health data registers to establish a research database and a shared decision-making tool in hip replacement. *BMC Musculoskeletal Disorders* 2016; **17**: 414.

166 SQLite. SQLite. https://www.sqlite.org/index.html (accessed Aug 2, 2020).

167 Garland A. Early Mortality After Total Hip Arthroplasty In Sweden. 2017. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-316989 (accessed June 1, 2020).

168 Google. Google Scholar. https://scholar.google.se/ (accessed Aug 2, 2020).

169 Russell-Rose T, Goach P, Willitts M, Thomas A. 2Dsearch. https://www.2dsearch.com (accessed Aug 2, 2020).

170 National library of medecine. PubMed. pubmed.gov (accessed Aug 2, 2020).

171 Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016; **5**: 210.

172 Institut de l'Information Scientifique et Technique. OpenGrey. http://www.opengrey.eu/ (accessed Aug 2, 2020).

173 Epistemikos Foundation. Epistemonikos: Database of the best Evidence-Based Health Care. https://www.epistemonikos.org/ (accessed Aug 2, 2020).

174 Lukacs PM, Burnham KP, Anderson DR. Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics* 2010; **62**: 117–25.

175 Uno H, Tian L, Horiguchi M, Cronin A, Battioui C, Bell J. survRM2: Comparing Restricted Mean Survival Time. 2020 https://CRAN.R-project.org/package=survRM2 (accessed Aug 2, 2020).

176 Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine* 2008; **89**: 289–300.

177 Gerds TA. Prodlim: Product-Limit Estimation for Censored Event History Analysis. 2019 https://CRAN.R-project.org/package=prodlim (accessed Aug 2, 2020).

178 Højsgaard S, Halekoh U, Yan J. The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software* 2005; **15**: 1–11.

179 Zhao L, Claggett B, Tian L *et al.* On the restricted mean survival time curve in survival analysis. *Biometrics* 2016; **72**: 215–21.

180 Hocking TD. Comparing namedCapture with other R packages for regular expressions. *The R Journal* 2019; **11**: 328–46.

181 Boeckxstaens P, De Sutter A, Vaes B, Degryse J-M. Should we keep on measuring multimorbidity? *Journal of Clinical Epidemiology* 2016; **71**: 113–4.

182 Mehta HB, Mehta V, Girman CJ, Adhikari D, Johnson ML. Regression coefficient-based scoring system should be used to assign weights to the risk index. *Journal of Clinical Epidemiology* 2016; **79**: 22–8.

183 Moons KGM, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? *Journal of Clinical Epidemiology* 2002; **55**: 1054–5.

184 Knottnerus JA, Tugwell P, Wells G. Editorial comment: Ratios should be multiplied, not added. *Journal of Clinical Epidemiology* 2016; **79**: 30.

185 Charlson ME, Wells M. Comment by M.E. Charlson and M. Wells. *Journal of Clinical Epidemiology* 2016; **79**: 29.

186 Gran JM, Wasmuth L, Amundsen EJ, Lindqvist BH, Aalen OO. Growth rates in epidemic models: Application to a model for HIV/AIDS progression. *Statistics in medicine* 2009; **28**: 221–39.

187 Kattan MW, Gerds TA. The index of prediction accuracy: An intuitive measure useful for evaluating risk prediction models. *Diagn Progn Res* 2018; **2**: 7.

188 Bjorgul K, Novicoff WM, Saleh KJ. Evaluating comorbidities in total hip and knee arthroplasty: Available instruments. *Journal of orthopaedics and traumatology* 2010; **11**: 203–9.

189 Bülow E, Nemes S, Rolfson O. Are the first or the second hips of staged bilateral THAs more similar to unilateral procedures? A study from the swedish hip arthroplasty register. *Clinical Orthopaedics and Related Research* 2020; **2020**: 11262–1270.

190 Thomas DR, Zhu P, Zumbo BD, Dutta S. On Measuring the Relative Importance of Explanatory Variables in a Logistic Regression. *J Mod App Stat Meth* 2008; **7**: 21–38.

191 Azen R, Traxel N. Using Dominance Analysis to Determine Predictor Importance in Logistic Regression. *Journal of Educational and Behavioral Statistics* 2009; **34**: 319–47.

192 Cava WL, Bauer C, Moore JH, Pendergrass SA. Interpretation of machine learning predictions for patient outcomes in electronic health records. *AMIA Annu Symp Proc* 2019; **2019**: 572–81.

193 McCaffrey D, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 2005. DOI:10.1037/1082-989X.9.4.403.

194 Jovanovic M, Radovanovic S, Vukicevic M, Poucke SV, Delibasic B. Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression. *Artificial Intelligence in Medicine* 2016; **72**: 12–21.

195 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Calster BV. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019; **110**: 12–22.

196 King G, Zeng L. Logistic regression in rare events data. *Political Analysis* 2001; **9**: 137–63.

197 Lunardon N, Menardi G, Torelli N. ROSE: A package for binary imbalanced learning. *The R Journal* 2014; **6**. DOI:10.32614/RJ-2014-008.

198 Ahmed I, Pariente A, Tubert-Bitter P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statistical Methods in Medical Research* 2018; **27**: 785–97.

199 Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 2012; **42**: 463–84.

200 Wang H, Xu Q, Zhou L. Large unbalanced credit scoring using lasso-logistic regression ensemble. *PLOS ONE* 2015; **10**: e0117844.

201 Pozzolo AD, Caelen O, Bontempi G, Johnson RA. Calibrating probability with undersampling for unbalanced classification. Cape Town, South Africa: IEEE, 2015: 159–66.

202 Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statist Med* 2016; **35**: 1159–77.

203 Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Stat Methods Med Res* 2020; **2020**. DOI:10.1177/0962280220921415.

204 Lindeløv JK. Mcp: An R Package for Regression With Multiple Change Points. Open Science Framework, 2020 DOI:10.31219/osf.io/fzqxv.

205 Liu B, Zhou C, Zhang X, Liu Y. A unified data-adaptive framework for high dimensional change point detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2020; **2020**. DOI:10.1111/rssb.12375.

206 Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine* 2019; **38**: 4051–65.

207 Grady D, SA B. Why is a good clinical prediction rule so hard to find? *Archives of Internal Medicine* 2011; **171**: 1701–2.

208 Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology* 2015; **68**: 25–34.

209 Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: Beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology* 2011; **64**: 993–1000.

210 Harrell FE. Regression modeling strategies : With applications to linear models, logistic and ordinal regression, and survival analysis, 2nd edn. Springer, 2015.

211 Kuhn M, Johnson K. Feature Engineering and Selection: A Practical Approach for Predictive Models, 1 edition. Chapman and Hall/CRC, 2019.

212 Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology* 2020; published online June. DOI:10.1016/j.jclinepi.2020.03.028.

213 van Smeden M, Groenwold RHH, Moons KG. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *Journal of Clinical Epidemiology* 2020; published online June 19. DOI:10.1016/j.jclinepi.2020.06.007.

214 Wang Y, Kung L, Byrd TA. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change* 2018; **126**: 3–13.

215 Lin DY. On the Breslow estimator. *Lifetime Data Anal* 2007; **13**: 471–80.

216 Zhang Y. A comparison of methods for estimating Restricted Mean Survival Time. 2018; published online Nov 20. `https://api.semanticscholar.org/CorpusID:207810064` (accessed Aug 2, 2020).

217 Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010; **21**: 13–5.

218 Wasey JO, Frank SM, Rehman MA. Icd: Efficient Computation of Comorbidities from ICD Codes Using Sparse Matrix Multiplication in R..

219 Davis SE, Lasko TA, Chen G, Matheny ME. Calibration Drift Among Regression and Machine Learning Models for Hospital Mortality. *AMIA Annu Symp Proc* 2017; **2017**: 625–34.

220 OHDSI. Observational Health Data Sciences and Informatics. `https://ohdsi.org/` (accessed Aug 2, 2020).

221 Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018; **25**: 969–75.

222 Cabitza F, Locoro A, Banfi G. Machine Learning in Orthopedics: A Literature Review. *Front Bioeng Biotechnol* 2018; **6**. DOI:10.3389/fbioe.2018.00075.

223 Olczak J, Fahlberg N, Maki A *et al.* Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures? *Acta Orthopaedica* 2017. DOI:10.1080/17453674.2017.1344459.

224 Bayliss L, Jones LD. The role of artificial intelligence and machine learning in predicting orthopaedic outcomes. *The Bone & Joint Journal* 2019; **101-B**: 1476–8.

225 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019; **3**. DOI:10.1186/s41512-019-0064-7.

226 Li G, Lip GYH, Marcucci M, Thabane L, Tian J, Levine MA. The number needed to treat for net effect (NNTnet) as a metric for measuring combined benefits and harms. *Journal of Clinical Epidemiology* 2020; published online June. DOI:10.1016/j.jclinepi.2020.05.031.