KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Updating and calibration of multinomial risk prediction models

Kirsten Van Hoorde

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

December 2014

Updating and calibration of multinomial risk prediction models

Kirsten VAN HOORDE

Examination committee:
Prof. dr. ir. H. Hens, chair
Prof. dr. ir. S. Van Huffel, supervisor
Prof. dr. B. Van Calster, supervisor
Prof. dr. D. Timmerman, co-supervisor
Prof. dr. A.-M. De Meyer
Prof. dr. ir. J.A.K. Suykens
Prof. dr. T. Bourne
(Imperial College, London)
Prof. dr. E.W. Steyerberg
(Erasmus MC, Rotterdam)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

December 2014

© 2014 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Kirsten Van Hoorde, Kasteelpark Arenberg 10 box 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

ISBN 978-94-6018-926-5

D/2014/7515/150

Dankwoord

Net zoals Rome niet op één dag gebouwd is komt een doctoraat niet op één dag of door één persoon tot stand. Een proefschrift is het resultaat van een lange periode vol inspanningen van vele mensen zowel rechtstreeks als onrechtstreeks. Daarom wil ik graag de tijd nemen om een aantal mensen te bedanken.

Nothing that's worthwhile is ever easy. Remember that.
Nicholas Sparks

Eerst en vooral wil ik mijn promotor, professor Sabine Van Huffel, van harte bedanken. Sabine, dank je voor je vertrouwen, voor de steun en de kansen die je me gaf. Je hebt mijn droom om te doctoreren mogelijk gemaakt. Verder gaat er ook een zeer grote dank naar mijn andere promotor, professor Ben Van Calster. Ben, ik leerde je kennen als assessor van mijn master thesis. Eenmaal gestart met mijn doctoraat bleken we bureaugenoten te zijn. Vervolgens stond je tijdens het aanvragen van mijn IWT-beurs klaar met hulp en advies ondanks het feit dat je juist papa werd. Je mail de dag voor mijn IWT-verdediging, met de quote van *Albert Einstein "If we knew what we were doing, it wouldn't be called research."*, zal ik nooit vergeten. Tijdens het samenstellen van mijn 'supervisory committee' was het niet meer dan logisch dat je mijn copromotor werd. Gaandeweg verhuisde je van de faculteit Ingenieurswetenschappen naar de faculteit Geneeskunde. Je werd mijn promotor. Ik heb het als een eer ervaren om samen met mijn (co)promotor het bureau te mogen delen en zo nauw samen te werken. Dank je voor al je vertrouwen en steun de voorbije jaren. Zonder jou zou dit doctoraat nooit de vorm gehad hebben die het nu heeft. Ook mijn copromotor professor Dirk Timmerman wil ik bedanken om voor de aangename en vruchtbare samenwerking tussen statistici en gynaecologen te zorgen. Dankzij deze samenwerking heb ik veel bijgeleerd de afgelopen jaren.

Daarnaast wil ik graag professor Suykens en professor De Meyer bedanken om te zetelen in mijn 'supervisory committee' en in de jury. Ook een speciaal woord van

dank gaat naar Yvonne Vergouwe en Ewout Steyerberg voor de vele discussies en uitwisselingen van ideeën, verder ook naar professor Steyerberg om te willen zetelen in mijn jury. Gedurende mijn doctoraat heb ik ook met diverse clinici mogen samenwerken. Dank aan de IOTA groep voor deze unieke kans (prof. Dirk Timmerman, prof. Lil Valentin, prof. Tom Bourne en prof. Antonia Testa). Professor Tom Bourne I would like to thank you for being a member of my jury. Tot slot, wil ik de voorzitter van mijn jury, professor Hugo Hens, bedanken.

Verder ben ik het Instituut voor Wetenschap en Technologie (IWT) Vlaanderen zeer dankbaar voor de financiële steun aan mijn onderzoek; ook Ida, Elsy, John, Wim, Mimi, Maarten en Liesbeth voor hun administratieve en logistieke steun en tot slot ook professor Joos Vandewalle, professor Vincent Rijmen en professor Hugo Van Hamme voor de mogelijkheid om oefenzittingen algebra te geven en voor de leuke ervaring!

Teamwork is the secret that make common people achieve uncommon result.
Ifeanyi Enoch Onuoha

Also the BioMed group deserves a thank you. Thank you for the birthday parties, the bachelor and wedding parties and so much more. Aileen, Adrian, Alexander, Amir, Anca, Ben, Bharath, Bogdan, Bori, Carolina, Devy, Diana, Dzemila, Griet, Ivan, Jan, Joachim, Katrien, Kris, Laure, Lieveke, Lieven, Maarten, Maria Isabel, Milica, Nico, Nicolas, Ninah, Otto, Rob, Rosy, Steven, Thomas, Tim, Vanya, Vladimir, Wang, Wout and Yipeng: thanks! Ik wil ook graag mijn bureaugenoten bedanken: Vanya, jij was mijn grote inspiratie. Als mijn thesisbegeleider wakkerde je mijn interesse voor een doctoraat aan, je luisterde vaak naar mijn vragen en hielp me verder met mijn onderzoek. Je stond altijd klaar met een luisterend oor. Laure, sorry voor mijn wiskundige uitleg als ik vast zat. Ik hou fijne herinneringen over aan onze conferenties en ritjes naar Gasthuisberg. Griet en Thomas, jullie waren de twee nieuwkomers in het bureau. Dank je voor alle babbels en je gezelschap. Verder wil ik ook nog enkele andere collega's bedanken die ondertussen meer dan collega's geworden zijn. Devy, ik herinner me nog als de dag van gisteren dat we begonnen aan ons doctoraat als twee vreemden. Samen ondergingen en overleefden we de IWT-oefensessies en al snel ontstond een vriendschap door de vele koffiepauzes, de babbels en de steun! Je hebt het parcours van mijn doctoraat een heel stuk aangenamer gemaakt. Ook al gaan onze professionele wegen een andere richting uit, ik ben er van overtuigd dat we elkaar niet uit het oog zullen verliezen. Katrien, je was mijn grote voorbeeld. Ik vergeet het antwoorden op mijn vele vragen, de gezellige etentjes en de terrasjes samen zeker niet. Steven (en Lies), op korte tijd veranderde collega's zijn in buren en vrienden wat resulteerde in

Nijn-sitten en in gezellige avonden babysitten op Ella en Milo. Ook de collega's in Gasthuisberg verdienen een dank je. Het was fijn om elke maandag met jullie te vergaderen en van gedachten te wisselen. Het was een fijne en unieke uitdaging om statistiek op een verstaanbare manier uit te leggen. Jeroen, dank je voor de fijne samenwerking en het luisterend oor.

There is nothing better than the encouragement of a good friend.
Katherine Hathaway

Natuurlijk verdienen ook mijn vrienden een dikke dankjewel voor alle vriendelijke woorden, voor jullie steun en aanmoedigingen, de mogelijkheid om te ontspannen en mijn werk even te vergeten. In het bijzonder wil ik Jelle en Anja bedanken voor de fijne anti-stress momenten!

The greatest gifts you can give your children are
the roots of responsibility and the wings of independence.
Denis Waitley

Verder wil ik nog enkele speciale mensen in mijn leven bedanken. Eerst en vooral mijn moeke en vake. Wie had, na het eerste jaar, kunnen denken of vermoeden dat mijn verhaal aan de universiteit op deze manier zou eindigen? Ik kan jullie niet genoeg bedanken voor jullie eindeloze geloof in mij en jullie onvoorwaardelijke steun. Steven en Maarten: ik weet dat jullie ondanks al jullie geplaag toch wel fier zijn op jullie kleine zus. Eveline, je maakte me hier op de gepaste momenten attent op! Bedankt voor alle steun, liefde en aanmoedigingen!

"How do you spell 'love'?" - Piglet
"You don't spell it ... you feel it." - Pooh
A.A. Milne

Last but not least wil ik mijn echtgenoot Willem bedanken voor alle steun! Willem, zonder jou zou ik hier niet gestaan hebben. Je was en bent nog steeds mijn maatje, mijn grote steun en toeverlaat. Je stond - als ik weer eens aan het twijfelen of piekeren was - voor me klaar en je wist me dat extra duwtje in de rug te geven. Echt super dankjewel en samen blijven we onze grootste droom najagen!

Kirsten

Leuven
December 2014

If we knew what we were doing,
it wouldn't be called research.

Albert Einstein

Abstract

Each day someone gets the diagnosis of having a tumor. At first, a decision has to be made whether the tumor should or should not be removed. For the removal of the tumor questions arise such as “Which is the best procedure?” and “Should the patient go to a specialized clinic?”. Each decision has physical, psychological and financial consequences. Clinicians make these decisions, which directly and indirectly influence many lives, each day. All of the available information should be taken into account to formulate a decision. These different kinds of information can and will have different impacts on the outcome and the decision.

Clinical prediction models use existing information (e.g. demographics and properties of the tumor) to predict a certain outcome (e.g. malignancy of a tumor). However, it can never be the purpose to use the prediction models to replace the clinicians. These models can and should support the clinicians in their decision making. Therefore the developed models need to be sufficiently robust and accurate, which corresponds with the correct discrimination of different outcome categories (i.e. discrimination), and the reliability of the predicted risks (i.e. calibration).

The development of robust and accurate prediction models requires the collection of a large dataset with respect to the number of patients and the number of variables. The performance of existing models - from the available research or literature - could be disappointing. A few possible reasons are: the model is being used in a different country, a new potential predictor is available, the protocol has changed, ... Instead of developing a new model, one could consider to update the existing model. Updating techniques start from the existing prediction model, which contains useful information e.g. with respect to relevant predictors, and alters the model slightly or completely. Updating the existing prediction model might be more resource-efficient than developing a new prediction model.

Clinical problems are often dichotomized for analysis - i.e. only two options are available, having the 'event' or not - although the underlying problem is polytomous - i.e. subdivided into many distinct subordinate parts. Most of the time a tumor is classified as being benign or malignant, leaving out classifications as borderline and invasive or ignoring even more specific distinctions as stage I invasive, stage II-IV invasive and metastatic cancer. A more specific differentiation of malignant tumors in two or more categories can help in determining the optimal surgical procedure, the correct referral of patients to specialized clinics and increase patient survival. For example, if a 25 year old woman with the desire of having a child is diagnosed with an ovarian tumor, techniques avoiding a later influence on fertility might be used during surgery if the tumor is diagnosed as borderline instead of invasive - both categories being part of the classification malignant.

In this dissertation we focus on polytomous prediction models that use multinomial logistic regression. Calibration measures and plots to assess these prediction models, based on existing dichotomous tools, are introduced. Furthermore, we suggest different nominal updating techniques that differ in complexity.

Korte inhoud

Elke dag krijgt wel iemand te horen dat hij/zij een tumor heeft. Vervolgens moet er een beslissing genomen worden om deze al dan niet weg te nemen. Bij het verwijderen van de tumor rijst de vraag wat de beste procedure is en indien het aangewezen is om de patiënt door te verwijzen naar een meer gespecialiseerd ziekenhuis. Met elke beslissing gaan er zowel fysische, psychologische als financiële gevolgen gepaard. Dagelijks maken clinici dit soort beslissingen, die rechtstreeks en onrechtstreeks invloed hebben op meerdere mensenlevens. Het is belangrijk dat bij het maken van de beslissing, alle beschikbare informatie in rekening gebracht wordt. Niet alle informatie heeft echter evenveel invloed op de uitkomst en dus op de beslissing.

Klinische predictiemodellen maken gebruik van de bestaande informatie (e.g. demografische factoren en eigenschappen van de tumor) om een bepaalde uitkomst (e.g. kwaadaardigheid van een tumor) te voorspellen. Het is geenszins de bedoeling dat deze modellen de clinici vervangen; zij kunnen hen evenwel ondersteunen in het nemen van hun beslissingen. Het is van cruciaal belang dat de ontwikkelde modellen voldoende robuust en accuraat zijn. Deze accuraatheid omvat enerzijds het correct onderscheiden van de verschillende categorieën van de uitkomst (i.e. discriminatie), en anderzijds de betrouwbaarheid van de voorspelde risico's (i.e. calibratie).

Het ontwikkelen van robuuste en accurate modellen vereist het verzamelen van een grote dataset wat het aantal patiënten en het aantal variabelen betreft. Bestaande modellen uit de voor handen zijnde literatuur, hebben niet altijd een optimale performantie. Enkele mogelijke redenen zijn: het model wordt gebruikt in een ander land dan waar het ontwikkeld werd, er is een potentiele nieuwe predictor beschikbaar of het protocol is ondertussen gewijzigd. Het is evenwel efficiënter om een bestaand model aan te passen, dan een nieuw model op te stellen. Bij updating wordt er vertrokken van het bestaande model, dat

onder andere nuttige informatie bevat wat betreft relevante predictoren, en worden er kleine of grote aanpassingen doorgevoerd. Op deze manier kunnen tijd en geld bespaard worden.

Veel klinische problemen worden gemodelleerd als het al dan niet hebben van een 'event', i.e. een dichotoom probleem, terwijl het onderliggend probleem vaak meerdere categorieën heeft, i.e. een polytoom probleem. Voor een tumor wordt meestal het onderscheid gemaakt tussen goedaardig en kwaadaardig, maar een tumor kan goedaardig, borderline of invasief zijn of zelfs nog specifieker kan het onderscheid gemaakt worden tussen goedaardig, borderline, stadium I invasief, stadium II-IV invasief en metastasen. De opsplitsing van kwaadaardige tumoren in twee of meer categorieën kan helpen in het bepalen van de optimale operatietechniek, het correct doorverwijzen van patiënten naar een gespecialiseerd ziekenhuis en de overlevingskansen van de patiënt vergroten. Bijvoorbeeld voor een vrouw van 25 jaar met een kinderwens en een eierstok tumor kan het een groot verschil maken indien het onderscheid goed- en kwaadaardig uitgebreid wordt tot goedaardig, borderline en invasief. Indien de vrouw gediagnosticeerd wordt met een borderline tumor kan er immers gebruik gemaakt worden van fertiliteitsparende technieken tijdens de operatie.

In dit doctoraat focussen we op polytome predictiemodellen die gemodelleerd worden door middel van multinomiale logistische regressie. We introduceren calibratiematen en figuren voor het evalueren van deze modellen vertrekkende van bestaande dichotome tools. Verder stellen we ook verschillende nominale updating technieken voor die verschillen in complexiteit.

Nomenclature

List of symbols

lp	linear predictor
π	probability
$\widehat{\theta}$	estimator of θ
$\mathbf{x}_i$	variable vector of patient i
x_{ij}	variable j of patient i
Y	outcome
y_i	outcome of patient i
L	likelihood function
l	log-likelihood function
$s(.)$	(vector) spline smoother
$\alpha \,\vert \beta = 1$	calibration intercept
$\alpha_i \,\vert \beta_{ii} = 1, \beta_{ij,j\neq i} = 0$	multinomial calibration intercept of category i (relative to the reference category)
β	calibration slope
$\beta_{ii} \,\vert \beta_{ij,j\neq i} = 0$	multinomial calibration slope of category i (relative to the reference category)
Δ	threshold for Nearest Shrunken Centroid

σ	kernel parameter of the Gaussian or radial basis function kernel
`k`	number of considered neighbors in `k` nearest neighbors
C	regularization parameter of Support Vector Machines
K	kernel function
$mtry$	number of candidate variables that are considered for Random Forests
$useKernel$	tuning parameter identifying whether a kernel or normal density estimate is considered for Naive Bayes

List of thesis specific symbols

k	number of considered outcome categories
METHOD M	updating method M
N	number of patients
n	number of predictors
π_i	probability of outcome i
lp_{ij}	lp of outcome i versus outcome j
lp_i	lp of outcome i (versus not i)
$\text{lp}_{m,ij}$	lp of outcome i versus outcome j using METHOD M
$\text{lp}_{m,i}$	lp of outcome i (versus not i) using METHOD M

List of basic operations

log	natural logarithm
$\|\|\cdot\|\|$	norm of $\cdot$
$\|a\|$	absolute value of a
$\prod$	product
$\sum$	sum

$\mathbf{x}^T$	transpose of $\mathbf{x}$
e	Euler's number and base of the natural logarithm

List of abbreviations

ADNEX	Assessment of Different NEoplasias in the adneXa
AFP	α-FetoProtein
AIC	Akaike's Information Criterion
ALN	Axillary Lymph Node
AUC	Area Under the Curve
B	Benefit
BIC	Bayesian Information Criterion
BOT	Borderline tumor
BT	Binary Tree (or sequential dichotomous model)
CI	Confidence Interval
CITL	Calibration-In-The-Large
df	Degrees of Freedom
DS	Discrimination Slope
ECI	Estimated Calibration Index
EP	Ectopic Pregnancy
EPV	Events Per Variable
ER	Estrogen Receptor
FP	False Positive
FPUL	Failed Pregnancy of Unknown Location
H	Harm
*h*CG	human Chorionic Gonadotropin
hCG0	human Chorionic Gonadotropin at presentation

hCGm	average of hCG0 and hCG48
hCGr	ratio of hCG48 and hCG0
hCGrc	a centered version of the ratio of hCG48 and hCG0 (i.e. hCGr-1.17)
hCG48	human Chorionic Gonadotropin after 48 hours
HER	Human Epidermal growth factor Receptor 2
H-L	Hosmer-Lemeshow
HSROC	Hierarchical Summary Receiver Operating Curve
ICSI	Intra-Cytoplasmatic Sperm Injection
IDI	Integrated Discrimination Improvement
ILC	Invasive Lobular Carcinoma
ILR	Individualized Logistic Regression
IOTA	International Ovarian Tumor Analysis
IUP	Intra-Uterine Pregnancy
IVF	In-Vitro-Fertilization
kNN	k Nearest Neighbors
LR	Logistic Regression
MFP	Multivariable Fractional Polynomials
ML	Machine Learning
MLR	Multinomial Logistic regression
NaB	Naive Bayes
NB	Net Benefit
NRI	Net Reclassification Improvement
NSC	Nearest Shrunken Centroid
PC	Pairwise Coupling
PDI	Polytomous Discrimination Index

PR	Progesterone Receptor
prog	Progesterone at presentation
PUL	Pregnancy of Unknown Location
RF	Random Forest
RMI	Risk of Malignancy Index
ROC	Receiver Operating Characteristic
ROMA	Risk of Ovarian Malignancy Algorithm
R^2	R-squared; measure of explained variance
SLN	Sentinel Lymph Node
SR	Simple Rules
SVM	Support Vector Machines
TP	True Positive
1-V-R	1-vs-rest
RBF	Radial Basis Function

List of mathematical models and scoring systems

ADNEX	Multinomial logistic regression model to predict benign, borderline, stage I invasive, stage II-IV invasive and metastatic cancer of an ovarian tumor
LR2	Dichotomous logistic regression model to predict malignancy of an ovarian tumor
M4	Multinomial logistic regression model to predict failed pregnancy (FPUL), intra-uterine pregnancy (IUP) and ectopic pregnancy (EP) of a pregnancy of unknown location (PUL)
RMI	Multimodal scoring system to predict malignancy of an ovarian tumor
ROMA	Biomarker algorithm to predict malignancy of an ovarian tumor

SR	Ultrasound rules to predict malignancy of an ovarian tumor

List of abbreviations for data and hospitals

EORTC	European Organization for Research and Treatment of Cancer
MRC	Medical Research Council
QCCH	Queen Charlotte and Chelsea's Hospital
SGH	St George's Hospital

Contents

Abstract vii

Korte inhoud ix

Nomenclature xi

Contents xvii

List of Figures xxiii

List of Tables xxvii

1 Introduction 1

1.1 Problem formulation 1

1.1.1 Model validation 2

1.1.2 Model updating 2

1.2 Aims of this thesis 3

1.3 Chapter-by-chapter overview 3

1.4 Data 5

1.4.1 Diagnosis of ovarian tumors 5

1.4.2 Prediction of resected residual mass of testicular cancer 8

1.4.3 Prediction of the outcome of pregnancies of unknown location . . . 8
1.5 Personal contributions . . . 12
1.5.1 Methodological research . . . 12
1.5.2 Clinical projects . . . 13
1.6 Summary . . . 17

2 Dichotomous risk prediction models 19
2.1 Model development . . . 19
2.1.1 Introduction . . . 19
2.1.2 Dichotomous logistic regression model . . . 20
2.1.3 Some remarks . . . 22
2.2 Types of model validation . . . 22
2.3 Model performance . . . 23
2.3.1 Evaluation of predictions . . . 23
2.3.2 Evaluation of classification . . . 31
2.3.3 Evaluation of model comparison . . . 34
2.4 Model updating . . . 36
2.4.1 Reference method . . . 36
2.4.2 Recalibration methods . . . 38
2.4.3 Revision methods . . . 38
2.4.4 Extension methods . . . 39
2.4.5 Redevelopment method . . . 40
2.5 Conclusion . . . 41

3 Multinomial risk prediction models 43
3.1 Introduction . . . 43
3.2 Approaches . . . 44

3.2.1 Multinomial logistic regression (MLR) 44
3.2.2 Sequential dichotomous model/binary tree (BT) 47
3.2.3 Pairwise coupling (PC) 49
3.2.4 Individualized logistic regression (ILR) 49
3.2.5 1-vs-rest (1-V-R) 50
3.3 Multinomial discrimination 51
3.3.1 A set of measures 52
3.3.2 One overall measure 52
3.3.3 Visualization 53
3.4 Application: comparing different methods to develop multinomial risk prediction models 53
3.5 Conclusion 59

4 Multinomial calibration 61

4.1 Introduction 61
4.2 Methods 62
4.2.1 Nominal logistic recalibration framework 63
4.2.2 Nominal non-parametric recalibration framework 65
4.2.3 Nominal calibration plots 65
4.2.4 Estimation of calibration intercepts and slopes 66
4.3 Study set-up 67
4.3.1 Diagnosis of ovarian tumors 67
4.3.2 Diagnosing of residual mass after chemotherapy for testicular cancer 68
4.4 Results 68
4.4.1 Calibration-in-the-large and calibration plots 68
4.4.2 Calibration intercepts and slopes 74
4.5 Discussion 74

4.5.1 Nominal versus dichotomous calibration 75
4.5.2 Calibration plots 75
4.5.3 Calibration intercepts and slopes 76
4.5.4 Guidelines 76
4.6 What is new? 79
4.7 Application: multinomial risk prediction model for ovarian tumors 80
4.8 Conclusion 87

5 Multinomial updating 89
5.1 Introduction 89
5.2 Dichotomous updating methods 90
5.2.1 Patients and methods 90
5.2.2 Results 97
5.2.3 Discussion 104
5.2.4 What is new? 106
5.3 Nominal updating methods 107
5.3.1 Patients and methods 107
5.3.2 Results 115
5.3.3 Discussion 124
5.3.4 What is new? 126
5.4 Conclusion 127

6 Calibration of multinomial or multiclass risk predictions 129
6.1 Introduction 129
6.2 Patients and methods 130
6.2.1 Study set-up 130
6.2.2 Case study 131
6.2.3 Performance evaluation 131

6.2.4 Logistic regression models . . . 133
6.2.5 Machine learning models . . . 134
6.2.6 Multiclass prediction models and combinations of binary prediction models . . . 136
6.2.7 Overview of the implemented approaches . . . 137
6.2.8 Tuning of hyperparameters . . . 137
6.2.9 Software . . . 139
6.3 Results . . . 139
6.3.1 Binary outcome . . . 139
6.3.2 Multiclass outcome . . . 140
6.4 Discussion . . . 148
6.4.1 Visualization versus quantification of calibration . . . 148
6.4.2 Tuning: logistic regression versus machine learning . . . 149
6.4.3 Choice of algorithms . . . 149
6.4.4 Further practical limitations . . . 149
6.5 What is new? . . . 150
6.6 Conclusion . . . 151

7 Overview and future work 153

7.1 Overview personal contributions . . . 153
7.2 Future work . . . 155
7.2.1 Sample size issues . . . 155
7.2.2 Multinomial risk prediction models . . . 156
7.2.3 Multinomial performance measures . . . 157
7.2.4 Multinomial updating . . . 157
7.2.5 Ordinal outcome . . . 158

A `R` code 159

A.1 Multinomial calibration function 159

B SAS code **165**

B.1 Multinomial calibration macro 165

C Simulation study **171**

D Multiclass calibration plots **179**

D.1 Logistic regression models 179

D.2 Machine learning models 180

Bibliography **185**

Curriculum Vitae **211**

Publication list **213**

List of Figures

1.1 Overview of chapters. 4

2.1 Receiver Operating Characteristic (ROC) curve. Solid black line: actual test; dashed black line: ideal test; solid gray line: no predictive value (i.e. random). 26

2.2 Discrimination plot with for each outcome group (a) a box plot and (b) a histogram with the predicted probabilities for outcome Y=1. 27

2.3 Parametric calibration plot showing predicted probabilities and observed proportions for both outcome categories in case of underestimation (a), overestimation (b), underfitting (c) and overfitting (d) or a combination of both (e-h). The red curve corresponds with Y= 1 and the green curve with Y= 0. 29

2.4 Decision curves. Solid gray line: prediction model; dashed black line: assume all patients have the outcome (i.e. treat all); solid black line: assume none of the patients have the outcome (i.e. treat none). 33

2.5 Relative utility curve. 33

2.6 Reclassification graph for cases (a) with and (b) without the outcome. 34

2.7 Integrated Discrimination Improvement (IDI). 35

3.1 Multinomial logistic regression model in case of an outcome with k categories (i.e. A-K). 45

3.2 Schematic overview of the various approaches for the case study of ovarian tumors with outcomes benign, borderline and invasive: (a) multinomial LR model, (b) the three different possible sequential dichotomous models, (c) pairwise coupling LR model, (d) individualized LR model with benign as reference category and (e) 1-vs-rest LR model. 46

3.3 Some examples of different possible binary trees in case of an outcome with k categories (i.e. A-K). 48

3.4 Polytomous discrimination plot: for each observed outcome category, boxplots of the predicted probabilities for the k outcome categories (e.g. k=5 and outcome categories A-E) are given. In total $k \cdot k$ boxplots are shown. The gray horizontal lines may indicate the prevalence or baseline probability of each outcome category. 54

3.5 Boxplots of polytomous discrimination index (PDI) of the various approaches for different events per variable (i.e. EPV 5, 10 and 20) (a) without and (b) with variable selection for the validation data. 56

4.1 Parametric nominal calibration plot for the validation data of the ovarian tumor case study for each outcome category separately (panel (a), (b) and (c)) and overall (panel (d)). 71

4.2 Non-parametric nominal calibration plot considering vector splines with 2 *df* for the validation data of the ovarian tumor case study. Cubic spline smoothers are added for each outcome category. 72

4.3 Parametric (panel (a)) and non-parametric (panel (b)) nominal calibration plot considering vector splines with 2 *df* for the validation data of the testicular cancer case study. Cubic spline smoothers are added for each outcome category. 72

4.4 Non-parametric calibration plots for the ovarian tumor (a-c) and the testicular cancer (d-f) case study for different degrees of smoothing (2 *df* in (a) and (d), 3 *df* in (b) and (e), 4 *df* in (c) and (f)). The reference categories were respectively benign tumors and benign tissue for the ovarian tumor case study and the testicular cancer case study. 73

4.5 'Strong' and 'weak' calibration. 77

4.6 Non-parametric calibration plots for different choices of reference category. For the ovarian tumor study the reference categories are benign (a), borderline (b) and invasive (c). For the testicular cancer study the reference categories are benign tissue (d), mature teratoma (e) and viable cancer (f). 78

4.7 Forest plot with AUCs for the different centers of the validation data for the discrimination between benign and malignant tumors. (NC means not computed) 84

4.8 Calibration plot of the ADNEX model for each outcome category on the validation data. Perfect agreement between predicted probabilities (x-axis) and observed probabilities (y-axis) corresponds with the diagonal line. Below the calibration plot the distribution of the predicted probabilities is shown in a histogram. 85

4.9 Discrimination plot of the ADNEX model on the pooled data (N=5909) after retraining. Boxplots of the predicted probabilities for each outcome category stratified for the actual observed outcome category are presented. The gray horizontal lines show the baseline probability of each outcome category. 86

5.1 Data structure: development and validation data the latter subdivided in updating and test set. 90

5.2 Logistic calibration curves for the original sequential dichotomous models. The plots use the average intercept and slope adjustments for METHOD 2 with EPV 5 over the 500 updating sets. Panel (a) gives the curves for ovarian tumor study, panel (b) for the testicular cancer study. 102

5.3 Calibration intercept and slope (with 95% CI) on the validation data. 118

5.4 Non-parametric calibration plot for the different updating methods considering vector splines with 2 *df* for the validation data of the `SGH` case study. Cubic spline smoothers are added for each outcome category. 119

5.5 Non-parametric calibration plot for the different updating methods considering vector splines with 2 *df* for the validation data of the `QCCH` case study. Cubic spline smoothers are added for each outcome category. 122

6.1 The three different considered sequential dichotomous models. . 137

6.2 Calibration plot of the validation data for (a) dichotomous logistic regression and (b) binary naive bayes. 141

6.3 Calibration plot of the validation data for the binary `k` nearest neighbors with (a) maximizing accuracy as selection criterion and (b) minimizing logloss as selection criterion. 141

6.4 Multiclass calibration plot with smoother of the validation data for (a) pairwise coupling logistic regression model, (b) binary tree support vector machines with logloss as optimization method, (c) random forest and (d) k nearest neighbors with logloss as optimization method. 145

6.5 Smoothed calibration plot of the validation data for (a) nested dichotomous logistic regression models, (b) binary tree support vector machines with logloss as optimization method, (c) multiclass random forest and (d) multiclass `k` nearest neighbors with logloss as optimization method. 146

6.6 Smoothed calibration plot of the validation data for (a-b) multiclass `k` nearest neighbors and (c-d) multiclass naive bayes with accuracy and logloss as optimization method. 147

C.1 Schematic overview of the 12 settings used in the simulation study. 172

D.1 Smoothed calibration plot of the validation data for (a) multinomial logistic regression, (b) nested dichotomous logistic regression models and (c) dichotomous logistic regression models combined via pairwise coupling. 179

D.2 Smoothed calibration plot of the validation data for (a-b) binary tree support vector machines, (c-d) support vector machines combined via pairwise coupling, (e-f) `k` nearest neighbors, (g-h) random forest, (i-j) naive bayes and (k-l) nearest shrunken centroid with accuracy and logloss as optimization method. . . 180

List of Tables

1.1 Description of the different phases of the IOTA data in terms of recruitment periods and patients 5

1.2 Description of the centers included in the IOTAstudy. 7

1.3 Considered variables in the IOTA study with their description . 9

1.4 Hospitals, number of patients, years of treatment (range) in the development and validation data of testicular cancer 10

1.5 Considered testicular cancer variables with their description . . . 11

2.1 Updating methods. For methods with an asterisk (*) shrinkage towards METHOD 2 may be considered. 37

2.2 Updating methods: formulae. For methods with an asterisk (*) shrinkage towards METHOD 2 may be considered. 40

3.1 Average PDI with 95% confidence intervals for the different approaches, without and with backward variable selection, on the development (Dev) and validation (Val) data. 55

3.2 Average PDI with 95% confidence intervals for the three different sequential dichotomous models or binary trees in case of an EPV of 20. 57

3.3 Average c-statistics with 95% confidence intervals for the dichotomous risk prediction models on the validation data. . . 57

3.4 Average calibration-in-the-large for benign (Be), borderline (Bo) and invasive (In) in % with 95% confidence intervals for the validation data. 58

3.5 Percentage of number of variables selected in each approach for different EPV. The average number of variables is given in the last column. 59

4.1 Description and descriptive statistics for the case studies for each outcome category separately. 69

4.2 Assessment of calibration-in-the-large in the validation data. . . 70

4.3 Calibration intercepts α_i when fixing β_{ii} at 1 and $\beta_{ij,j\neq i}$ at 0 and calibration slopes β_{ii} when fixing $\beta_{ij,j\neq i}$ at zero for the validation data with their 95% Wald confidence intervals. 74

4.4 Calibration slopes β_{ii} when fixing $\beta_{ij,j\neq i}$ at zero (with 95% Wald confidence intervals) and p-values obtained on the validation data for different choices of reference category. 79

4.5 Descriptive statistics of the ten a priori considered predictors stratified for outcome category in the pooled data set (N=5909). 83

4.6 The AUC and sensitivity and specificity of the ADNEX model for distinguishing benign and malignant tumors on the development, the validation and the pooled data (after retraining). For sensitivity and specificity different cut-offs are considered (i.e. a probability greater than or equal to the cut-off is classified as malignant). The exact Binomial 95% CI are reported. 83

4.7 Polytomous discrimination measures (i.e. conditional pairwise c-statistics and polytomous discrimination index (PDI)) with 95% CI of the ADNEX model on the development, the validation data and the pooled data (after retraining). The random performance corresponds with a c-statistic of 0.5 and a PDI of 0.2 (i.e. $1/k$). 84

5.1 Description and descriptive statistics for the case studies (i.e. ovarian tumor and testicular cancer). 92

5.2 Model coefficients for the original dichotomous prediction models, and average coefficients for the updated prediction models obtained on the 500 updating sets. For model redevelopment (METHOD R), only coefficients that differed from zero were considered to compute the average, i.e. coefficients from selected predictors. 94

5.3 Sizes of development, updating and test set 97

5.4 Polytomous discrimination as assessed by the PDI on the updating and test data. The values are averaged over the 500 divisions of the validation data into updating and test data. The 5th and 95th percentiles of these results are shown between square brackets. 98

5.5 C-statistics for the sequential dichotomous models averaged over the 500 updating and test data. The 5th and 95th percentiles of these results are shown between square brackets. 99

5.6 Calibration-in-the-large results on the test data, expressed as the difference in the observed prevalence (in %) and the expected prevalence (in %). The values are averaged over the 500 divisions of the validation data into updating and test data. The 5th and 95th percentiles of these results are shown between square brackets. 100

5.7 Average of the estimated updating parameters for the different considered updating methods. 101

5.8 Frequency of effect re-estimation for METHOD 3 for the 500 updating sets. 103

5.9 Frequency of variable selection and mean number of selected variables for model redevelopment (METHOD R) for the 500 updating sets. 104

5.10 Description and descriptive statistics for development data M4, updating and validation data of SGH (i.e. temporal updating) and QCCH (i.e. external updating). 110

5.11 Updating methods. 111

5.12 Polytomous discrimination index on the updating and validation data. 115

5.13 Conditional pairwise c-statistics on the updating and validation data. 116

5.14 Calibration-in-the-large (i.e. difference in observed prevalence and expected prevalence in percentage) on the validation data. 117

5.15 Calibration intercept and slope (with 95% CI) on the validation data. 121

6.1 Descriptive statistics of the ovarian tumor case study (N=5909). 132

6.2 Overview of the implemented multiclass approaches. 138

6.3 Machine learning models with their respectively considered grid for the different hyperparameters. 138

6.4 Discrimination (c-statistic) and calibration (mean calibration (in %) of the risk of malignancy and estimated calibration index (ECI)) for the binary prediction models for the validation data. 140

6.5 Discrimination (M-index and polytomous discrimination index PDI) and calibration (estimated calibration index ECI) for the multiclass prediction models for the validation data. 142

6.6 Conditional pairwise c-statistics for the multiclass prediction models for the validation data. 143

6.7 Mean calibration (in %) for the multinomial prediction models for the validation data. 144

C.1 Details for the 12 settings with respect to development and validation data. The settings with a star are reference settings with identical event rates and predictor distributions for the development and validation data. 174

C.2 Pairwise c-statistics for development and validation data. The settings with a star are reference settings with identical event rates and predictor distributions for the development and validation data. 175

C.3 Estimated calibration slopes and chi-squared test statistic (2 *df*) for different choices of reference category. The settings with a star are reference settings with identical event rates and predictor distributions for the development and validation data. 177

Chapter 1

Introduction

This Chapter starts with the problem formulation (Section 1.1), followed by the general aims of this thesis (Section 1.2). In Section 1.3 the structure of the thesis is outlined and visualized. The data of the considered case studies are discussed in Section 1.4. Finally, my personal contributions and a summary of clinical research that will not be discussed in this dissertation is given in Section 1.5.

1.1 Problem formulation

In medicine, clinicians need to diagnose patients and make different decisions (e.g. choice of optimal treatment). Their decisions are often based on implicit or explicit risk probabilities. In the current era of 'evidence based medicine', where decisions about the care of individual patients are based on the best evidence currently available, the focus goes towards explicit risks [156]. Another new tendency in this era is shared decision making. Clinicians and patients sit together and discuss risks and benefits and based on this discussion a decision is made incorporating the preferences of the patient [25, 34, 47, 76, 126]. It might be that two patients with the same medical facts opt for a different treatment.

Clinical risk prediction models can be used in order to present the evidence to clinicians and patients. These models predict the outcome based on patient specific characteristics such as demographics or clinical variables. The development of robust and accurate risk prediction models to assist clinicians in their decision making and diagnosing patients is a resource-demanding task. It

requires the collection of large datasets with respect to sample size and variables.

In the medical literature, the main focus goes to dichotomous prediction models distinguishing patients with and without an 'event' although many clinical problems are in fact polytomous and contain more than two levels. E.g. we may consider malignancy of a tumor, but a tumor can be benign, borderline or invasive.

1.1.1 Model validation

The use of risk prediction models in the process of decision making implies that the model should be evaluated with respect to performance. Model performance consists of discrimination, calibration and usefulness [169, 174]. Discrimination considers the ability to distinguish different outcome categories (e.g. having an 'event' or not), calibration covers the reliability of the predicted risks and usefulness focuses on assessing the clinical utility of a risk prediction model or the added value of a certain marker. A plethora of performance measures exist for dichotomous outcomes, such as the area under the receiver operating characteristic (ROC) curve (AUC) (discrimination), calibration plots (calibration), and decision curves (usefulness) [31, 66, 174, 220, 222]. Extensions to polytomous models are not readily available. Most research focused on discrimination measures [61, 192, 194–196].

1.1.2 Model updating

The performance of an existing model may be not optimal in a different setting (e.g. different hospital, availability of new markers/predictors). This often leads researchers to develop a completely new model. Since the existing model contains useful information (e.g. relevant predictors or the effect of predictors) this is inefficient. It is recognized that it is easier and more efficient to adapt an existing dichotomous model than to start from scratch [82, 168]. Time and money can be saved by model updating since smaller samples are needed, allowing the model development to be more resource-efficient. This is even more important for a polytomous model: the number of cases in the smallest category is often small due to low prevalence influencing the required sample size. Typically larger datasets are needed to build robust polytomous models. Updating methods for dichotomous risk prediction models, based on logistic regression, have already been suggested [83, 116, 117, 124, 167, 168]. In terms of performance the most important aspect for model updating is calibration, i.e. the correctness of predicted risks, since a different setting often implies a difference in case-mix or prevalence of the outcome [214].

1.2 Aims of this thesis

The main purpose of this dissertation project is extending evaluation of calibration and updating methods from dichotomous risk prediction models towards polytomous risk prediction models. Calibration assesses the correspondence between predicted probabilities and observed proportions, while for updating an existing prediction model is updated in order to improve predictions from future patients in a new and different setting. These methods will facilitate the development of polytomous decision support systems for clinicians. More specifically, we will focus on risk prediction models based on logistic regression and nominal polytomous or multinomial outcomes. In order to achieve this goal three steps are required:

1. The modeling of a multinomial outcome as a combination of dichotomous logistic regression models (e.g. sequential dichotomous model or binary tree (BT)) [52, 186]. In this way, existing updating techniques for logistic regression can be considered and evaluated.
2. The extension of existing dichotomous calibration tools towards multinomial outcomes.
3. The generalization of the dichotomous updating techniques towards polytomous outcomes and more specific multinomial logistic regression.

1.3 Chapter-by-chapter overview

Chapter 1 describes the problem formulation and the aims of this thesis. Furthermore, case studies and personal contributions concerning methodological research as well as clinical projects are discussed. The dichotomous logistic regression (LR) model is discussed in **Chapter 2**. First, the model building is clarified. Second, different types of validation (i.e. internal and external) as well as different performance measures are described. Third, model updating techniques going from simple recalibration methods towards more complex methods are considered. The next three chapters look in more detail into the extension of these topics towards multinomial outcomes. In **Chapter 3**, approaches to model multinomial outcomes are summarized. These can either be a combination of dichotomous prediction models or a multinomial prediction model. Furthermore, existing multinomial discrimination measures (among others the extension of discrimination measures towards nominal polytomous outcomes proposed by Van Calster et al.) are summarized [195, 196, 199]. The topic of multinomial performance calibration is tackled in **Chapter 4**,

multinomial calibration measures and plots are proposed, focusing on models developed using multinomial logistic regression [206]. **Chapter 5** considers updating techniques for multinomial outcomes. Starting from the dichotomous updating techniques in Chapter 2 a model with multinomial outcome is updated via sequential dichotomous modeling (i.e. combining dichotomous LR models) [205]. Afterwards multinomial updating techniques are proposed. As an extension of Chapter 4, the assessment of the calibration of multinomial risk models irrespective of the adopted algorithm (e.g. support vector machines (SVM) or other machine learning (ML) models) is discussed in **Chapter 6**. The final **Chapter 7**, presents conclusions and ideas for future work.

A schematic overview of the different chapters in this dissertation is shown in Figure 1.1.

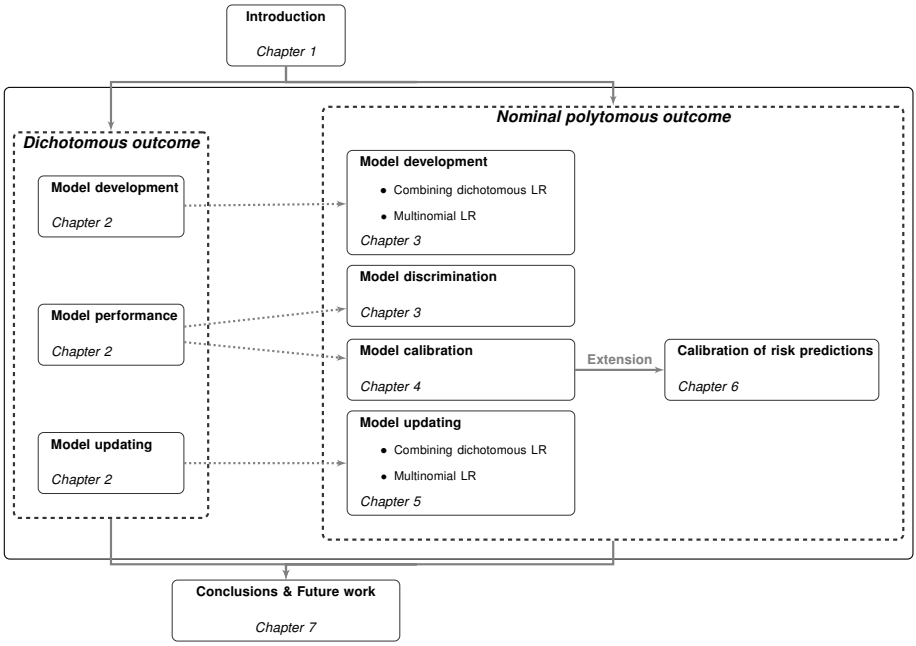


Figure 1.1: Overview of chapters.

1.4 Data

In this dissertation three different case studies are considered: the diagnosis of ovarian tumors (Section 1.4.1), the accurate prediction of resected residual mass of testicular cancer (Section 1.4.2) and the prediction of the outcome of pregnancies of unknown location (Section 1.4.3).

1.4.1 Diagnosis of ovarian tumors

The International Ovarian Tumor Analysis (IOTA) study group is an international multicenter group collecting prospectively data of women with at least one persistent adnexal mass that requires surgery. IOTA uses a standardized ultrasound examination protocol, terms and definitions [184]. The data were collected between 1999 and 2012 in different phases. The inclusion criteria for the study were presentation with at least one persistent adnexal mass, examination with trans-vaginal ultrasound by one of the principal investigators and an informed consent (oral or written depending on the requirements of the local ethics committee) prior to the ultrasound scan and surgery. The exclusion criteria were refusal of trans-vaginal ultrasonography, pregnancy at the time of presentation and more than 120 days between the ultrasound examination and the surgical removal of the mass. In case of multiple masses, the mass with the most complex ultrasound morphology was included (if similar morphology, the largest mass or the one most easily accessible by ultrasound) [183,185,204]. The integrity of the data was ensured by built-in automatic checks, manual data review and cleaning by clinicians and statisticians. The data were collected in different phases as shown in Table 1.1. The database contains 5914 patients from 24 centers in 10 countries (Table 1.2) [88,183,185,203,204].

Table 1.1: Description of the different phases of the IOTA data in terms of recruitment periods and patients

Phase	Years	Patients
1	1999-2002	1066
1b	2002-2005	507
2	2005-2007	1938
3	2009-2012	2403

The different centers contributed to different phases, thereby including a different number of patients (Table 1.2). The centers also differ with respect to their type of center. Some centers are tertiary oncologic referral centers, others are tertiary not-oncologic and the remaining are regional or public hospitals. The

latter two groups are sometimes combined and considered as other hospitals (see variable oncology referral center in Table 1.3).

A good pre-operative diagnosis of the ovarian tumor is crucial for several aspects. First, the patient needs a correct referral to specialized centers and clinicians since this improves the recovery and survival rate. Second, the decision of the optimal treatment or surgical procedure depends on the type of tumor [193]. Third, clinicians want to avoid overly invasive management or surgery with too many risks in case of a benign tumor. Fourth, in case of a malignant tumor a good diagnosis before surgery can avoid rupture and residual disease. In the literature, most of the time the distinction between benign and malignant tumors is made [90, 91, 180, 182, 183, 192, 203, 204]. Benign masses can often be treated with a conservative, expectant management or a minimally invasive surgery [24]. In this dissertation the final diagnosis of ovarian tumors is divided into three or five tumor groups. For the former a distinction between benign, borderline and invasive tumors can be made; while for the latter the invasive tumors are subdivided into stage I invasive tumors, stage II-IV invasive tumors and metastatic cancer. In case of borderline malignancies, less invasive surgical methods are typically preferred (e.g. fertility-sparing methods in young women) [40]. The type and quality of surgery affects prognosis for invasive cancers. A more conservative management may be considered for stage I invasive cancer compared with the late stage disease. Finally, for cancers metastasized to the ovary the treatment may depend on the primary cancer site [70].

These data were used in this dissertation in four different ways. First, a model distinguishing benign, borderline and invasive tumors is updated. For this research, data from phase 1, 1b and 2 were considered. The tertiary referral centers for oncology were considered as development data (N=2037) while the regional or public hospitals were used for updating and validation (N=1107). Second, different multinomial models differentiating benign, borderline and invasive tumors are compared considering all 5914 patients. Third, a model with five categories (i.e. benign, borderline, stage I invasive tumors, stage II-IV invasive tumors and metastatic cancer) was built. The data of phase 1, 1b and 2 were considered as development data (N=3511) while phase 3 was reserved for validation data (N=2403). Afterwards the model was refitted using all available data. Fourth, the data from all phases (N=5914) were randomly split into training and test data and used to assess calibration of dichotomous and multinomial risk prediction models. The variables considered in this dissertation are summarized in Table 1.3. The serum CA125 is measured according to the local protocol and is missing for some women. For the development of the multinomial risk prediction model, serum CA125 is considered and multiple imputation is used to deal with this missingness while for the calibration of dichotomous and multinomial risk prediction models only the first imputation

Table 1.2: Description of the centers included in the IOTAstudy.

Center code	City	Country	Centertype	Contributed to Phase				# patients
				1	1b	2	3	
GBE	Genk	Belgium	Regional			X	X	428
LBE	Leuven	Belgium	Tertiary Oncologic Referral	X	X	X	X	931
OCA	Ontaria	Canada	Regional			X		12
BCH	Beijing	China	Tertiary Oncologic Referral			X		73
PCR	Prague	Czech Republic	Tertiary Oncologic Referral			X	X	354
MFR	Maurepas	France	Regional	X				64
BFR	Paris	France	Regional	X				80
BIT	Bologna	Italy	Tertiary Not-Oncologic			X		135
BIT	Bologna	Italy	Tertiary Oncologic Referral				X	213
FIT	Florence	Italy	Regional				X	21
CIT	Milano	Italy	Tertiary Oncologic Referral			X	X	312
MIT	Milano	Italy	Tertiary Not-Oncologic	X		X	X	223
VIT	Milano	Italy	Regional			X		21
OIT	Monza	Italy	Tertiary Oncologic Referral	X		X	X	402
GIT	Napoli	Italy	Tertiary Oncologic Referral			X	X	15
NIT	Napoli	Italy	Tertiary Not-Oncologic	X		X	X	103
RIT	Roma	Italy	Tertiary Oncologic Referral	X	X	X	X	787
SIT	Sardinia	Italy	Regional			X	X	261
UDI	Udinese	Italy	Tertiary Oncologic Referral			X	X	64
LPO	Lublin	Poland	Tertiary Oncologic Referral			X	X	285
BSP	Barcelona	Spain	Tertiary Not-Oncologic				X	37
LSW	Lund	Sweden	Tertiary Oncologic Referral			X	X	77
MSW	Malmö	Sweden	Regional	X	X	X	X	777
SSW	Stockholm	Sweden	Tertiary Oncologic Referral				X	120
KUK	London	UK	Tertiary Oncologic Referral	X		X		119

is used [159, 163].

1.4.2 Prediction of resected residual mass of testicular cancer

The data set about testicular cancer data consists of patients that have been treated with cisplatin-based chemotherapy for advanced nonseminomatous testicular germ cell cancer [186, 219]. The surgically resected retroperitoneal lymph nodes may still contain tumor cells [186, 219]. These resected masses may contain necrosis or fibrosis (benign tissue), teratoma elements, or viable cancer (residual tumor). An accurate prediction of outcome is important to decide the optimal management strategy. In case of teratoma or viable cancer, the mass should be surgically resected. For the latter, also additional chemotherapy is needed. The data are collected between 1979 and 1999. The exclusion criteria for this study were patients with histologic pure seminoma in the primary tumor without elevated prechemotherapy serum tumor markers, patients with elevated levels of the serum tumor markers α-fetoprotein (AFP) or human chorionic gonadotropin (hCG) at the time of surgery and patients with resection after relapse [219]. The development data (N=544), from three large tertiary referral centers (i.e. Academic Hospital Groningen, Norwegian Radium Hospital and Memorial Sloan Kettering Cancer Center) and several smaller Dutch centers, are mainly treated before 1990 [166]. The validation data (N=550) consists of patients which were collected more recently. These patients were either treated in a tertiary referral center (i.e. Indiana University), a German center (i.e. Klinikum Grosshadern München) or from the European Organization for Research and Treatment of Cancer/Medical Research Council (`EORTC`/`MRC`) trial. The latter data contain only patients with good prognosis disease [41, 55, 218]. The number of patients and years of treatment for development and validation data can be seen in Table 1.4 [219].

This case study is considered for two research topics. For updating purposes, only the development data without Indiana University (N=502) and the validation data from Indiana University (N=273) were considered [219]. In the context of evaluation of calibration all the available development and validation data were used. The considered variables with their description are given in Table 1.5.

1.4.3 Prediction of the outcome of pregnancies of unknown location

The last case study considers women presenting with a pregnancy of unknown location (PUL) [29, 94, 189]. A pregnancy of unknown location is defined as a situation in which a woman has a positive pregnancy test although there

Table 1.3: Considered variables in the IOTA study with their description

Variable	Description	Unit	Type
Age	age of the patient	years	continuous
Propsol	ratio of the maximum diameters of the largest solid component and the lesion, i.e. proportion solid tissue if present		continuous
Bilateral	masses on both ovaries	binary (0 no; 1 yes)	binary
Ascites	presence of ascites	binary (0 no; 1 yes)	binary
Irregularwall	irregular internal cyst wall	binary (0 no; 1 yes)	binary
Hormtherapy	current use of hormonal therapy	binary (0 no; 1 yes)	binary
Shadows	presence of acoustic shadows	binary (0 no; 1 yes)	binary
CA125	level of CA125 tumor marker	U/ml	continuous
Famhistovca	family members with history of ovarian cancer	binary (0 no; 1 yes)	binary
Lesdmax	maximal diameter of the lesion	mm	continuous
Locules10	presence of more than 10 cyst locules	binary (0 no; 1 yes)	binary
Papnr	number of papillary projections	0 zero 1 one 2 two 3 three 4 more than three	ordinal
Oncology referral center	tertiary referral center with a specific gynecological oncological unit	binary	

Table 1.4: Hospitals, number of patients, years of treatment (range) in the development and validation data of testicular cancer

	Development data		Validation data	
	N	Years	*N*	Years
Academic Hospital Groningen	137	1979-1993	17	1995
Small Dutch centers	118	1979-1991	34	1992-1995
Norwegian Radium Hospital	127	1980-1990	23	1995
Klinikum Grosshadern München			98	1980-1996
Indiana University	42	1979-1985	273	1986-1999
Memorial Sloan Kettering Cancer Center	120	1977-1988		
EORTC/MRC Trial			105	1980-1999

are no signs of intra- or extra-uterine pregnancy on ultrasound and also no remnants of miscarriage [27, 28]. Recruitment of this data took place between 2002 and 2007 at St George's Hospital (SGH) in London and 2009 and 2013 in Queen Charlotte and Chelsea's Hospital (QCCH) . Three possible outcomes are considered: failed PUL (FPUL) , intra-uterine pregnancy (IUP) and ectopic pregnancy (EP) . The last group is the smallest and most important group. An ectopic pregnancy can result in a rupture and is dangerous for the mother [86]. For FPUL and IUP expectant management is appropriate [10, 58]. The follow-up (number of blood tests, ultrasound scans) depends on the level of suspicion of an ectopic pregnancy. Unnecessary follow-up might be psychologically and financially burdensome.

These data are considered for the topic of multinomial updating and randomly divided (2:1) into an updating and a validation set. Exclusion criteria were missing HCG measurement at presentation (HCG0) or 48 hours later (HCG48), lost to follow-up or a time interval between the two HCG measurements that was not within reasonable limits (i.e. 1-3 days). Furthermore, only women with an initial HCG level of more than 25 were considered, since in clinical practice this is the criterion of being pregnant. For SGH and QCCH respectively 1422 and 873 women were retained. The variables of interest are HCG at presentation (i.e. HCG0), HCG 48 hours later (i.e. HCG48), initial progesterone (i.e. *prog*) and age. Missingness in *prog* and age were imputed via predictive mean matching regression [159, 188].

Table 1.5: Considered testicular cancer variables with their description

Variable	Description	Unit	Type
Teratoma	the presence of teratoma elements in the primary tumor	yes/no	binary
AFP	pre-chemotherapy levels of the serum tumor markers α-fetoprotein (AFP)	normal/elevated	binary
hCG	pre-chemotherapy levels of the serum tumor markers human chorionic gonadotropin (hCG)	normal/elevated	binary
Residual mass size	the maximum diameter of the residual mass measured on computerized tomography after chemotherapy	mm	continuous
Change in diameter	the change in diameter induced by chemotherapy (per 10 %)	mm	continuous

1.5 Personal contributions

In this section I provide an overview of my personal contributions as well as an overview of clinical studies that will not be discussed in this dissertation.

1.5.1 Methodological research

The first topic of my PhD was using the existing dichotomous updating techniques in the case of a nominal outcome (Chapter 5). The nominal outcome was modeled as a combination of dichotomous logistic regression models. A sequential dichotomous model or binary tree was built. For the case study of ovarian tumors, first a distinction between benign and malignant (borderline and invasive) tumors was made, afterwards borderline and invasive tumors are discriminated. Another considered case study was testicular cancer. This work was presented at the 34th Annual Conference of the International Society of Clinical Biostatistics (ISCB) in Bergen in August 2012 and published in the *Journal Clinical Epidemiology* [205].

For the second topic calibration measures for multinomial outcomes were investigated (Chapter 4). Dichotomous calibration measures such as calibration-in-the-large, calibration intercept and calibration slope as well as (non-) parametric calibration plots were extended to multinomial outcomes. These calibration measures and plots were illustrated and discussed for the case studies of ovarian tumors and testicular cancer. Part of this work was presented at the Eastern North-American Region (ENAR) conference in Orlando in March 2013. This research resulted in a paper in *Statistics in Medicine* [206].

As third topic, the extension of dichotomous updating techniques towards multinomial outcomes modeled via multinomial logistic regression was considered (Chapter 5). For the case study of pregnancies of unknown location, several updating methods were applied and compared. This resulted in a paper that is submitted to *Statistics in Medicine*.

In the fourth and last topic not only logistic regression models but also machine learning models were considered. In this research, we started from the non-parametric recalibration framework and plots introduced in the second topic and assessed the calibration of dichotomous and multinomial risks modeled via logistic regression as well as machine learning. A paper concerning this topic is submitted to *Journal of Biomedical Informatics*.

Furthermore, I was daily supervisor of a thesis 'A comparison of different approaches to develop prediction models for multinomial outcomes' where different approaches were considered to develop multinomial prediction models

(Chapter 3). I presented this work at the 36th Annual Conference of ISCB in Vienna in August 2014.

1.5.2 Clinical projects

During my PhD I collaborated with several clinicians which resulted in a better understanding of the needs of clinicians with respect to risk prediction models and their interpretability and applicability. The challenge of explaining advanced statistics to clinicians learned me a lot. Below, a short overview of the different studies is given.

Breast cancer

A first topic, concerning survival analysis of breast cancer, was work in progress of a doctoral researcher who left our research group BioMED. This resulted in a publication in *Breast Cancer Research Treatment* [211]. The research question was whether the lobular histology, i.e. invasive lobular carcinoma (ILC) or not, has an effect on axillary lymph node (ALN) involvement. Patients with a primary operable invasive breast cancer at the Multidisciplinary Breast Center of the University Hospitals Leuven were considered. 4486 women were collected between 2000 and 2009 of which 4292 consecutive patients were eligible for this study. Reasons for exclusion were: neoadjuvant systemic therapy, no breast cancer surgery, pathologic assessment outside University Hospitals Leuven, primary metastatic disease, missing data and the combination of ER negative and PR positive. A subgroup of these patients (N=1506) underwent a sentinel lymph node (SLN) procedure. A univariate as well as multivariate analysis was done. From the former, we concluded that ILCs were slightly more likely ALN positive than non-ILCs. For the subgroup of SLNs the reverse was seen: ILCs were less likely to have ALN involvement compared to non-ILCs. The latter, multivariate analysis was done to determine whether the higher incidence is a consequence of lobular histology or other covariates that interact with lymph node involvement. The other considered covariates were: age tumor size, grade, multifocality, estrogen receptor (ER), progesterone receptor (PR) and HER (i.e. human epidermal growth factor receptor 2) status. The conclusion from this analysis was that in case of correction for the other covariates ILC tumors are less likely to have ALN involvement. For the subgroup that underwent SLN procedure the same conclusion was drawn.

Pregnancies of unknown location

A difference in nomenclature for PUL is seen in the United Kingdom (UK) and the United States (US) [11,12]. In the US more severe criteria are considered in order to locate a pregnancy site. For an intra-uterine pregnancy (IUP) a cystic structure compatible with a gestation sac should be visible in the uterus and a yolk sac and/or embryo should be seen. Similarly, an ectopic pregnancy (EP) is generally not definitively diagnosed with ultrasound if a gestation sac is not visible in the adnexa. In contrast, a more restricted definition is used in the UK. An IUP is presumed in the presence of a probable gestation sac even if a yolk sac or embryo have not been visualized, and an EP if only a homogenous mass is seen. This more restricted definition of PUL is sometimes referred to as true PUL. The question is whether pregnancies can reliably be diagnosed as an IUP or EP when a yolk sac or embryo has not be seen. In this study, the existing polytomous prediction model M4 was applied to the broad and restricted definitions from respectively US and UK considering 339 patients [29]. Diagnostic performance as well as number of patient visits and blood tests were compared. The concluding statement was that care must be taken since the subjective impression of UK lead to wrong diagnoses. This has impact on clinical practice as well as psychological well-being. In general, the broad definitions of PUL lead to fewer misdiagnoses but also more clinical visits and blood tests. A manuscript based on this research is submitted to *Obstetrics & Gynecology*.

A second topic concerns the time interval between two hCG measurements. In the prediction model M4 the predictors hCG at initial visit and hCG ratio are considered [29]. For the latter the ratio of the first and second hCG measurement are computed assuming a time interval of 48 hours. In practice this time interval may vary. The goal of this research was investigating whether the time interval influences the prediction of the PUL outcome. Of the 427 PUL, 58 were EP, 275 FPUL and 94 IUP. Of the patients 44% returned between 45 and 51 hours and 66% returned within 42 and 54 hours. From this study, we concluded that timing can have an influence but due to the clustered nature of our data around 48 hours this is not seen in our data. In order to investigate the influence of time on the prediction of PUL and the ability of triaging at an earlier stage, women should be asked to have a blood sample at e.g. 12, 24, 36 and 48 hours.

The third topic tackles the issue of accurately defining a group of PUL as low risk. If this would be feasible, follow-up could safely be reduced for these women and resources could be assigned to those women who have a higher risk of an EP. The goal of this study was validation of M4 with 5% cut-off to reliably classify a group of PUL as low risk. The performance and usefulness were temporarily (i.e. more recent data) and geographically (i.e. data from other centra) validated.

If a PUL is classified as high risk in case the risk of an EP is above 5% and otherwise as low risk, 70% of the PULs at the external units are selected for reduced follow-up. This classification was correct for 97% of the women. Of the EPs at the external units 88% were correctly assigned to extensive follow-up. This protocol could safely reduce the follow-up in the majority of PUL and is published in *Human Reproduction* [189].

A final topic consists of a two-stage test for PUL where the purpose is to distinguish EP from FP and IUP. The aim is to reduce unnecessary follow-up. Already after the initial visit, some women will be classified as low risk. For the remaining women a second, follow-up visit is required to decide whether these women have a low or high risk of having an EP.

In-vitro-fertilization pregnancies

Another clinical topic involves a population that underwent in-vitro-fertilization (IVF) with or without intra-cytoplasmatic sperm injection (ICSI) in fresh or thawed (cryo) cycles. Women presenting at the university fertility center in Leuven between September 2008 and September 2010 underwent a routine blood sample at 4 and 5 weeks gestation. The values of hCG, progesterone, follicle stimulating hormone and oestradiol were measured. The aim of this study was predicting the viability of the pregnancy at 12 weeks based on serum biomarkers. An extra research question was whether information about the embryo transfer (e.g. day of transfer and number of embryos) may improve the predictions. The database consisted of 447 women of which 170 had undergone ICSI, 155 a fresh IVF cycle and 122 had cryo cycles. Two kinds of missingness were encountered for the biomarkers in this database. Some of the values were completely missing, while for others the lab reported '>x' or '<x' due to the used assay-kit. One hundred multiple imputation datasets were considered accounting for these two types of missingness [159, 163].

Ovarian tumors

This last clinical topic involves ovarian tumors. First of all, a comparison was made between the Risk of Ovarian Malignancy Algorithm (ROMA) and an ultrasound based prediction model LR2. The first is a combination of two markers (i.e. CA125 and HE4) while the latter is a logistic regression model including six variables [123, 185]. These variables are: age of the patient, the presence of ascites, the presence of blood flow within a papillary projection, maximal diameter of the solid component, irregular internal cyst wall and presence of acoustic shadows. The discriminative ability of both models, with

respect to predicting malignancy, was assessed and compared at completion of the study and did not influence the process of decision making. For this study, 216 benign and 144 malignant tumors were considered. The test performance of LR2 was better than ROMA irrespective of stratification for menopausal status. This research resulted in a journal paper in *Gynecologic Oncology* [91].

Second, a systematic review and meta-analysis of presurgical diagnosis of adnexal tumors was executed and published in *Human Reproductions Update* [90]. For this research mathematical models and scoring systems designed to predict pre-operatively ovarian cancer containing at least two variables were considered. Nineteen different prediction models were evaluated and compared. For the meta-analysis we considered those prediction models with external validation in at least two different studies accounting for at least 1000 adnexal masses. In case of multicenter studies, the results per center were considered if available. A graphical representation of the individual study results was given by plotting sensitivity and specificity with their 95% confidence intervals. For those models with a common cut-off point to define a positive test result for the different studies, a pooled summary estimate of the sensitivity and specificity with their corresponding 95% confidence interval was obtained using a bivariate random effects model [90, 109, 115]. In case of variation in cut-off points for the different studies a bivariate hierarchical summary receiver operating curve (HSROC) model was fitted [108, 109, 115]. Furthermore, the prediction model LR2 and the scoring system Simple Rules (SR) were compared with the Risk of Malignancy Index (RMI) in pre- and postmenopausal women on two multicenter cohorts. Our systematic review and meta-analysis showed that the best tests to distinguish benign and malignant tumors in a pre-operative setting are LR2 and SR. A summary of this paper in Dutch was published in *Nederlands Tijdschrift voor Obstetrie en Gynecologie* [89] and a summary in English has been submitted to *The Obstetrician & Gynecologist*.

Third, I was involved in the development of a new multiclass ADNEX (Assessment of Different NEoplasias in the adneXa) model differentiating five tumor groups (cf. Chapter 4). For this paper all the available IOTA data were considered. A multinomial logistic regression model was developed and validated on different phases of the data. Afterwards the model was refitted on all data. The prediction model contains nine predictors: the age of the patient, CA125, maximal diameter of the lesion, proportion of solid tissue, more than 10 cyst locules, number of papillary projections, acoustic shadows, ascites and type of center. This ADNEX model may improve the decision making and reduce morbidity. This research is published in the *BMJ* [197]. I was also involved in the development of the apps and the website of this prediction model (`http://www.iotagroup.org/adnexmodel/site%20iota.html`).

1.6 Summary

In this Chapter, the issues and aims of this thesis were mentioned. Afterwards, a chapter-to-chapter overview was given. The case studies used in this dissertation were explained. To end, personal contributions concerning methodological and clinical research are stated.

Chapter 2

Dichotomous risk prediction models

*This Chapter gives an overview of the dichotomous logistic regression (*LR*) model. First, in Section 2.1, the development of a dichotomous* LR *model is introduced. Subsequently, Section 2.2 describes how a prediction model can and should be validated before being used in clinical practice. Further, Section 2.3 describes measures to assess model performance on the level of the predictions or the classification as well as model comparisons. Finally, the topic of model updating is introduced in Section 2.4.*

2.1 Model development

2.1.1 Introduction

A prediction model needs to capture the underlying structure of the data. It might be that this structure is not sufficiently discovered (i.e. underfitting) or that the model is too specific for this population (i.e. overfitting). The general aim is that the model still needs to be applicable for new patients that are similar to the training data. The challenge is finding a relatively simple model, with respect to method as well as predictors, that still fits the data. Hand stated that "Simple methods typically yield performance almost as good as more sophisticated methods..." as well as "... the apparent superiority of more sophisticated methods may be something of an illusion" [59]. Many medical

applications use a prediction model based on logistic regression. Machine learning methods might be an alternative but the question remains when these methods outperform an established approach such as logistic regression in practice [16]. Recent studies reported similar performance of logistic regression versus machine learning techniques [190, 193, 201]. Steyerberg et al. mentioned that they envision that logistic regression modeling will remain the default modeling approach to probability estimation in medical risk prediction [173]. In this dissertation we will consider logistic regression, which allows clinical interpretation [173, 235].

2.1.2 Dichotomous logistic regression model

A dichotomous logistic regression model has a dichotomous or binary outcome, representing the presence ($\mathrm{y}_i = 1$) or absence ($\mathrm{y}_i = 0$) of an 'event', contrary to a linear regression model with a continuous outcome [2, 4, 71, 79, 95, 209]. Therefore both models differ with respect to the choice of parametric model and its assumptions. An advantage of the logistic regression model is that due to the use of a logistic distribution the probabilities are automatically between zero and one as required [4, 37]. In standard (dichotomous) logistic regression (LR), the purpose is to assess the influence of one or more explanatory variables, which can be numeric and/or categorical, on the dichotomous outcome variable. In this way, a clinical interpretation of the model is possible. The aim is to predict the probability of outcome $\mathrm{y}_i = 1$, given a set of n independent variables, $P(\mathrm{Y} = 1|\mathbf{x}) = \pi_1(\mathbf{x})$ where $\mathbf{x}^T = [\mathrm{x}_1\ \mathrm{x}_2\ \mathrm{x}_3\ ...\ \mathrm{x}_n]$. The logit of the logistic regression model can be written as follows:

$$\text{logit}\left[\pi_1(\mathbf{x})\right] = \log\left(\frac{\pi_1(\mathbf{x})}{1-\pi_1(\mathbf{x})}\right) = a + \sum_{j=1}^{n} b_j \mathrm{x}_j = a + b^T\mathbf{x} = \mathrm{lp}_1, \qquad (2.1)$$

where a is the intercept and $b = [b_1\ b_2\ ...\ b_n]^T$ a vector weighting the effect of each predictor in the linear combination. The sign of b_j indicates whether $\pi_1(\mathbf{x})$ increases or decreases as x_j increases. The higher $|b_j|$ the steeper the climb or descent. A b_j close to zero means that the effect of the predictor x_j on the outcome is very low. The output of the model $\pi_1(\mathbf{x})$ is a probability between zero and one (due to the logit transformation) and is obtained with the following formula:

$$\pi_1(x) = \frac{\exp(a + b^T\mathbf{x})}{1 + \exp(a + b^T\mathbf{x})}. \qquad (2.2)$$

Fitting this LR model means that we need to obtain an estimated $\widehat{\theta} = \left[\widehat{a}\ \widehat{b}_1\ \widehat{b}_2\ ...\ \widehat{b}_n\right]^T$. This $\widehat{\theta}$ is attained via maximum likelihood, returning the

values which maximize the probability of obtaining the observed set of data. Suppose that we have a sample of N (with N the number of observations in our training data) independent observations $(\mathbf{x}_i, \mathrm{y}_i)$ with $i = 1, 2..., N$. The probability of having an event is denoted as $\pi_1(\mathbf{x}_i)$. The variable y_i indicates the presence or absence of a disease and is therefore zero or one (e.g. having a malignant tumor). If patient i with variables $\mathbf{x}_i$ has an event ($\mathrm{y}_i = 1$), the contribution to the likelihood is $\pi_1(\mathbf{x}_i)$; while otherwise the contribution to the likelihood is $1 - \pi_1(\mathbf{x}_i)$. Considering all N observations, the following likelihood function is obtained:

$$L(\theta) = \prod_{i=1}^{N} \pi_1(\mathbf{x}_i)^{\mathrm{y}_i} (1 - \pi_1(\mathbf{x}_i))^{1-\mathrm{y}_i} \tag{2.3}$$

Equation 2.3 expresses the probability of the training data as a function of the unknown parameters. The corresponding log likelihood function is given by:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N} \left(\mathrm{y}_i \log\left[\pi_1(\mathbf{x}_i)\right] + (1 - \mathrm{y}_i) \log\left[1 - \pi_1(\mathbf{x}_i)\right]\right) \\ &= \sum_{i=1}^{N} \left(\mathrm{y}_i \log[\exp(a + b^T \mathbf{x}_i)] - \log[1 + \exp(a + b^T \mathbf{x}_i)]\right) \\ &= \sum_{i=1}^{N} \left(\mathrm{y}_i \left(a + b^T \mathbf{x}_i\right) - \log[1 + \exp(a + b^T \mathbf{x}_i)]\right) \end{aligned} \tag{2.4}$$

If we now take the derivative of the log likelihood function with respect to θ, the following $n + 1$ likelihood equations are obtained:

$$\begin{cases} \dfrac{\delta l(\theta)}{\delta a} = \displaystyle\sum_{i=1}^{N} \mathrm{y}_i - \frac{e^{a+b^T \mathbf{x}_i}}{1 + e^{a+b^T \mathbf{x}_i}} = \sum_{i=1}^{N} \left(\mathrm{y}_i - \pi(\mathbf{x}_i)\right) = 0 \\ \dfrac{\delta l(\theta)}{\delta b_1} = \displaystyle\sum_{i=1}^{N} \mathrm{y}_i x_{i1} - \mathrm{x}_{i1} \frac{e^{a+b^T \mathbf{x}_i}}{1 + e^{a+b^T \mathbf{x}_i}} = \sum_{i=1}^{N} \mathrm{x}_{i1} \left(\mathrm{y}_i - \pi(\mathbf{x}_i)\right) = 0 \\ \ldots \\ \dfrac{\delta l(\theta)}{\delta b_n} = \displaystyle\sum_{i=1}^{N} \mathrm{y}_i x_{in} - \mathrm{x}_{in} \frac{e^{a+b^T x_i}}{1 + e^{a+b^T x_i}} = \sum_{i=1}^{N} \mathrm{x}_{in} \left(\mathrm{y}_i - \pi(\mathbf{x}_i)\right) = 0 \end{cases} \tag{2.5}$$

Since the equations in 2.5 are nonlinear, the solution $\widehat{\theta}$ of these equations is obtained via an iterative re-weighted least squares procedure.

2.1.3 Some remarks

An important point in the development of a robust and accurate logistic regression model is the required sample size or the number of considered predictors. In the literature, it is suggested that for a dichotomous logistic regression model 10 events (i.e. patients belonging to the smallest outcome category) are required for each predictor (i.e. events per variable, EPV) [35, 64, 65, 134, 164, 172, 223]. The simulation study of Peduzzi et al. showed that in case of an EPV below 10, regression estimates were biased, variability was increased, confidence intervals were too wide and had non proper coverage and significance testing was too conservative while for an EPV of 10 or greater no major problems were observed [134]. The rule of thumb of EPV$\geq$ 10 indicates that caution is required when the dataset at hand has an EPV below 10 [134,223].

Another aspect concerning model development is its assumptions and variable selection. The variables of a logistic regression model are assumed to be linear in the logit and additive, interaction terms should be explicitly specified [79]. The variable selection for a prediction model can be done based on clinical knowledge or via a selection procedure (e.g. stepwise, best subset). In the literature criticisms concerning the performance of automated variable selection techniques are reported [64, 164, 177]. Variable selection for clinical prediction models should never be solely based on a selection procedure but also be discussed and evaluated together with clinicians.

2.2 Types of model validation

Before risk prediction models can be used, validation in clinical practice with respect to their performance is needed. Different types of validation can be considered: internal, temporal and external validation [5, 6, 87].

Internal validation

Splitting the data random into a training and test set is a common approach for internal validation [75, 165]. The training set is used for development while the developed model is validated on the test data. Due to the similarity of both data sets, the results tend to be optimistic.

In case of data-splitting, a non-random split (e.g. by center or by time) may be preferred. An alternative for data-splitting is cross-validation or bootstrapping [46, 165]. This type of validation still cannot provide information about the model performance in another data set.

External validation

Considering patients from a similar population in other center(s) or countries is called external validation. In this way, the generalizability of the model towards other centra or countries is assessed (geographical transportability) [87, 97, 116, 217].

Temporal validation

A third type of validation is internal in terms of population, but external in terms of time [87, 116]. Therefore this temporal validation falls in-between internal and external validation [6]. The model is evaluated on subsequent patients from the same center(s) and assesses temporal transportability [97]. Although similarities can be seen with internal validation with a split based on time, temporal validation is done on a separate dataset.

2.3 Model performance

A crucial aspect in the process of model building is the evaluation of the model performance [5]. The performance of a dichotomous LR model can be assessed at different levels: the predictions on itself, the classification based on these predictions (i.e. specifying a threshold or cut-off and evaluating whether the predicted risks are above or below) and the added value when comparing two models.

2.3.1 Evaluation of predictions

Performance measures of the predictions on itself can cover discrimination and calibration jointly or assess them separately. In this section some common measures are described.

Overall performance

R-squared. The explained variation (R^2) evaluates discrimination and calibration [54, 63, 114, 118, 174]. This measure indicates the proportion of variation of the outcome Y explained by the considered model, or $1-R^2$ indicates the proportion of unexplained variation [127]. In logistic regression often Nagelkerke's R^2 is used [127]:

$$R^2 = \frac{1 - \left[\frac{L(\theta_0)}{L(\theta_M)}\right]^{2/N}}{1 - L\left(\theta_0\right)^{2/N}}, \tag{2.6}$$

with $L(\theta_0)$ the (maximized) likelihood of the null model (i.e. the intercept only model), $L(\theta_M)$ the (maximized) likelihood of the model and N the number of patients. Another frequently used alternative is McFadden's R^2:

$$R^2 = 1 - \frac{l(\theta_M)}{l(\theta_0)}, \tag{2.7}$$

where $l(\theta_M)$ is the (maximized) log-likelihood of the model, and $l(\theta_0)$ the (maximized) log-likelihood of the null model (i.e. the intercept only model) [114, 118].

Brier score. The Brier score, computed as the squared difference between the actual dichotomous outcome and the predicted probabilities, is a possible alternative and a quadratic scoring rule: $(Y - \widehat{\pi})^2$ [54]. A perfect model (i.e. corresponding predicted probabilities and actual outcome) results in a Brier score of zero. The maximum Brier score depends on the average prediction. Therefore, a scaled Brier score is proposed:

$$\text{Brier}_{\text{scaled}} = 1 - \frac{\text{Brier}}{\text{Brier}_{\text{max}}}, \tag{2.8}$$

where $\text{Brier}_{\text{max}} = \text{mean}(\pi)(1 - \text{mean}(\pi))$.

Discrimination

Whether a model assigns higher risk predictions to patients with an outcome than those without an outcome is assessed via discrimination [169, 174].

Area under the Receiver Operating Characteristic (ROC) Curve (AUC). The most common metric is the area under the ROC curve (Figure 2.1) [62–64, 130,

132, 145, 221]. In case of a binary outcome, this AUC corresponds with the concordance-statistic, c-statistic or c-index. In the ROC curve the sensitivity (or true positive (TP) rate) is plotted versus 1 - specificity (or false positive (FP) rate). For the example of predicting the malignancy of a tumor, TP corresponds with malignant tumors correctly classified as malignant, while FP corresponds with benign tumors wrongly classified as malignant. The TP-rate expresses the proportion of predicted malignancies amongst the actual malignant tumors. Similarly, the FP-rate indicates the proportion of predicted malignancies amongst the actual benign tumors. Sensitivity and specificity are computed and plotted for all possible cut-offs. The AUC corresponds with the probability that given two subjects (one with a benign and one with a malignant tumor) the model assigns a higher probability of malignancy to the latter. In case of a perfect test the resulting AUC is one, while in case of a random test the AUC is 0.5. This metric is a rank-order statistic and is thus insensitive towards systematic errors in predicted risks that maintain the order (i.e. calibration) [138,198]. Furthermore, it does not incorporate misclassification costs [122, 169, 171, 222]. Finally, the c-statistic can be disappointing due to incorrect regression coefficients but the c-statistic may also be influenced by the considered case-mix [214]. This case-mix refers to differences in patient characteristics which can be reflected by the considered predictors as well as by non-considered predictors and due to several reasons (e.g. other inclusion criteria). Vergouwe et al. suggested a case-mix corrected c-statistic to disentangle the case-mix effect from incorrect regression coefficients [214].

Discrimination slope (DS). Another simple measure to assess discrimination is the discrimination slope [228]. This performance measure computes the absolute difference in average predictions for patients with and without the outcome. For our example this means the difference in average probability of patients with a benign tumor and average probability of patients with a malignant tumor. A visualization with a boxplot and histogram is possible (Figure 2.2). The less overlap between the boxplots or histograms, the better the discrimination between both outcome categories.

Calibration

A model with good discriminative ability may still have poor correspondence between predicted and observed probabilities. The predicted risk should be reliable and accurate [167,169,174].

Calibration-in-the-large (CITL). Calibration can be assessed by looking at an overall measure, such as calibration-in-the-large which indicates the difference

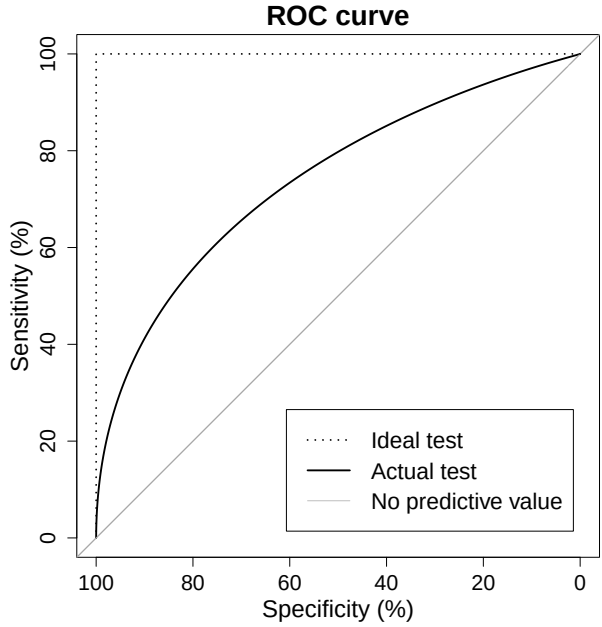
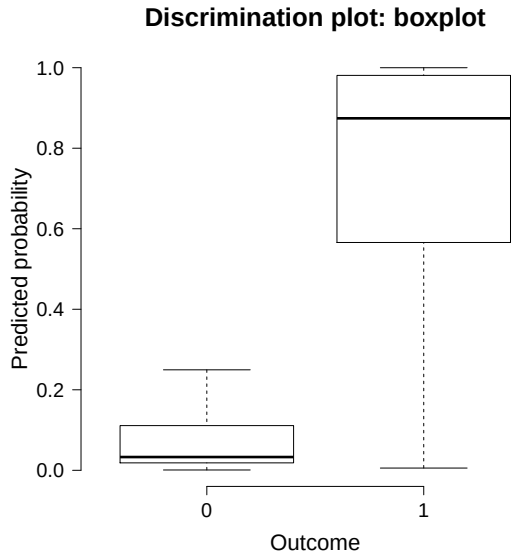


Figure 2.1: Receiver Operating Characteristic (ROC) curve. Solid black line: actual test; dashed black line: ideal test; solid gray line: no predictive value (i.e. random).

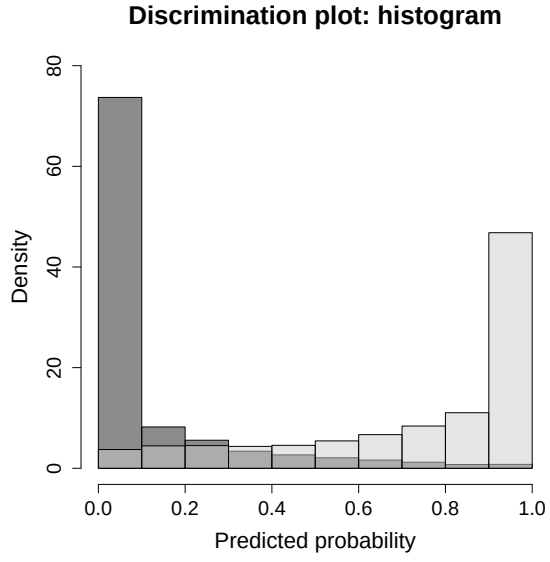
between the observed prevalence and the mean predicted probability [117]. This population-based level does not guarantee calibrated risks at the individual patient level.

Hosmer-Lemeshow (H-L) test-statistic. More interesting is the accuracy of the predicted risks for actual patients. Telling a patient that (s)he has a risk of 2% of a malignant tumor should correspond with an actual risk of 2% and not 20% or even more. The Hosmer-Lemeshow test statistic is a goodness-of-fit test which assesses whether the observed event rates match the expected event rates in g subgroups of the patient population [78, 79]. These subgroups are usually based on deciles of predicted risks. Drawbacks of this test are arbitrariness of subgroup construction, low power and instability [63, 136, 217].

Calibration intercept and slope. The calibration intercept indicates systematically too high or too low predictions, while the calibration slope expresses predictions



(a)



(b)

Figure 2.2: Discrimination plot with for each outcome group (a) a box plot and (b) a histogram with the predicted probabilities for outcome Y=1.

that are too extreme or not extreme enough [36]. These measures are obtained via the logistic recalibration framework:

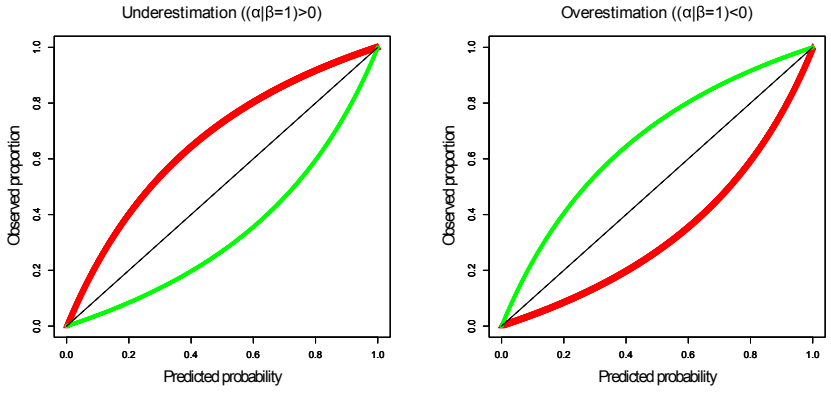
$$\log\left(\frac{P(\mathrm{Y}=1)}{1-P(\mathrm{Y}=1)}\right)=\log\left(\frac{P(\mathrm{Y}=1)}{P(\mathrm{Y}=0)}\right)=\log\left(\frac{\pi_1}{\pi_0}\right)=\alpha+\beta\ \mathrm{lp}_1. \quad (2.9)$$

with Y the outcome and lp_1 the linear predictor of the prediction model for the validation data. The outcome Y is regressed on this linear predictor lp_1 (see Equation 2.1) through a logistic regression model. The calibration slope β indicates whether the model coefficients are underfitted ($\beta > 1$) or overfitted ($\beta < 1$). The latter may reflect the need for shrinkage of the coefficients [32,210]. The intercept in case of a calibration slope of one (i.e. $\alpha\,|\beta = 1$) reflects overestimation ($\alpha\,|\beta = 1 < 0$) or underestimation ($\alpha\,|\beta = 1 > 0$). Perfectly calibrated risks correspond with a calibration intercept of zero and a calibration slope of one. A joint likelihood ratio chi-squared test can be considered to test the calibration intercept and slope ($df = 2$) [117]. The difference in log-likelihood between the original prediction model and the model with intercept and slope adaptation is computed and evaluated with two degrees of freedom. Separate tests for β and $\alpha\,|\beta = 1$ can be conducted in case of significance [117].

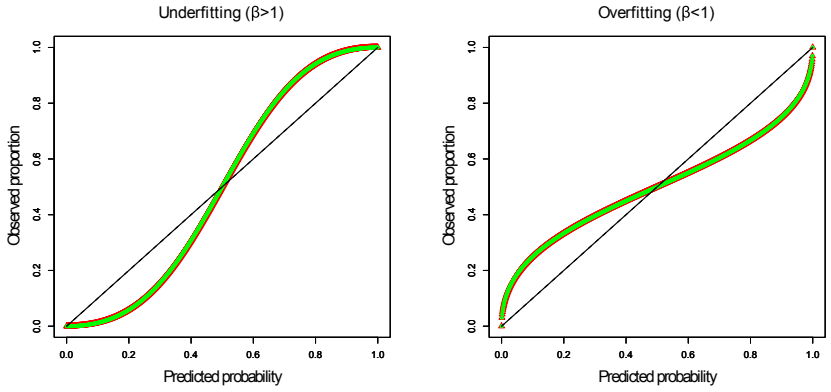
Calibration plot. The relation between predicted probabilities (x-axis) and observed probabilities (y-axis) can be visualized by a calibration plot [7, 63, 64, 128, 129, 167, 174]. Since the outcome is binary only zero and one values are available for the y-axis. The observed proportion of patients with the outcome can be estimated and will be plotted on the y-axis. Two types of calibration plots exist: parametric and non-parametric [174]. The former considers the logistic recalibration framework assuming a linear relationship of the linear predictor lp_1 with the outcome [36]. The more flexible non-parametric approach considers the model

$$\log\left(\frac{P(\mathrm{Y}=1)}{1-P(\mathrm{Y}=1)}\right)=\log\left(\frac{P(\mathrm{Y}=1)}{P(\mathrm{Y}=0)}\right)=\log\left(\frac{\pi_1}{\pi_0}\right)=\alpha'+\beta'\ s(\mathrm{lp}_1), \quad (2.10)$$

with $s(.)$ a loess or spline smoother [7, 39, 67, 167]. The calibration plot can indicate whether predictions are systematically too low or too high, but also too extreme or not extreme enough or a combination of both. In Figure 2.3, the different situations are shown for parametric calibration plots by plotting the probabilities of both outcome categories as introduction for the multinomial calibration plots that are presented later.

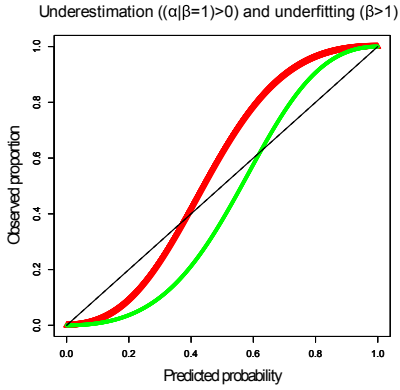


(a) Underestimation (b) Overestimation

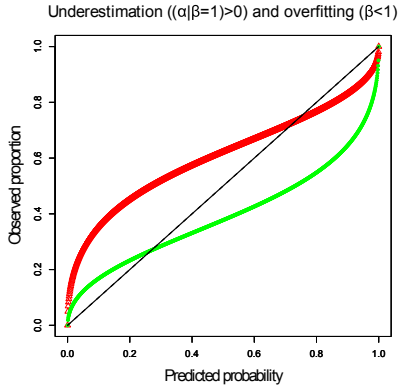


(c) Underfitting (d) Overfitting

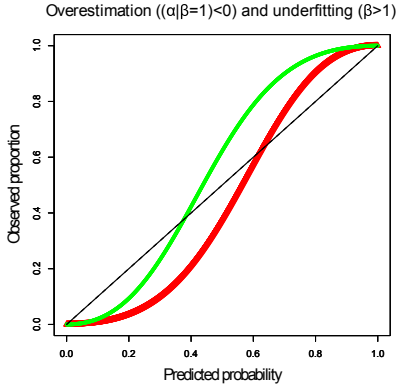
Figure 2.3: Parametric calibration plot showing predicted probabilities and observed proportions for both outcome categories in case of underestimation (a), overestimation (b), underfitting (c) and overfitting (d) or a combination of both (e-h). The red curve corresponds with Y= 1 and the green curve with Y= 0.



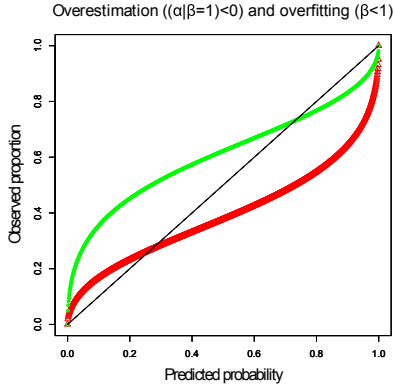
(e) Underestimation and underfitting



(f) Underestimation and overfitting



(g) Overestimation and underfitting



(h) Overestimation and overfitting

Continued Figure 2.3: Parametric calibration plot showing predicted probabilities and observed proportions for both outcome categories in case of underestimation (a), overestimation (b), underfitting (c) and overfitting (d) or a combination of both (e-h). The red curve corresponds with Y= 1 and the green curve with Y= 0.

2.3.2 Evaluation of classification

The performance of the prediction model can be assessed by considering the actual predicted risks but also by the classification. If the risk is above the threshold, the patients is classified as having the event ($Y = 1$), while a risk below the threshold is classified as not having the event ($Y = 0$). This cut-off or threshold can be determined in different ways. Below a data-driven and a decision-analytic approach are discussed [169].

Cut-off for classification

Data-driven. The ROC curve plots all pairs of sensitivity and specificity for all possible cut-offs. One approach to determine the cut-off is by maximizing the Youden's index [233]. This measure is defined as sensitivity+specificity-1 and is maximized in the upper left corner of the ROC curve. This point corresponds with the prevalence of the outcome on the condition that the model is well calibrated and the ROC curve is concave [73].

Decision-analytic. Another approach is to start from the clinical context and consider the implications of a true or a false classification [81, 110, 133, 151]. In order to determine the cut-off, the harm (H) of a false-positive is divided by the benefit (B) of a true-positive classification [174]. The first may result in overtreatment. Suppose that clinicians indicate that the harm of a FP (diagnosing and treating a benign tumor as malignant) is 9 times less important than finding an extra TP (diagnosing and treating a malignant tumor as malignant). This corresponds with a cut-off of 10% and a 1:9 H:B ratio. Giving the harm and the benefit an equal weight (i.e. a 1:1 H:B ratio) results in a cut-off of 50%.

Performance measures for classification

Some common statistical measures to evaluate the classification are accuracy, Youden's index and ROC with 1 cut-off. Alternative decision analytic measures are net benefit (NB), decision curves and relative utility.

Accuracy. The proportion of patients that are correctly classified corresponds with the accuracy. Misclassification error indicates the inverse (i.e. proportion of misclassified patients). This general measure can be misleading in case of unbalanced outcome categories [5]. Presume that our training sample has 95 % benign tumors and only 5% malignant tumors. Classifying all patients as having

a benign tumor will result in an accuracy of 0.95, while a model that tries to predict the malignant tumors might have a lower accuracy despite being a better model. This measure assumes equal misclassification costs for classifying a malignant tumor as benign and a benign tumor as malignant while this is not the case.

Youden's index. This single test statistic corresponds with sensitivity+specificity-1 [233]. An alternative would be to construct an ROC curve based on this sensitivity and specificity (i.e. ROC with one cut-off). The AUC then corresponds with the average of sensitivity and specificity.

Net Benefit (NB). The quality of classification can be expressed as the net benefit where the FP are weighted:

$$\text{NB} = \frac{(\text{TP} - w\ \text{FP})}{N} \tag{2.11}$$

with N the total number of patients [110, 144]. A weight of 1 corresponds with equally weighting FP and TP or a H:B ratio of 1:1. The weight w corresponds with the H:B ratio. Hence, a cut-off of 10% corresponds with a H:B ratio of 1:9 and $w = 0.11$.

Decision curves. Vickers and Elkin proposed to consider a range of cut-offs instead of one specific cut-off [222]. The NB for the different cut-offs can be plotted in a decision curve (Figure 2.4) [8]. These curves can be used to determine a cut-off to categorize patients with and without an outcome as well as weighing the harm of FP and TP [137, 174].

Relative utility (RU). Baker et al. introduced a measure that provides information on how much the obtained risk predictions contribute to clinical utility relative to perfect prediction [8, 9, 137]. The RU is the improvement in net benefit of the considered model, with a chosen threshold, with respect to the best baseline strategy ("treat all" or "treat none") divided by the net benefit of perfect prediction [200]. If the considered threshold is below the prevalence "treat all" is the best baseline strategy, while for a threshold larger than or equal to the prevalence the best baseline strategy is "treat none" [200]. A more detailed description of relative utility as well as its formula can be found in Baker et al. [8]. Similar as for NB, a plot of the RU for the different cut-offs can be made (Figure 2.5).

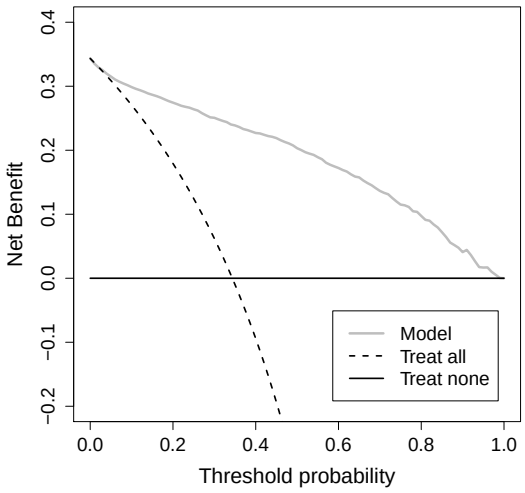


Figure 2.4: Decision curves. Solid gray line: prediction model; dashed black line: assume all patients have the outcome (i.e. treat all); solid black line: assume none of the patients have the outcome (i.e. treat none).

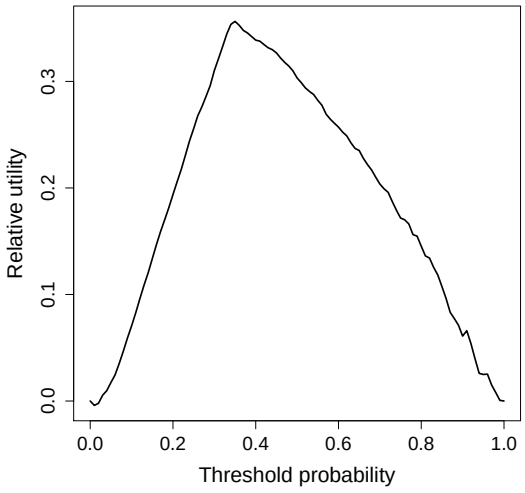


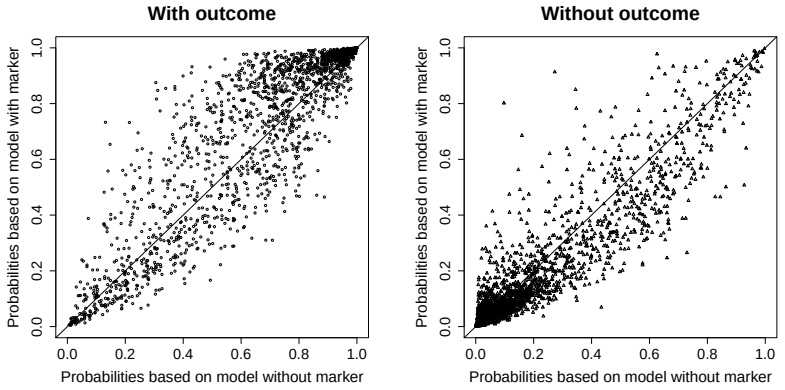
Figure 2.5: Relative utility curve.

2.3.3 Evaluation of model comparison

Besides evaluating the predictions or classifications on itself, it might be the purpose to investigate the incremental value of a new marker or the improvement of a new prediction model compared with an existing model [121]. One option is computing the difference in measures of both models (e.g. Δ AUC or Δ NB), possible alternative measures are reclassification, net reclassification improvement (NRI) and integrated discrimination improvement (IDI).

Reclassification. The incremental value of a new marker or prediction model can be expressed as the changes in classification using a new marker or a new prediction model [30, 121]. A reclassification table can be made indicating the changes in classification.

Net reclassification improvement. A remark concerning the reclassification table is that the changes need to be appropriate (i.e. higher risk for those with the outcome and lower risk for those without the outcome) [105, 121, 138, 144]. The NRI measures the improvement in reclassification. It computes the difference in proportion of individuals moving up minus those moving down for those with the outcome and the reverse for those without outcome (proportion moving down minus proportion moving up). The changes in predictions for those with and without the outcome can be visualized in a reclassification graph (Figure 2.6).



(a) (b)

Figure 2.6: Reclassification graph for cases (a) with and (b) without the outcome.

Integrated Discrimination Improvement. Finally, also the IDI can be considered which integrates the net reclassification index over all possible cut-offs [121,144]. This measure computes the increase in average predicted probability of the outcome for those with the outcome, as well as the decrease in the average predicted probability of the outcome for those without the outcome (Figure 2.7). IDI takes the sum of both in order to assess the improvement in discrimination of both models. The IDI corresponds with the difference in discrimination slopes. There are many interpretations for the IDI statistic [138,140,143,146].

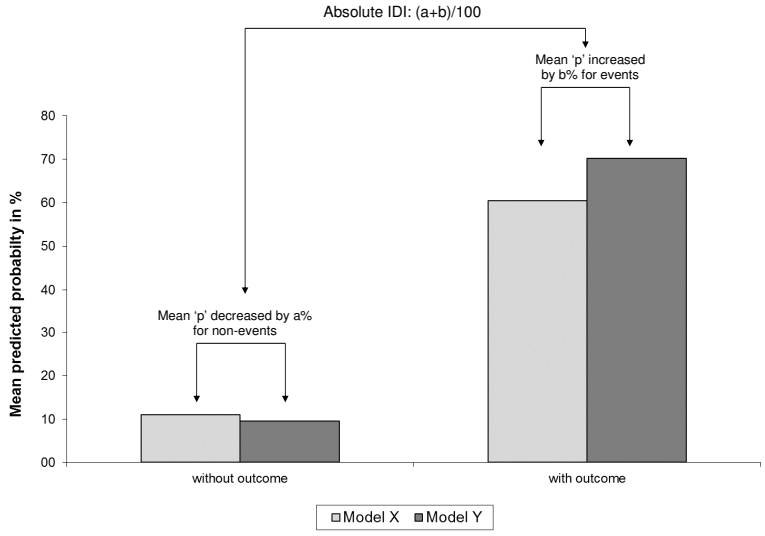


Figure 2.7: Integrated Discrimination Improvement (IDI).

Marker versus model perspective

For the evaluation of prediction models, a distinction can be made between quantifying the incremental value of an added marker ('marker perspective') and the comparison of the predicted risks of two prediction models to support clinical decision making ('model perspective') [104]. A reported drawback of the measures introduced above is that calibration of the prediction model influences the performance of these measures (i.e. non-proper measures) which might lead to misleading results [74,105,138,140,142,147]. Leening et al. indicate that the impact of miscalibration depends on the considered perspective [104]. For the marker perspective, two different situations are possible: (1) a model with and without the added marker may be developed on the same data and (2) the

added value of a marker on an existing model may be assessed. For the former we do not expect to see miscalibration, unless the considered dataset is not sufficiently large, while for the latter the existing model may be miscalibrated on the considered data resulting in spurious results [104]. In case of model perspective, comparing a new model with an existing model or comparing multiple existing models, the evaluation of model discrimination and calibration is crucial and the discussed measures should be considered as complementary measures [104].

2.4 Model updating

The external validity of an existing prediction model might be disappointing due to different reasons (e.g. differences in setting, differences in measurement methods, patient characteristics) [119]. Mostly this results in the development of a new model, while model updating might be a sensible alternative. The prediction model still contains useful information with respect to relevant predictors and predictor effects. It might even be more efficient to update the prediction model instead of developing a new prediction model. Several methods can be considered to update an existing dichotomous logistic regression model going from simple methods to more extensive methods (Table 2.1) [167, 168]. The formulae of the different considered methods are grouped in Table 2.2. Often simple methods such as recalibration are enough; more extensive methods should only be considered in case of sufficiently large validation samples, preferably in combination with shrinkage [66, 82, 83, 167, 168, 186].

2.4.1 Reference method

The simplest method, METHOD 0, is a reference method where the original prediction model without any adjustments is considered:

$$\log\left(\frac{\pi_1(\mathbf{x})}{1-\pi_1(\mathbf{x})}\right) = \log\left(\frac{\pi_1(\mathbf{x})}{\pi_0(\mathbf{x})}\right) = a + \sum_{i=1}^{n} b_i \mathrm{x}_i = a + b^T\mathbf{x} = \mathrm{lp}_{0,1} \quad (2.12)$$

with a and b the intercept and coefficients of the original prediction model. The vector $\mathbf{x}$ obtains the values of the predictors of the original prediction model in the new data set. The first subscript of the linear predictor $\mathrm{lp}_{0,1}$ indicates the number of method, while the second indicates that the risks of outcome $Y = 1$ are modeled.

Table 2.1: Updating methods. For methods with an asterisk (*) shrinkage towards METHOD 2 may be considered.

METHOD		Description
0	Reference	The original prediction model *without* any adjustments.
1	Recalibration	The original prediction with *intercept* adjustment.
2		The original prediction model with *intercept* and *slope* adjustments.
3(*)	Revision	METHOD 2 with testing whether the effect of *each individual predictor* is *significantly* different in the new data.
4(*)		The original prediction model with re-estimation of the *intercept* and the *coefficients of the original individual predictors.*
5(*)	Extension	METHOD 3 with testing whether the *effect of new individual predictors* is *significant* in the new data.
6(*)		METHOD 4 with testing whether the *effect of new individual predictors* is *significant* in the new data.
7(*)		The original prediction model with re-estimation of the *intercept*, the *coefficients* of the *original individual predictors* and the coefficients of the *new additional predictors.*
R	Redevelopment	Building a new model from scratch for the new data.

2.4.2 Recalibration methods

The first recalibration method (i.e. METHOD 1) consists of updating the intercept of the prediction model:

$$\log\left(\frac{\pi_1(\mathbf{x})}{\pi_0(\mathbf{x})}\right) = \alpha + \mathrm{lp}_{0,1} = \mathrm{lp}_{1,1}. \tag{2.13}$$

The original linear predictor $\mathrm{lp}_{0,1}$ is considered as offset and an intercept α is estimated. In this way, the calibration-in-the-large is corrected to make sure that the average predicted probability corresponds with the observed incidence. METHOD 2 allows for an additional correction in the slope which is a single overall adjustment factor to update the coefficients of the predictors by considering $\mathrm{lp}_{0,1}$ as predictor and estimating an intercept α and slope β:

$$\log\left(\frac{\pi_1(\mathbf{x})}{\pi_0(\mathbf{x})}\right) = \alpha + \beta\ \mathrm{lp}_{0,1} = \mathrm{lp}_{2,1}. \tag{2.14}$$

2.4.3 Revision methods

For revision the coefficients of the predictors are considered for re-estimation. METHOD 3 starts from METHOD 2 followed by likelihood ratio tests to assess whether the effect of each individual predictors needs to be updated:

$$\mathrm{lp}_{3,1} = \mathrm{lp}_{2,1} + \sum_{j=1}^{n} \delta_{j|p\leq 0.05}\ \mathrm{x}_j, \tag{2.15}$$

with $\delta_{j|p\leq 0.05}$ equal to zero if $p > 0.05$ for the likelihood ratio test of predictor x_j and otherwise the difference in effect for predictor x_j. Although 0.05 is commonly used [83, 167, 168], other p-values such as $p \leq 0.15$ or $p \leq 0.20$ are valid alternatives and might even be better [79]. A forward re-estimation procedure, starting with the predictor with the strongest difference, is followed until none of the tests is statistically significant. The last revision method is METHOD 4 and re-estimates the intercept and the coefficients of the individual predictor on the new data:

$$\mathrm{lp}_{4,1} = a + \sum_{j=1}^{n} b_j\ \mathrm{x}_j. \tag{2.16}$$

METHOD 3 corresponds with METHOD 2 in case of a significance level of zero (i.e. no coefficients re-estimated), and with METHOD 4 in case of a significance level of 1 (i.e. all coefficients re-estimated) for every likelihood ratio test instead of

0.05. The higher the significance level or p-value for METHOD 3 the smaller the difference between METHOD 3 and 4 will be. Therefore we considered 0.05 as significance level. For both methods shrinkage can be used to prevent overfitting. A possible shrinkage factor is the heuristic shrinkage factor [168, 210]:

$$c = \frac{(\text{model } \chi^2 - df)}{\text{model } \chi^2}. \tag{2.17}$$

Model χ^2 refers to the difference in -2 log-likelihood between the model with re-estimated parameters (i.e. METHOD 3 or METHOD 4) and the recalibrated model (i.e. METHOD 2) and df corresponds with the difference in degrees of freedom between those two models. The shrunken linear predictor $\text{lp}^s_{i,1}$ ($i = 3$ or 4) is obtained as follows:

$$\begin{aligned} \text{lp}^s_{i,1} &= \text{lp}_{2,1} + \widehat{c}\left[\text{lp}_{i,1} - \text{lp}_{2,1}\right] \\ &= (1-\widehat{c})\,\text{lp}_{2,1} + \widehat{c}\,\text{lp}_{i,1} \\ &= (1-\widehat{c})\left(\alpha + \beta \left[a + \sum_{j=1}^{n} b_j\ \text{x}_j\right]\right) + \widehat{c}\left[a' + \sum_{j=1}^{n} b'_j\ \text{x}_j\right] \end{aligned} \tag{2.18}$$

with $\widehat{c}$ the estimated heuristic shrinkage factor, a' and b' the individual regression coefficients of the considered method i (e.g. $b'_j = \beta_j + \delta_{j|p\leq 0.05}$ for METHOD 3). The individual regression coefficients are shrunken towards the recalibrated values [83, 168, 210]. Afterwards, the intercept can be re-estimated to correct for the calibration-in-the-large. There are also alternative shrinkage methods such as Lasso and Ridge [53]

2.4.4 Extension methods

The following three methods consider additional predictors. METHOD 5 is an extension of METHOD 3, while METHOD 6 is an extension of METHOD 4:

$$\text{lp}_{5,1} = \text{lp}_{2,1} + \sum_{j=1}^{n} \delta_{j|p\leq 0.05}\ \text{x}_j + \sum_{j=n+1}^{l} \gamma_{j|p\leq 0.05}\ \text{x}_j, \tag{2.19}$$

$$\text{lp}_{6,1} = a + \sum_{j=1}^{n} b_j\ \text{x}_j + \sum_{j=n+1}^{l} \gamma_{j|p\leq 0.05}\ \text{x}_j, \tag{2.20}$$

where $\text{x}_{n+1}, ..., \text{x}_l$ are additional predictors, $\delta_{j|p\leq 0.05}$ equals zero if $p > 0.05$ for the likelihood ratio test of predictor x_j and otherwise the difference in effect

Table 2.2: Updating methods: formulae. For methods with an asterisk (*) shrinkage towards METHOD 2 may be considered.

METHOD	Description
Reference	
0	$\text{lp}_{0,1} = a + \sum_{i=1}^{n} b_i \text{x}_i = a + b^T \mathbf{x}$
Recalibration	
1	$\text{lp}_{1,1} = \alpha + \text{lp}_{0,1}$
2	$\text{lp}_{21} = \alpha + \beta\ \text{lp}_{0,1}$
Re-estimation	
3(*)	$\text{lp}_{3,1} = \text{lp}_{2,1} + \sum_{j=1}^{n} \delta_{j\|p\leq 0.05}\ \text{x}_j$
4(*)	$\text{lp}_{4,1} = a + \sum_{j=1}^{n} b_j\ \text{x}_j$
Extension	
5(*)	$\text{lp}_{5,1} = \text{lp}_{2,1} + \sum_{j=1}^{n} \delta_{j\|p\leq 0.05}\ \text{x}_j + \sum_{j=n+1}^{l} \delta_{j\|p\leq 0.05}\ \text{x}_j$
6(*)	$\text{lp}_{6,1} = a + \sum_{j=1}^{n} b_j\ \text{x}_j + \sum_{j=n+1}^{l} \delta_{j\|p\leq 0.05}\ \text{x}_j$
7(*)	$\text{lp}_{7,1} = a + \sum_{j=1}^{l} b_j\ \text{x}_j$
Redevelopment	
R	$\text{lp}_{R,1} = a + \sum_{i=1}^{l} b_i \text{x}_i = a + b^T \mathbf{x}$

for predictor x_j and similarly $\gamma_{j|p\leq 0.05}$ equals zero if $p > 0.05$ for the likelihood ratio test of predictor x_j and otherwise the effect for predictor x_j. Analogously, also here another p-value such as $p \leq 0.15$ or $p \leq 0.20$ could be considered [79]. Further METHOD 7, also an extension of METHOD 4, estimates coefficients for all the original predictors and all the additional predictors:

$$\text{lp}_{7,1} = a + \sum_{j=1}^{l} b_j\ \text{x}_j. \tag{2.21}$$

Similar as for the revision methods, also for these methods shrinkage can be considered. The coefficients of the predictors of the original prediction model are shrunken towards their recalibrated values, while the coefficients of the new predictors are shrunken towards zero.

2.4.5 Redevelopment method

The final method (i.e. METHOD R) builds a completely new model. This may be in a similar way as the original prediction model (e.g. same variable selection procedure) but is not necessary.

2.5 Conclusion

This chapter describes the dichotomous logistic regression model. Different types of validation together with performance measures are discussed. A distinction can be made between measures to assess the performance of the predictions, the classifications based on these predictions and the improvement of markers or models. Different updating techniques for a dichotomous logistic regression model can be used going from simple recalibration methods towards more complex methods such as extension or redevelopment.

Chapter 3

Multinomial risk prediction models

*In the previous Chapter the dichotomous logistic regression model is introduced. This Chapter discusses approaches to model a multinomial outcome, i.e. an unordered outcome with $k\,(>2)$ categories. We compare the common multinomial logistic regression (*MLR*), Section 3.2.1, with approaches that obtain multinomial risk probabilities by combining different dichotomous logistic regression models. A first combination approach is the sequential dichotomous model or binary tree, i.e. a 'nested sequence' of dichotomous prediction models (Section 3.2.2). Afterwards, dichotomous models for each pair of outcome categories are combined via pairwise coupling (Section 3.2.3). Section 3.2.5 discusses 1-vs-rest dichotomous prediction models. The last considered approach is individualized logistic regression (*ILR*), where the different categories $j (j \neq i)$ are modeled with respect to a chosen reference category. This approach is very similar to the* MLR *approach, the difference is that* ILR *fits the k-1 models individually, while* MLR *fits them simultaneously. In Section 3.3, measures to assess the multinomial discriminative ability are introduced. Finally, in Section 3.4, the different proposed approaches are compared with respect to discrimination, using some of the methods discussed in Section 3.3, and calibration.*

3.1 Introduction

Risk prediction models for diagnostic or prognostic outcomes are useful tools for clinical decision support [120]. Most commonly, a dichotomous outcome is

considered. However, especially in diagnostic problems, a differential diagnosis often includes more levels than categorization of subjects as 'diseased' vs 'non-diseased'. In the case of ovarian tumor diagnosis the dichotomous problem consists of predicting malignancy of the tumor, even though optimal management does not only depend on this simplification. For borderline malignancies, often less invasive fertility-sparing methods are preferred, while for invasive tumors it is important to avoid rupture and residual disease and more invasive techniques are required since type and quality of surgery affect prognosis. [40, 40, 70, 193].

3.2 Approaches

A polytomous outcome can be modeled all-at-once or by combining different dichotomous prediction models. The different possible approaches will first be described and thereafter be clarified using a worked out example concerning the diagnosis of ovarian tumors as benign, borderline or invasive ($k = 3$). The considered dichotomous and multinomial prediction models are logistic regression models.

3.2.1 Multinomial logistic regression (MLR)

The most common approach for modeling a nominal outcome is multinomial logistic regression also known as baseline category logit model [2, 4, 79, 80] (Figure 3.1). This model is an extension of the dichotomous logistic regression model towards a nominal outcome. Suppose, outcome category k is considered as reference category R. The k-1 submodels comparing each category i ($i \neq k$) with the reference category are:

$$\begin{cases} \log\left(\dfrac{\pi_1}{\pi_k}\right) = a_1 + b_1^T\mathbf{x} = \mathrm{lp}_{1k} \\ \log\left(\dfrac{\pi_2}{\pi_k}\right) = a_2 + b_2^T\mathbf{x} = \mathrm{lp}_{2k} \\ \dots \\ \log\left(\dfrac{\pi_{k-1}}{\pi_k}\right) = a_{k-1} + b_{k-1}^T\mathbf{x} = \mathrm{lp}_{k-1k} \end{cases} \tag{3.1}$$

The choice of reference category (i.c. $i = k$) is important for the interpretation of the regression coefficients b_i but does not affect the predicted probabilities.

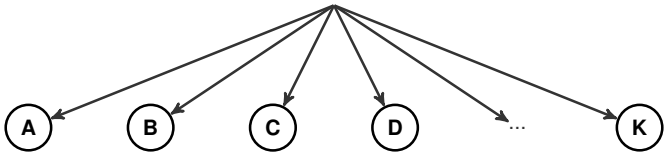


Figure 3.1: Multinomial logistic regression model in case of an outcome with k categories (i.e. A-K).

We denote the linear predictor for category i versus category k as lp_{ik}. Similarly as for dichotomous LR the estimates of the model parameters a_i and b_i^T $(i = 1, ..., k-1)$ are obtained by maximizing the log-likelihood which is in this case a polytomous log-likelihood [79]. The polytomous probability of category i is

$$\widehat{\pi}_i = \frac{e^{\text{lp}_{ik}}}{\sum_{j=1}^{k} e^{\text{lp}_{jk}}}, \tag{3.2}$$

where the linear predictor of the reference category lp_{kk} equals zero.

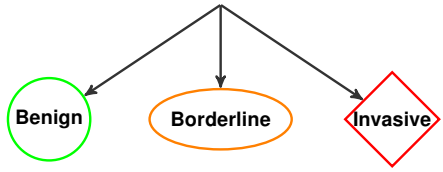
Multinomial logistic regression: worked out example

For the diagnosis of ovarian tumors as benign (Be), borderline (Bo) and invasive (In), the two dichotomous submodels of MLRare:

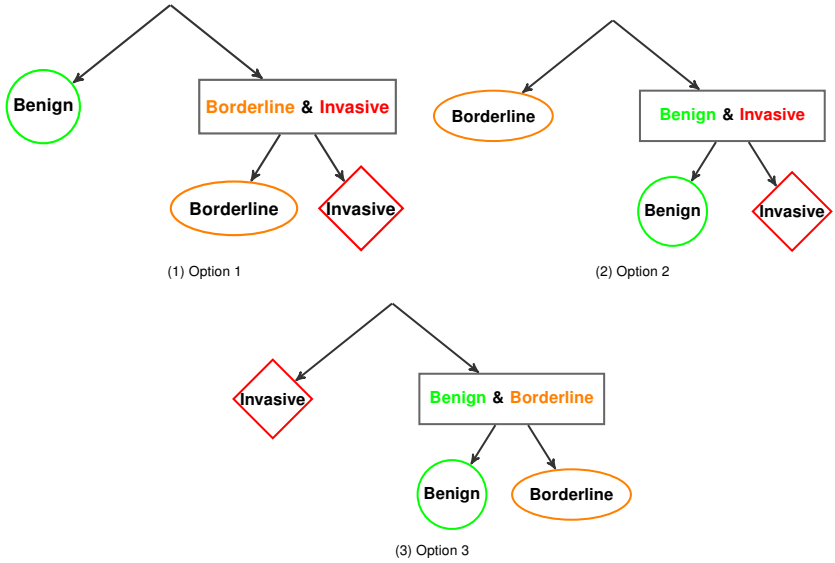
$$\begin{cases} \log\left(\dfrac{\pi_{Bo}}{\pi_{Be}}\right) &= (a_{Bo} + b_{Bo}^T \mathbf{x}) = \text{lp}_{BoBe}, \\ \log\left(\dfrac{\pi_{In}}{\pi_{Be}}\right) &= (a_{In} + b_{In}^T \mathbf{x}) = \text{lp}_{InBe}. \end{cases} \tag{3.3}$$

These submodels are fitted simultaneously. The polytomous predictions are:

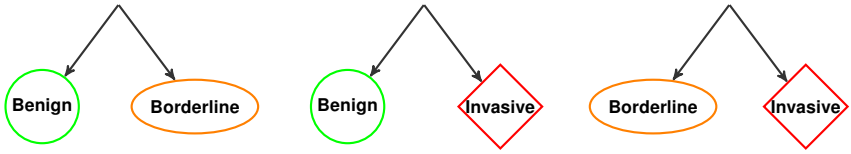
$$\begin{cases} \widehat{\pi}_{Be} = \dfrac{1}{1 + e^{\text{lp}_{BoBe}} + e^{\text{lp}_{InBe}}}, \\ \widehat{\pi}_{Bo} = \dfrac{e^{\text{lp}_{BoBe}}}{1 + e^{\text{lp}_{BoBe}} + e^{\text{lp}_{InBe}}}, \\ \widehat{\pi}_{In} = \dfrac{e^{\text{lp}_{InBe}}}{1 + e^{\text{lp}_{BoBe}} + e^{\text{lp}_{InBe}}}. \end{cases} \tag{3.4}$$



(a) Multinomial logistic regression

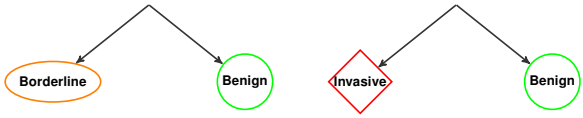


(b) Sequential dichotomous modeling

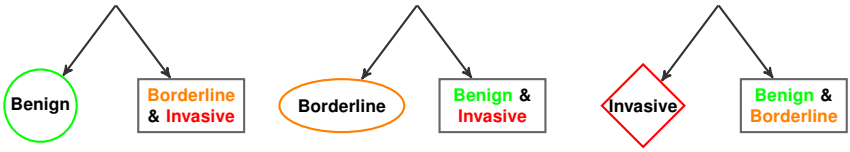


(c) Pairwise coupling

Figure 3.2: Schematic overview of the various approaches for the case study of ovarian tumors with outcomes benign, borderline and invasive: (a) multinomial LR model, (b) the three different possible sequential dichotomous models, (c) pairwise coupling LR model, (d) individualized LR model with benign as reference category and (e) 1-vs-rest LR model.



(d) Individualized logistic regression



(e) 1-vs-rest

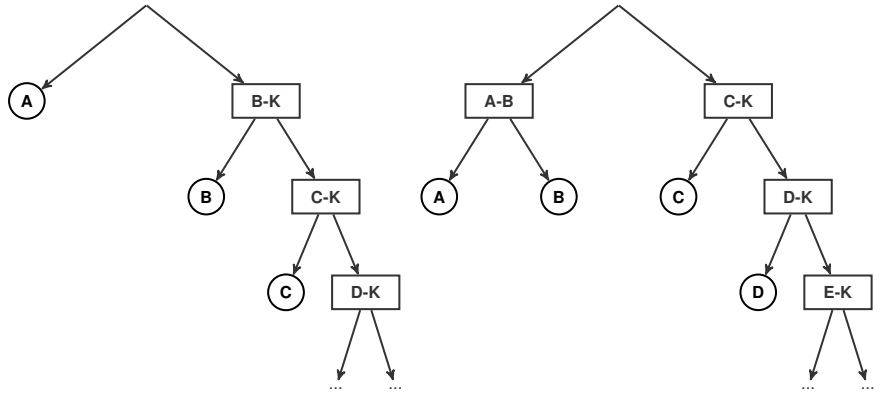
Continued Figure 3.2: Schematic overview of the various approaches for the case study of ovarian tumors with outcomes benign, borderline and invasive: (a) multinomial LR model, (b) the three different possible sequential dichotomous models, (c) pairwise coupling LR model, (d) individualized LR model with benign as reference category and (e) 1-vs-rest LR model.

3.2.2 Sequential dichotomous model/binary tree (BT)

The first combination approach, a sequential dichotomous model (also labeled binary tree, consecutive dichotomous modeling, nested dichotomies) consists of k-1 dichotomous prediction models in case of an outcome with k categories [45, 52, 154]. The categories can be combined in different ways (Figure 3.3). Clinical judgment, based on differences between categories concerning treatment consequences and clinical profile, can be used to decide about which categories should be combined. An alternative might be to average the result of all possible sequences although this number rapidly increases. In general for an outcome with k categories, $(2k-3)!!$ different sequences are possible. This double factorial corresponds with the product of all odd integers from 1 to $2k-3$:

$$\prod_{i=1}^{k}(2i-3). \tag{3.5}$$

E.g. for an outcome with 3 categories $3 \cdot 1 = 3$ different trees can be constructed (Figure 3.2b), while an outcome with 6 categories results in $9 \cdot 7 \cdot 5 \cdot 3 \cdot 1 = 945$ possible different trees.



(a) (b)

Figure 3.3: Some examples of different possible binary trees in case of an outcome with k categories (i.e. A-K).

Sequential dichotomous model: worked out example

For our example of ovarian tumor diagnosis, three possible sequences can be constructed (Figure 3.2b). The clinically most relevant tree distinguishes first benign and malignant tumors (i.e. borderline and invasive) and second borderline and invasive tumors (i.e. Option 1 in Figure 3.2b):

$$\log\left(\frac{\pi_{Be}}{\pi_{Bo}+\pi_{In}}\right) = \log\left(\frac{\pi_{Be}}{1-\pi_{Be}}\right) = (a_{Be} + b_{Be}^T\mathbf{x}) = \mathrm{z}_{Be},$$

$$\log\left(\frac{\pi_{Bo}}{\pi_{In}}\right) = (a_{BoIn} + b_{BoIn}^T\mathbf{x}) = \mathrm{z}_{BoIn}, \tag{3.6}$$

with π_{Be}, π_{Bo} and π_{In} the risk probabilities of having a benign, borderline or invasive tumor. The estimated probabilities for benign are obtained via the first model:

$$\widehat{\pi}_{Be} = \frac{e^{\mathrm{z}_{Be}}}{1+e^{\mathrm{z}_{Be}}}, \tag{3.7}$$

while the probabilities for borderline and invasive are obtained by multiplying the risk for a malignant tumor from the first model with the risk for a borderline

or invasive tumor from the second model:

$$\widehat{\pi}_{Bo} = \frac{1}{1+e^{\mathrm{z}_{BeBo}}} \cdot \frac{e^{\mathrm{z}_{BoIn}}}{1+e^{\mathrm{z}_{BoIn}}},$$

$$\widehat{\pi}_{In} = \frac{1}{1+e^{\mathrm{z}_{BeBo}}} \cdot \frac{1}{1+e^{\mathrm{z}_{BoIn}}}. \tag{3.8}$$

3.2.3 Pairwise coupling (PC)

A second possible approach is a pairwise coupling model where all pairs (i, j) are modeled against each other resulting in probabilities $\widehat{\pi_{ij}}$ [68]. The k polytomous probabilities are obtained by solving:

$$\pi_i = \sum_{j=1, j\neq i}^{k} \frac{\pi_i + \pi_j}{k-1} \widehat{\pi}_{ij} \quad \forall j \text{ with } \sum_{i=1}^{k} \pi_i = 1, \pi_i \geq 0 \tag{3.9}$$

This is a pairwise coupling approach of Wu et al. [224].

Pairwise coupling: worked out example

For our example, pairwise coupling consists of 3 dichotomous prediction models (Figure 3.2c):

$$\log\left(\frac{\pi_{Be}}{\pi_{Bo}}\right) = (a_{BeBo} + b_{BeBo}^T \mathbf{x}) = \mathrm{z}_{BeBo}, \tag{3.10}$$

$$\log\left(\frac{\pi_{Be}}{\pi_{In}}\right) = (a_{BeIn} + b_{BeIn}^T \mathbf{x}) = \mathrm{z}_{BeIn},$$

$$\log\left(\frac{\pi_{Bo}}{\pi_{In}}\right) = (a_{BoIn} + b_{BoIn}^T \mathbf{x}) = \mathrm{z}_{BoIn}.$$

The pairwise probabilities $\widehat{\pi}_i = e^{\mathrm{z}_i}/(1+e^{\mathrm{z}_i})$ $(i = BeBo, BeIn, BoIn)$ are combined using Wu et al. to obtain the polytomous probabilities for benign, borderline and invasive [224].

3.2.4 Individualized logistic regression (ILR)

A third option is individualized logistic regression [13]. This approach consists of developing k-1 dichotomous prediction models of each category i with respect to a chosen reference category. This reference category can be chosen based on the prevalence.

Individualized logistic regression: worked out example

The ILR model consists of 2 dichotomous prediction models (Figure 3.2d). For our example we consider benign tumors, with the largest prevalence, as reference category. The formulae of the logistic regression models are:

$$\log\left(\frac{\pi_{Bo}}{\pi_{Be}}\right) = (a_{BoBe} + b_{BoBe}^T \mathbf{x}) = \mathrm{z}_{BoBe}, \tag{3.11}$$

$$\log\left(\frac{\pi_{In}}{\pi_{Be}}\right) = (a_{InBe} + b_{InBe}^T \mathbf{x}) = \mathrm{z}_{InBe}.$$

Note that z_{BoBe} and z_{InBe} correspond with $-\mathrm{z}_{BeBo}$ and $-\mathrm{z}_{BeIn}$ of Equation 3.10. Equation 3.11 can be rewritten as $\pi_{Bo} = \pi_{Be} \cdot e^{\mathrm{z}_{BoBe}}$ and $\pi_{In} = \pi_{Be} \cdot e^{\mathrm{z}_{InBe}}$. Furthermore, these two submodels correspond with the models fitted for multinomial logistic regression. The difference between both approaches is that for MLR the submodels are fitted simultaneously, while for ILR the models are fitted separately. Since the probabilities of the tree outcome categories sum to 1, the polytomous probabilities are:

$$\begin{cases} \widehat{\pi}_{Be} &= \dfrac{1}{1 + e^{\mathrm{z}_{BoBe}} + e^{\mathrm{z}_{InBe}}}, \\ \widehat{\pi}_{Bo} &= \dfrac{e^{\mathrm{z}_{BoBe}}}{1 + e^{\mathrm{z}_{BoBe}} + e^{\mathrm{z}_{InBe}}}, \\ \widehat{\pi}_{In} &= \dfrac{e^{\mathrm{z}_{InBe}}}{1 + e^{\mathrm{z}_{BoBe}} + e^{\mathrm{z}_{InBe}}}. \end{cases} \tag{3.12}$$

The difference between MLR (Section 3.2.1) and ILR is that the former fits the k-1 submodels simultaneously, while the latter fits them independently.

3.2.5 1-vs-rest (1-V-R)

The fourth and last approach that combines dichotomous logistic regression models is 1-vs-rest logistic regression. Each outcome category is compared with the remaining categories, resulting in k dichotomous LR models [14, 152]. The obtained probabilities for the different outcome categories do not sum up to one, due to individual model building, and therefore need a normalization.

1-vs-rest: worked out example

For the 1-vs-rest logistic regression model, three dichotomous prediction models are built (Figure 3.2e):

$$\begin{aligned}
\log\left(\frac{\pi_{Be}}{\pi_{Bo}+\pi_{In}}\right) &= \log\left(\frac{\pi_{Be}}{1-\pi_{Be}}\right) = (a_{Be}+b_{Be}^T\mathbf{x}) = \mathrm{z}_{Be},\\
\log\left(\frac{\pi_{Bo}}{\pi_{Be}+\pi_{In}}\right) &= \log\left(\frac{\pi_{Bo}}{1-\pi_{Bo}}\right) = (a_{Bo}+b_{Bo}^T\mathbf{x}) = \mathrm{z}_{Bo},\\
\log\left(\frac{\pi_{In}}{\pi_{Be}+\pi_{Bo}}\right) &= \log\left(\frac{\pi_{In}}{1-\pi_{In}}\right) = (a_{In}+b_{In}^T\mathbf{x}) = \mathrm{z}_{In}.
\end{aligned} \tag{3.13}$$

The standardized polytomous probabilities are obtained by dividing 1-vs-rest probabilities for benign, borderline and invasive (i.e. $\widehat{\pi}_i = e^{\mathrm{z}_i}/(1+e^{\mathrm{z}_i})$ with $i = BeRest, BoRest, InRest$) by the sum of all three probabilities:

$$\begin{cases}
\widehat{\pi}_{Be} &= \dfrac{\widehat{\pi}_{BeRest}}{\widehat{\pi}_{BeRest}+\widehat{\pi}_{BoRest}+\widehat{\pi}_{InRest}},\\
\widehat{\pi}_{Bo} &= \dfrac{\widehat{\pi}_{BoRest}}{\widehat{\pi}_{BeRest}+\widehat{\pi}_{BoRest}+\widehat{\pi}_{InRest}},\\
\widehat{\pi}_{In} &= \dfrac{\widehat{\pi}_{InRest}}{\widehat{\pi}_{BeRest}+\widehat{\pi}_{BoRest}+\widehat{\pi}_{InRest}}.
\end{cases} \tag{3.14}$$

3.3 Multinomial discrimination

For the assessment of the discriminative ability of a multinomial risk prediction model either a set of dichotomous measures or one overall polytomous measure can be used [199]. Similarly as for a dichotomous outcome, the discrimination can be visualized by means of a discrimination box plot. In this Section we discuss multinomial discrimination measures that will be used in the remainder of this dissertation. The topic of multinomial calibration is covered in detail in Chapter 4 since multinomial calibration tools were not yet suggested in the literature, with the exception of suggestions for a multinomial Hosmer-Lemeshow variant [56, 138, 148].

3.3.1 A set of measures

In case of a multinomial outcome a 1-vs-rest approach or a pairwise approach can be considered in order to obtain dichotomous c-statistics.

1-vs-rest c-statistics. The outcome for each category i can be dichotomized, the 1-vs-rest c-statistic is obtained by evaluating the multinomial risk of category i versus the risk of the remaining categories. The latter is the sum of the multinomial risks of the remaining categories or 1 minus the multinomial risk of category i.

Pairwise c-statistics. An alternative is evaluating pairs of outcome categories. For the pairwise c-statistic of category i versus j, only cases of category i or j are considered. Since the outcome has more than two categories, the risk of category i and j do not sum up to one. There are different ways to compute pairwise c-statistics. A first option is averaging the c-statistic c_{ij} obtained with π_i and the c-statistic c_{ji} obtained with π_j [61]. A second option is computing the c-statistic of $\pi_i - \pi_j$ [131]. A third option is conditional pairwise c-statistics [199]. For these c-statistics the conditional risk is considered (i.e. $\pi_i/(\pi_i + \pi_j)$). The pairwise c-statistics may reveal poor or excellent discrimination between pairs of categories that are hidden in the 1-vs-rest c-statistics. In this dissertation, we make use of the conditional pairwise c-statistics.

3.3.2 One overall measure

We will discuss two overall multinomial discrimination measures. First, the M-index by Hand and Till, a common approach in machine learning, is considered [61]. Second, the polytomous discrimination index (PDI) by Van Calster et al. is discussed [196]. Other overall measures exist such as volume under the surface by Mossman [125], an averaged 1-vs-rest c-statistics mentioned by Ferri and colleagues [49] and Provost and Domingos [150] and a measure similar to the M-index proposed by Obuchowski and colleagues [131].

M-*index.* The pairwise c-statistics can be combined to an overall discrimination measure. The M-index uses the average of the pairwise c-statistics c_{ij} and c_{ji} (i.e. $c_{(ij)}$) and is obtained by:

$$\text{M-index} = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} c_{(ij)}. \tag{3.15}$$

This measure indicates the probability of correctly distinguishing two cases from two randomly selected categories. Random and perfect discrimination

corresponds with an M-index of respectively 0.5 and 1.

*Polytomous discrimination index (*PDI*).* A polytomous extension of the c statistic, assessing a set of k cases each belonging to a different category is the PDI. The PDI counts for how many outcome categories the predicted probability is highest for the case belonging to that category. For the formula of PDI some notations need to be introduced: N_i indicates the number of cases in outcome category i and $\boldsymbol{\pi}_{ni} = [\pi_{1,ni} \ ... \ \pi_{k,ni}]^T$ the vector with predicted risks of the nth case from category i. The PDI can be computed as follows:

$$\text{PDI} = \frac{1}{kN_1 \cdots N_k} \sum_{n_1=1}^{N_1} \cdots \sum_{n_k=1}^{N_k} \left[\sum_{i=1}^{k} C_i(\boldsymbol{\pi}_{n_1 1}, \cdots, \boldsymbol{\pi}_{n_k k}) \right]. \tag{3.16}$$

where $C_i(\boldsymbol{\pi}_{n_1 1}, \cdots, \boldsymbol{\pi}_{n_k k})$ equals 1 if $\pi_{i,n_i i} > \pi_{i,n_j j}$ for all $j \neq i$ and zero otherwise. The interpretation is that a correct classification of a set $n_1 1, \cdots, n_k k$ (i.e. n_1th patient of category 1, ..., n_kth patient of category k) occurs when the probability of outcome category i is the highest for the case with outcome i. The PDI can be interpreted as the probability to correctly identify the case from a randomly selected outcome category from a group of cases who all belong to a different category or the average proportion correctly classified cases in a set of k cases [196]. The PDI corresponds to $1/k$ in case of random performance and 1 for perfect performance [196].

3.3.3 Visualization

The discriminative potential of a multinomial risk prediction model can be visualized in a discrimination plot (Figure 3.4) [199]. The predicted probability for each outcome category is given stratified by the actual observed outcome category. Additionally to these boxplots the actual observed prevalence or baseline probabilities for each outcome group can be indicated.

3.4 Application: comparing different methods to develop multinomial risk prediction models

The various approaches to develop a multinomial risk prediction model based on logistic regression are compared with respect to discrimination and calibration on a case study of ovarian tumors. The IOTA data from all phases were considered

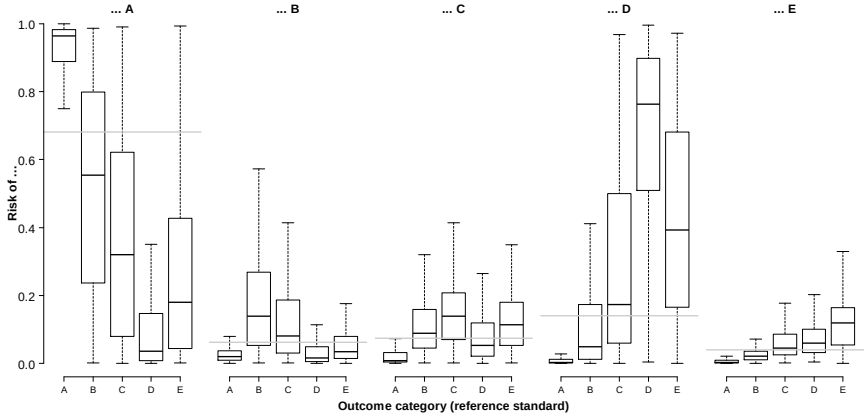


Figure 3.4: Polytomous discrimination plot: for each observed outcome category, boxplots of the predicted probabilities for the k outcome categories (e.g. k=5 and outcome categories A-E) are given. In total $k \cdot k$ boxplots are shown. The gray horizontal lines may indicate the prevalence or baseline probability of each outcome category.

(see Section 1.4.1)[1]. The final diagnosis was either benign, borderline or invasive. 3981 tumors of our considered data were benign (67%), 339 borderline (6%) and 1592 invasive (27%). Two clinical variables - patient age and current use of hormonal therapy - and nine ultrasound variables - ascites, bilaterality of the mass, maximum diameter of the lesion and of the solid component, number of papillary structures, height of the largest papillary structure, presence of acoustic shadows, presence of ten or more locules and irregular internal cyst wall - were used (see Section 1.4.1 and Table 1.3).

The risk prediction models were derived on development data and evaluated on separate validation data. We considered the sequential dichotomous logistic regression model with first benign vs rest (BTbe), and the average of all three possible sequences (BTav), pairwise coupling logistic regression (PC), individualized logistic regression (ILR), 1-vs-rest logistic regression (1-V-R) and finally multinomial logistic regression (MLR). The data were randomly split into development (N=3937) and validation data (N=1975). The split was repeated 300 times and average results were reported. Different subsets of the development data (i.e. EPV 5, 10 or 20), without altering the size of the validation data, were considered in order to assess the influence of sample size on the results. Since

[1]From the 5914 included patients in the IOTA data, two tumors were removed due to further data review.

the borderline tumors form the smallest group, an EPV of 5 means that for each considered predictor there should be 5 borderline tumors in the database. The considered sample sizes of the development data are therefore respectively 958, 1916 and 3833. Furthermore, for each dichotomous and multinomial logistic regression model a full model and a model with backward variable selection ($\alpha = 0.05$) were developed to investigate the effect of variable selection on the different approaches.

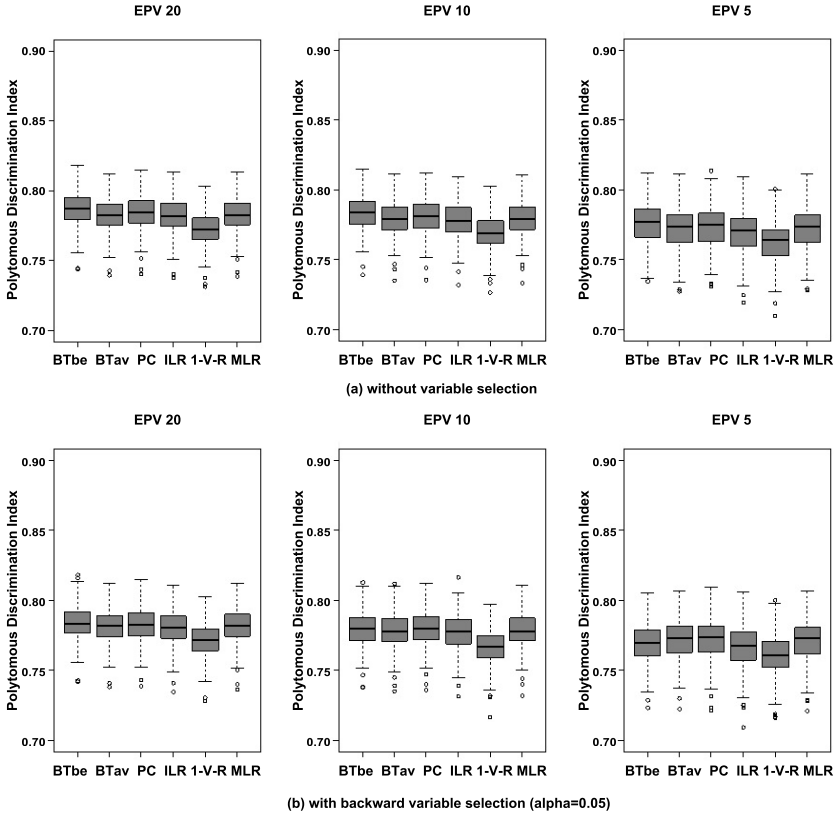
For the assessment of discriminative ability (i.e. how well the categories can be distinguished) the polytomous discrimination index (PDI) is calculated [196]. In case of 3 outcome categories the PDI ranges between 0.33 (1/3) and 1. The differences between observed proportions and predicted risks (i.e. calibration) are captured in the calibration-in-the-large.

Results

Discrimination. One of the main findings concerning discrimination is that 1-vs-rest LR had a poorer performance compared to the other approaches (Figure 3.5 and Table 3.1) [199]. The reason may be that pooling categories into a rest group obscures relevant differences between individual categories. For the same reason, the performance of sequential dichotomous modeling depends on the chosen sequence. The clinically most meaningful tree, with first benign vs rest and followed by borderline vs invasive, resulted in a higher discrimination than borderline vs rest followed by benign vs invasive (shown for EPV=20 in Table 3.2). Comparing the discrimination of dichotomous prediction models showed that the model with the lowest c-statistic was borderline vs rest (i.e. benign and invasive) (Table 3.3). Apparent performance was overoptimistic when EPV=5 and with EPV=20 optimism was notably smaller yet not negligible (Table 3.1). Lower EPV also resulted in larger variability.

Table 3.1: Average PDI with 95% confidence intervals for the different approaches, without and with backward variable selection, on the development (Dev) and validation (Val) data.

Variable selection	EPV	BTbe		BTav	
		Dev	Val	Dev	Val
Without	20	0.797 [0.783; 0.809]	0.787 [0.766; 0.810]	0.792 [0.779; 0.804]	0.783 [0.761; 0.806]
	10	0.801 [0.774; 0.822]	0.783 [0.761; 0.806]	0.797 [0.770; 0.819]	0.779 [0.755; 0.803]
	5	0.809 [0.768; 0.850]	0.776 [0.745; 0.803]	0.806 [0.763; 0.846]	0.772 [0.743; 0.800]
Backward	20	0.794 [0.781; 0.806]	0.784 [0.763; 0.807]	0.791 [0.778; 0.803]	0.782 [0.759; 0.805]
	10	0.796 [0.768; 0.817]	0.779 [0.756; 0.803]	0.795 [0.768; 0.817]	0.778 [0.754; 0.801]
	5	0.798 [0.754; 0.840]	0.769 [0.742; 0.799]	0.799 [0.759; 0.839]	0.772 [0.743; 0.799]



Abbreviations: BTbe, sequential dichotomous LR model or binary tree with first split benign vs malignant; BTav, average over all possible three sequential dichotomous LR models; PC, pairwise coupling LR model; ILR, individualized LR model; 1-V-R, 1-vs-rest LR model; MLR, multinomial LR model.

Figure 3.5: Boxplots of polytomous discrimination index (PDI) of the various approaches for different events per variable (i.e. EPV 5, 10 and 20) (a) without and (b) with variable selection for the validation data.

Continued Table 3.1: Average PDI with 95% confidence intervals for the different approaches, without and with backward variable selection, on the development and validation data.

Variable selection	EPV	PC		ILR	
		Dev	Val	Dev	Val
Without	20	0.794 [0.781; 0.807]	0.785 [0.763; 0.807]	0.791 [0.779; 0.804]	0.782 [0.760; 0.806]
	10	0.799 [0.772; 0.821]	0.781 [0.758; 0.804]	0.795 [0.768; 0.817]	0.778 [0.753; 0.802]
	5	0.808 [0.765; 0.849]	0.773 [0.742; 0.801]	0.800 [0.758; 0.841]	0.769 [0.739; 0.800]
Backward	20	0.793 [0.780; 0.805]	0.783 [0.761; 0.807]	0.790 [0.777; 0.801]	0.781 [0.758; 0.805]
	10	0.796 [0.769; 0.818]	0.780 [0.756; 0.802]	0.792 [0.763; 0.815]	0.777 [0.752; 0.801]
	5	0.799 [0.757; 0.839]	0.772 [0.741; 0.796]	0.791 [0.748; 0.831]	0.767 [0.733; 0.794]
Variable selection	EPV	1-V-R		MLR	
		Dev	Val	Dev	Val
Without	20	0.782 [0.769; 0.795]	0.773 [0.750; 0.797]	0.792 [0.779; 0.804]	0.783 [0.761; 0.807]
	10	0.787 [0.760; 0.811]	0.770 [0.746; 0.794]	0.797 [0.770; 0.819]	0.779 [0.755; 0.802]
	5	0.796 [0.755; 0.833]	0.762 [0.733; 0.789]	0.805 [0.763; 0.846]	0.772 [0.743; 0.799]
Backward	20	0.782 [0.769; 0.794]	0.772 [0.749; 0.796]	0.791 [0.778; 0.804]	0.782 [0.760; 0.805]
	10	0.785 [0.757; 0.808]	0.767 [0.742; 0.791]	0.795 [0.768; 0.818]	0.779 [0.754; 0.801]
	5	0.788 [0.747; 0.830]	0.761 [0.730; 0.789]	0.801 [0.759; 0.841]	0.771 [0.742; 0.797]

Abbreviations: BTbe, sequential dichotomous LR model or binary tree with first split benign vs malignant; BTav, average over all possible three sequential dichotomous LR models; PC, pairwise coupling LR model; ILR, individualized LR model; 1-V-R, 1-vs-rest LR model; MLR, multinomial LR model.

Table 3.2: Average PDI with 95% confidence intervals for the three different sequential dichotomous models or binary trees in case of an EPV of 20.

	Sequential dichotomous model		
Fig 3.3	First split	Development	Validation
Option 1	Benign vs rest	0.797 [0.783;0.808]	0.787 [0.766;0.810]
Option 2	Borderline vs rest	0.779 [0.766;0.792]	0.770 [0.747;0.793]
Option 3	Invasive vs rest	0.792 [0.778;0.803]	0.782 [0.760;0.806]

Table 3.3: Average c-statistics with 95% confidence intervals for the dichotomous risk prediction models on the validation data.

(a) 1-vs-1 dichotomous prediction models

Variable selection	EPV	Be vs Bo	Be vs In	Bo vs In
Without	20	0.875 [0.844;0.902]	0.955 [0.949;0.962]	0.866 [0.838;0.895]
	10	0.870 [0.841;0.900]	0.955 [0.948;0.961]	0.860 [0.828;0.890]
	5	0.860 [0.928;0.954]	0.953 [0.779;0.882]	0.847 [0.937;0.960]
Backward	20	0.873 [0.842;0.901]	0.955 [0.949;0.962]	0.863 [0.836;0.891]
	10	0.868 [0.835;0.899]	0.955 [0.948;0.961]	0.858 [0.828;0.887]
	5	0.854 [0.798;0.890]	0.952 [0.945;0.960]	0.848 [0.812;0.880]

Abbreviations: Be, benign; Bo, borderline; In, invasive.

Continued Table 3.3: Average c-statistics with 95% confidence intervals for the dichotomous risk prediction models on the validation data.

(b) 1-vs-rest dichotomous prediction models

Variable selection	EPV	Be vs rest	Bo vs rest	In vs rest
Without	20	0.938 [0.929;0.945]	0.806 [0.769;0.843]	0.945 [0.937;0.952]
	10	0.937 [0.928;0.944]	0.799 [0.765;0.836]	0.944 [0.936;0.951]
	5	0.935 [0.925;0.943]	0.787 [0.725;0.829]	0.942 [0.935;0.950]
Backward	20	0.937 [0.929;0.945]	0.803 [0.769;0.840]	0.944 [0.937;0.952]
	10	0.937 [0.929;0.944]	0.787 [0.742;0.828]	0.943 [0.935;0.951]
	5	0.934 [0.925;0.942]	0.765 [0.707;0.810]	0.941 [0.935;0.949]

Abbreviations: Be, benign; Bo, borderline; In, invasive.

Calibration. The different considered approaches lead to similar calibration-in-the-large values (Table 3.4). Lower EPV resulted in larger variability of calibration-in-the-large.

Table 3.4: Average calibration-in-the-large for benign (Be), borderline (Bo) and invasive (In) in % with 95% confidence intervals for the validation data.

	EPV		BTbe	BTav	PC	ILR	1-v-R	MLR
Without variable selection	20	Be	-0.1 [-1.4;1.2]	-0.2 [-1.4;1.1]	-0.1 [-1.4;1.2]	-0.1 [-1.5;1.2]	-0.1 [-1.4;1.2]	-0.1 [-1.4;1.2]
		Bo	-0.1 [-0.4;0.4]	0.0 [-0.3;0.5]	-0.0 [-0.4;0.4]	0.1 [-0.32;0.51]	0.0 [-0.3;0.4]	0.0 [-0.4;0.5]
		In	0.1 [-1.1;1.3]	0.1 [-1.2;1.3]	0.2 [-1.1;1.3]	0.0 [-1.2;1.2]	0.1 [-1.2;1.3]	0.0 [-1.2;1.2]
	10	Be	0.0 [-1.3;1.5]	-0.1 [-1.4;1.4]	-0.0 [-1.3;1.4]	-0.0 [-1.4;1.4]	-0.0 [-1.4;1.4]	0.0 [-1.3;1.5]
		Bo	-0.1 [0.6;0.5]	-0.0 [-0.5;0.5]	-0.1 [-0.6;0.5]	0.0 [-0.5;0.5]	-0.0 [-0.4;0.5]	-0.0 [-0.5;0.5]
		In	0.1 [-1.3;1.5]	0.1 [-1.3;1.4]	0.1 [-1.3;1.5]	0.0 [-1.3;1.4]	0.0 [-1.3;1.3]	-0.0 [-1.4;1.3]
	5	Be	0.0 [-1.7;1.7]	-0.0 [-1.8;1.7]	0.0 [-1.8;1.7]	0.0 [-1.8;1.7]	-0.0 [-1.8;1.6]	0.0 [-1.8;1.8]
		Bo	-0.1 [-0.8;0.6]	-0.0 [-0.6;0.6]	-0.1 [-0.8;0.5]	-0.0 [-0.9;0.7]	0.0 [-0.6;0.6]	-0.0 [-0.7;0.6]
		In	-0.1 [-1.6;1.7]	0.0 [-1.6;1.7]	0.1 [-1.5;1.8]	0.0 [-1.7;1.8]	0.0 [-1.6;1.6]	-0.0 [-1.7;1.6]
Backward variable selection	20	Be	-0.1 [-1.3;1.2]	-0.1 [-1.4;1.1]	-0.1 [-1.4;1.2]	-0.1 [-1.4;1.2]	-0.1 [-1.4;1.2]	-0.1 [-1.4;1.2]
		Bo	-0.0 [-0.4;0.4]	0.0 [-0.3;0.5]	-0.0 [-0.4;0.4]	0.1 [-0.3;0.5]	0.0 [-0.3;0.4]	0.0 [-0.4;0.5]
		In	0.1 [-1.1;1.2]	0.1 [-1.2;1.2]	0.1 [-1.1;1.3]	0.0 [-1.2;1.2]	0.0 [-1.2;1.2]	0.0 [-1.2;1.2]
	10	Be	0.0 [-1.3;1.5]	-0.0 [-1.3;1.4]	-0.0 [1.4;1.4]	0.0 [-1.4;1.4]	0.0 [-1.3;1.4]	0.0 [-1.3;1.5]
		Bo	-0.1 [-0.6;0.5]	-0.0 [-0.5;0.5]	-0.1 [-0.6;0.4]	-0.0 [-0.5;0.5]	-0.0 [-0.4;0.5]	-0.0 [-0.5;0.5]
		In	0.1 [-1.3;1.4]	0.0 [-1.3;1.4]	0.1 [-1.2;1.4]	-0.0 [-1.3;1.3]	-0.0 [-1.4;1.3]	-0.0 [-1.3;1.3]
	5	Be	0.0 [-1.7;1.7]	0.1 [-1.7;1.8]	0.1 [-1.7;1.8]	0.2 [-1.6;2.0]	0.2 [-1.6;2.0]	0.0 [-1.8;1.8]
		Bo	-0.1 [-0.8;0.5]	-0.1 [-0.6;0.5]	-0.2 [-0.8;0.4]	-0.1 [-1.1;0.6]	-0.0 [-0.6;0.5]	-0.0 [-0.7;0.6]
		In	0.1 [-1.6;1.8]	0.0 [-1.6;1.6]	0.1 [-1.6;1.7]	-0.1[-1.8;1.6]	-0.1[-1.8;1.5]	-0.0[-1.6;1.7]

Variable selection. For the backward variable selection, a variable is not considered in an approach if it is not selected in any of the models of this certain approach. E.g. variable x is not selected for pairwise coupling if it was not retained in the models benign vs borderline, benign vs invasive and borderline vs invasive. After backward variable selection, 1-vs-rest selected the most variables although this approach had the poorest discrimination performance (Table 3.5). MLR and ILR ended up with the most parsimonious prediction models yet had very good discriminative ability.

Table 3.5: Percentage of number of variables selected in each approach for different EPV. The average number of variables is given in the last column.

	Percentage of number of selected variables					Average #
	7	8	9	10	11	variables
EPV 20						
Sequential dichotomous model	-	-	27.0	53.0	20.0	9.9
Pairwise coupling LR	-	-	42.3	45.7	12.0	9.7
Individualized LR	-	-	60.3	36.7	3.0	9.4
1-vs-rest LR	-	-	0.3	43.7	56.0	10.6
Multinomial LR	-	-	52.7	41.3	6.0	9.5
EPV 10						
Sequential dichotomous model	-	1.3	39.3	45.7	13.7	9.7
Pairwise coupling LR	-	2.0	43.7	42.7	11.7	9.6
Individualized LR	-	4.0	68.0	24.7	3.3	9.3
1-vs-rest LR	-	0.3	29.3	54.3	16.0	9.9
Multinomial LR	-	2.7	73.0	22.0	2.3	9.2
EPV 5						
Sequential dichotomous model	0.7	15.7	44.0	24.0	5.7	9.3
Pairwise coupling LR	-	10.7	47.7	36.7	5.0	9.4
Individualized LR	0.3	20.0	64.7	14.0	1.0	9.0
1-vs-rest LR	-	7.0	45.7	40.7	6.7	9.5
Multinomial LR	3.3	25.0	60.0	11.0	0.7	8.8

3.5 Conclusion

In this Chapter different approaches to model a multinomial outcome are discussed. Combinations of dichotomous prediction models can be differentiated from an all-at-once multinomial prediction model. For the former, sequential dichotomous modeling, pairwise coupling, individualized logistic regression and 1-vs-rest logistic regression are covered; while for the latter the most common approach is multinomial logistic regression. The methods differ with respect to number of fitted models, the requirement of a reference category and the straightforwardness of obtaining probabilities. The multinomial logistic regression model is a reference approach with which the other approaches should be compared. As application we compared the different suggested approaches for the diagnosis of ovarian tumors. We noted that the approaches had very similar calibration-in-the-large but differed with respect to selected variables and discriminatory ability. The combination of 1-vs-rest models was not a good approach in our case study. Other approaches showed similar discrimination, but when variable selection was performed MLR and ILR selected fewer variables.

Chapter 4

Multinomial calibration

*In this Chapter, the topic of multinomial calibration is covered. Multinomial calibration statistics and (non-) parametric calibration plots are introduced. We focus on models developed using multinomial logistic regression (*MLR*). In Section 4.7 an application on the case study of ovarian tumors is considered: a multinomial risk prediction model is developed and evaluated [197]. The research concerning multinomial calibration can be found in Van Hoorde et al. 2014 [206].*

4.1 Introduction

Risk prediction models can be evaluated on the level of predictions. As mentioned in Chapter 2 a distinction can be made between the ability to separate different categories (discrimination) and the reliability of the predicted risks (calibration). For dichotomous outcomes, there are many tools to assess discrimination and calibration [174, 220] (see Section 2.3.1). We will discuss multinomial discrimination measures and extend dichotomous calibration statistics and calibration plots towards multinomial risk prediction models. In this Chapter, we focus on models developed using multinomial logistic regression where outcome Y with k categories is predicted using k-1 equations comparing each category i ($i = 2, \ldots, k$) with a reference category 1 using a set of predictors, resulting in k-1 linear predictors. This is the typical approach for predicting multinomial outcomes in medical statistics and epidemiology. For calibration, the Hosmer-Lemeshow (H-L) test statistic is not assessed due to the reported drawbacks [48,63,78,136,217]. The multinomial Hosmer-Lemeshow test statistic evaluates, similarly as for the dichotomous version, the correspondence

between the observed and estimated frequencies in g subgroups of the patient population [48]. The g subgroups are obtained by sorting according to the estimated probabilities $1 - \widehat{\pi}_0$, this corresponds with the dichotomous H-L test-statistic if $k = 2$ [48].

4.2 Methods

In this Section multinomial calibration tools are introduced. To assess calibration several aspects exist which differ in their level of generality. To start, the overall measure calibration-in-the-large and the logistic recalibration framework for multinomial prediction models are introduced. Then, the non-parametric recalibration framework as well as methods to construct parametric and non-parametric multinomial calibration plots from the recalibration results are proposed. The parametric recalibration framework can be used for the estimation and statistical testing of calibration intercepts and slopes. The proposed tools are illustrated on two case studies in the next Sections: one on ovarian tumor diagnosis and one on residual mass diagnosis in patients treated with cisplatin-based chemotherapy for testicular cancer.

We focus on the MLR model of outcome Y with k categories (see Section 3.2.1) [4]. The MLR model in case of reference category 1 consists of the following k-1 submodels:

$$\begin{cases} \log\left(\dfrac{P(\mathrm{Y}=2)}{P(\mathrm{Y}=1)}\right) = a_2 + \displaystyle\sum_{j=1}^{n} b_{j2}\mathrm{x}_j = \mathrm{lp}_{21} \\ \log\left(\dfrac{P(\mathrm{Y}=3)}{P(\mathrm{Y}=1)}\right) = a_3 + \displaystyle\sum_{j=1}^{n} b_{j3}\mathrm{x}_j = \mathrm{lp}_{31} \\ \ldots \\ \log\left(\dfrac{P(\mathrm{Y}=k)}{P(\mathrm{Y}=1)}\right) = a_k + \displaystyle\sum_{j=1}^{n} b_{jk}\mathrm{x}_j = \mathrm{lp}_{k1} \end{cases} \tag{4.1}$$

with n the number of considered predictors ($j = 1, ..., n$).

It is straightforward to investigate the overall calibration-in-the-large (CITL) for MLR models by assessing the differences of the observed and expected (i.e. average predicted probability) event rate for each outcome category. Possibly the most pragmatic and straightforward approach to assess calibration beyond CITL is to use dichotomous calibration tools to independently investigate each of

the k categories. In this way, the calibration of the probability of every category i ($i = 1, ..., k$) would be assessed relative to the probability of 'not i':

$$\log\left(\frac{P(\mathrm{Y}=i)}{1-P(\mathrm{Y}=i)}\right) = \alpha_i + \beta_i \log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) \tag{4.2}$$

with $\widehat{\pi}_i$ the estimated probability of category i obtained from the MLR model. This method is flawed as, contrary to our proposed recalibration framework, this pragmatic approach results in imperfect calibration for the development data. In addition, the result of the recalibration for each category does not sum to one for each patient, such that normalization would be required. Therefore we searched for an alternative approach which is unfortunately more complicated. First, the methodology of the proposed approach is described and afterwards, a worked-out example with $k = 3$ is considered.

4.2.1 Nominal logistic recalibration framework

We propose a nominal logistic recalibration framework where the outcome is regressed on the linear predictors $\{\mathrm{lp}_{21}, \mathrm{lp}_{31}, ..., \mathrm{lp}_{k1}\}$ derived from the prediction model through an MLR fit with the following k-1 submodels:

$$\begin{cases} \log\left(\dfrac{P(\mathrm{Y}=2)}{P(\mathrm{Y}=1)}\right) = \alpha_2 + \displaystyle\sum_{j=2}^{k} \beta_{2j}\mathrm{lp}_{j1} \\ \log\left(\dfrac{P(\mathrm{Y}=3)}{P(\mathrm{Y}=1)}\right) = \alpha_3 + \displaystyle\sum_{j=2}^{k} \beta_{3j}\mathrm{lp}_{j1} \\ \dots \\ \log\left(\dfrac{P(\mathrm{Y}=k)}{P(\mathrm{Y}=1)}\right) = \alpha_k + \displaystyle\sum_{j=2}^{k} \beta_{kj}\mathrm{lp}_{j1} \end{cases} \tag{4.3}$$

with lp_{j1} the linear predictor for category j versus 1 of the MLR prediction model (Equation 4.1). Note that the reference category in the recalibration fit does not have to be the same as the reference category in the multinomial logistic regression model (Equation 4.1).

We can split up the lp_{j1} into what we call 'corresponding' linear predictors when $\text{lp}_{j1} = \text{lp}_{i1}$ and 'non-corresponding' linear predictors when $\text{lp}_{j1} \neq \text{lp}_{i1}$:

$$\begin{cases} \log\left(\dfrac{P(\text{Y}=2)}{P(\text{Y}=1)}\right) = \alpha_2 + \beta_{22}\text{lp}_{21} + \displaystyle\sum_{j=2, j\neq 2}^{k} \beta_{2j}\text{lp}_{j1} \\ \log\left(\dfrac{P(\text{Y}=3)}{P(\text{Y}=1)}\right) = \alpha_3 + \beta_{33}\text{lp}_{31} + \displaystyle\sum_{j=2, j\neq 3}^{k} \beta_{3j}\text{lp}_{j1} \\ \ldots \\ \log\left(\dfrac{P(\text{Y}=k)}{P(\text{Y}=1)}\right) = \alpha_k + \beta_{kk}\text{lp}_{k1} + \displaystyle\sum_{j=2, j\neq k}^{k} \beta_{kj}\text{lp}_{j1} \end{cases} \tag{4.4}$$

In case of perfect calibration the intercepts α_i are zero, the corresponding slopes β_{ii} are one, and the non-corresponding slopes $\beta_{ij,j\neq i}$ are zero (i.e. $\log\left(\frac{P(\text{Y}=i)}{P(\text{Y}=1)}\right) = \text{lp}_{i1}$, $i = 2, ..., k$).

Nominal logistic recalibration framework: worked out example

Consider a prediction model with $k = 3$ and two predictors x_1 and x_2. The prediction model is written as:

$$\begin{cases} \log\left(\dfrac{P(\text{Y}=2)}{P(\text{Y}=1)}\right) = a_2 + b_{12}\text{x}_1 + b_{22}\text{x}_2 = \text{lp}_{21} \\ \log\left(\dfrac{P(\text{Y}=3)}{P(\text{Y}=1)}\right) = a_3 + b_{13}\text{x}_1 + b_{23}\text{x}_2 = \text{lp}_{31} \end{cases} \tag{4.5}$$

The logistic recalibration model for the parametric calibration plots is written as:

$$\begin{cases} \log\left(\dfrac{P(\text{Y}=2)}{P(\text{Y}=1)}\right) = \alpha_2 + \beta_{22}\text{lp}_{21} + \beta_{23}\text{lp}_{31} \\ \log\left(\dfrac{P(\text{Y}=3)}{P(\text{Y}=1)}\right) = \alpha_3 + \beta_{32}\text{lp}_{21} + \beta_{33}\text{lp}_{31} \end{cases} \tag{4.6}$$

with lp_{21} the 'corresponding' linear predictor in the first submodel ($i = 2$) and lp_{31} in the second submodel ($i = 3$).

4.2.2 Nominal non-parametric recalibration framework

A more general non-parametric recalibration framework consists of applying a vector spline smoother on the linear predictors $\{\text{lp}_{21}, \text{lp}_{31}, ..., \text{lp}_{k1}\}$ and considers the following MLR model [56]:

$$\begin{cases} \log\left(\dfrac{P(\text{Y}=2)}{P(\text{Y}=1)}\right) = \alpha'_2 + \displaystyle\sum_{j=2}^{k} \beta'_{2j} s_{2j}(\text{lp}_{j1}) \\ \log\left(\dfrac{P(\text{Y}=3)}{P(\text{Y}=1)}\right) = \alpha'_3 + \displaystyle\sum_{j=2}^{k} \beta'_{3j} s_{3j}(\text{lp}_{j1}) \\ \dots \\ \log\left(\dfrac{P(\text{Y}=k)}{P(\text{Y}=1)}\right) = \alpha'_k + \displaystyle\sum_{j=2}^{k} \beta'_{kj} s_{kj}(\text{lp}_{j1}) \end{cases} \tag{4.7}$$

where lp_{j1} is the linear predictor for category j versus 1 of the MLR prediction model (Equation 4.1) and $s_j(.) = [s_{2j}(.)\ s_{3j}(.)\ ...\ s_{kj}(.)]^T$ a vector spline [177]. Vector splines are natural extensions of the cubic smoothing splines to vector responses [50,230–232]. The k-1 different component functions $s_{ij}(.)$ $(i = 2, ..., k)$ are natural cubic B-splines [50, 177].

Nominal non-parametric recalibration framework: worked out example

The non-parametric recalibration model for the non-parametric calibration plots is written as:

$$\begin{cases} \log\left(\dfrac{P(\text{Y}=2)}{P(\text{Y}=1)}\right) = \alpha'_2 + \beta'_{22} s_{22}(\text{lp}_{21}) + \beta'_{23} s_{23}(\text{lp}_{31}) \\ \log\left(\dfrac{P(\text{Y}=3)}{P(\text{Y}=1)}\right) = \alpha'_3 + \beta'_{32} s_{32}(\text{lp}_{21}) + \beta'_{33} s_{33}(\text{lp}_{31}) \end{cases} \tag{4.8}$$

with $s_2(.) = [s_{22}(.)\ s_{32}(.)]^T$ and $s_3(.) = [s_{23}(.)\ s_{33}(.)]^T$ vector splines.

4.2.3 Nominal calibration plots

Analogous to the dichotomous case, a parametric solution to estimate observed proportions is to use the results of the nominal logistic recalibration framework

(Equation 4.3). The observed proportions can be plotted versus the predicted probabilities in a parametric nominal calibration plot. We can either make a plot for each outcome category separately or combine it into one single overall calibration plot. Separate plots might often be visually more attractive.

A more general non-parametric nominal calibration plot is obtained by using the nominal non-parametric recalibration framework (Equation 4.7). The degree of smoothing in the vector splines $s(.)$ is set by the degrees of freedom. For our case study, we found that 2 degrees of freedom for the vector splines was sufficient (see Section 4.4). This approach can be implemented with the vgam package in `R` (`www.r-project.org`; see Appendix A) [229]. Using this recalibration model, the construction of non-parametric calibration plots is analogous to the parametric version.

4.2.4 Estimation of calibration intercepts and slopes

In the estimation of calibration intercepts and slopes, the non-corresponding linear predictors lp_{j1} from Equation 4.4 can be considered nuisance parameters, as they do not involve category i. The key parameters are α_i and β_{ii} and for estimating these parameters as well as for statistical testing we set $\beta_{ij,j\neq i}$ to 0. The calibration intercepts are obtained as the intercepts α_i when fixing β_{ii} at 1 and $\beta_{ij,j\neq 0}$ at 0. For statistical testing, we extend the approach from Miller et al. [117]. Perfect calibration can be tested using a likelihood ratio chi-squared test with $2 \cdot (k-1)$ degrees of freedom to test the joint null hypothesis that $\alpha_i = 0$ and $\beta_{ii} = 1$. Note that the calibration intercepts and slopes are to be interpreted relative to the category chosen as reference in the recalibration model. For our worked out example (Equation 4.6): a calibration intercept $\alpha_2|(\beta_{22} = 1 \& \beta_{23} = 0) > 0$ indicates underestimation of outcome category 2 relative to outcome category 1, while a calibration slope $\beta_{33}|\beta_{32} < 1$ suggests overfitting of outcome category 3 with respect to the reference category (i.e. category 1). Reference values to define acceptable calibration are arbitrary, and we further comment on this issue in Section 4.5. In any case, we suggest that the calibration intercepts and slopes are interpreted in combination with a calibration plot.

Estimation of calibration intercepts and slopes: worked out example

The calibration intercepts are obtained by fixing the corresponding slopes β_{22} and β_{33} at unity and the non-corresponding slopes β_{23} and β_{32} at zero:

$$\begin{cases} \log\left(\dfrac{P(\mathrm{Y}=2)}{P(\mathrm{Y}=1)}\right) = \alpha_2 + 1 \cdot \mathrm{lp}_{21} + 0 \cdot \mathrm{lp}_{31} \\ \log\left(\dfrac{P(\mathrm{Y}=3)}{P(\mathrm{Y}=1)}\right) = \alpha_3 + 0 \cdot \mathrm{lp}_{21} + 1 \cdot \mathrm{lp}_{31} \end{cases} \tag{4.9}$$

These intercepts show whether the predicted probabilities for each category i are over- or underestimated relative to the probabilities for category 1. The calibration slopes are obtained by setting the non-corresponding slopes β_{23} and β_{32} to zero:

$$\begin{cases} \log\left(\dfrac{P(\mathrm{Y}=2)}{P(\mathrm{Y}=1)}\right) = \alpha_2 + \beta_{22} \cdot \mathrm{lp}_{21} + 0 \cdot \mathrm{lp}_{31} \\ \log\left(\dfrac{P(\mathrm{Y}=3)}{P(\mathrm{Y}=1)}\right) = \alpha_3 + 0 \cdot \mathrm{lp}_{21} + \beta_{33} \cdot \mathrm{lp}_{31} \end{cases} \tag{4.10}$$

The slopes indicate whether there is evidence of overfitting (if below 1) or underfitting (if above 1) of the risks for each category i relative to the risks for category 1.

If the overall test of perfect calibration (4 *df*, i.e. $(k-1)\cdot 2$) is statistically significant, separate tests for the calibration intercepts and calibration slopes may be considered, each with 2 *df* (i.e. k-1 *df*). The corresponding `SAS` and `R` code is provided in Appendix A and B.

4.3 Study set-up

4.3.1 Diagnosis of ovarian tumors

We consider the diagnosis of ovarian tumors with three outcome categories ($k = 3$) for which the optimal management is different: benign, borderline malignant and invasive. More information concerning the data is given in Section 1.4.1. A prediction model is developed using multinomial logistic regression on data from 2037 women recruited at 11 tertiary referral centers for oncology in seven countries within the framework of the International Ovarian Tumor Analysis (IOTA) study (Table 1.2 and 4.1) [88, 183, 184]. The model

contains the following predictors: age, square root of the ratio of the maximum diameter of the solid component and the maximum diameter of the lesion, current use of hormonal therapy (yes/no), irregular internal cyst wall (yes/no), ascites (yes/no), acoustic shadows (yes/no) and presence of masses on both ovaries (yes/no) (Table 1.3). Calibration was evaluated on 1107 women recruited at 7 public hospitals in five countries, also as part of the IOTA study (Table 4.1). As tertiary referral centers usually have higher rates of malignancy than public hospitals, we expect bad calibration-in-the-large (Table 4.1).

4.3.2 Diagnosing of residual mass after chemotherapy for testicular cancer

A similar prediction model as in the previous case study was developed on data from 544 patients with residual mass after chemotherapy for cell cancer (see Section 1.4.2). Most of the patients were treated before 1990 (Table 1.4 and 4.1) [219]. Five predictors were used: presence of teratoma elements in the primary tumor (yes/no), prechemotherapy levels of the serum tumor markers α-fetoprotein (AFP) and human chorionic gonadotropin (hCG) (normal/elevated), the square root of the maximum diameter of the residual mass measured on computerized tomography after chemotherapy, and the change in diameter induced by chemotherapy are considered (Table 1.5). Calibration was evaluated on data from 550 more recently treated patients mostly treated in other centers after 1990 (Table 4.1). The validation data has fewer cases with benign tissue and more with teratoma (33% vs 45% benign, 55% vs 42% teratoma for validation vs development respectively).

4.4 Results

4.4.1 Calibration-in-the-large and calibration plots

For the ovarian tumor case study, calibration-in-the-large was moderately off target (Table 4.2). The probability of a borderline or invasive tumor was overestimated, while the risk of a benign tumor was underestimated. For the testicular cancer case study, the probability of viable cancer was overestimated the probability of a teratoma was underestimated (Table 4.2).

For the recalibration model, we use the category with highest event rate as reference category (benign tumors in the ovarian tumor study, benign tissue in the testicular cancer study). Parametric calibration plots for the ovarian case study are shown in Figure 4.1. Contrary to dichotomous calibration

Table 4.1: Description and descriptive statistics for the case studies for each outcome category separately.

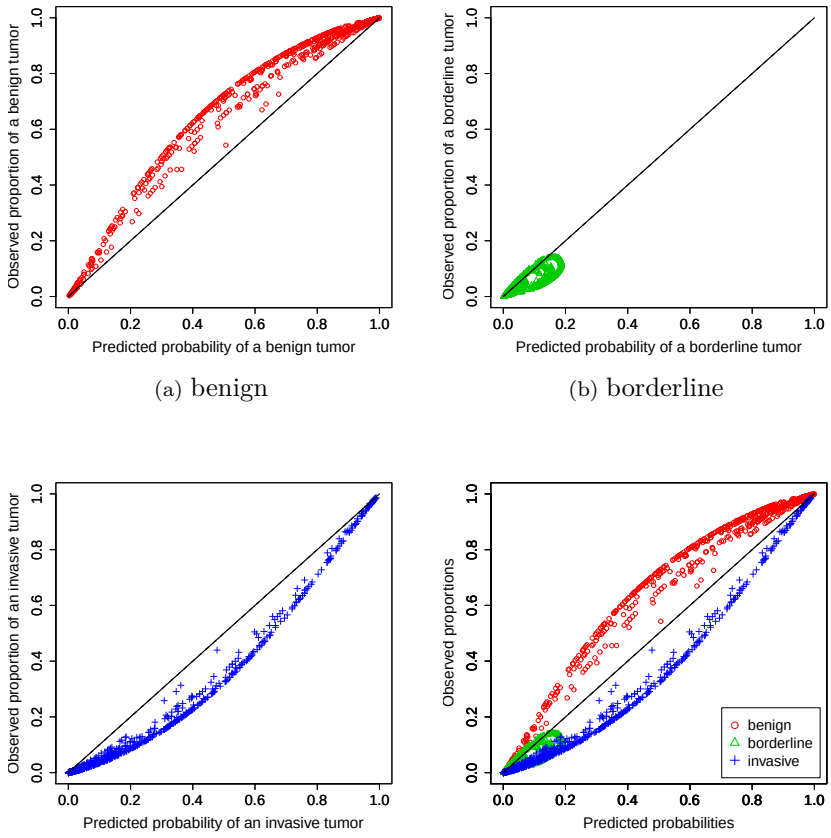
Ovarian tumor	Development set (n=2073)			Validation set (n=1107)		
	Benign	Borderline	Invasive	Benign	Borderline	Invasive
Outcome, N(%)	1315 (65%)	139 (7%)	583 (29%)	918 (83%)	41 (4%)	148 (13%)
Variable, N(%) or mean (SD)						
Age	44 (16)	49 (17)	56 (14)	44 (15)	52 (15)	59 (14)
Square root (propsol)	0.23 (0.36)	0.48 (0.31)	0.79 (0.26)	0.22 (0.33)	0.50 (0.25)	0.80 (0.22)
Bilateral	169 (13%)	18 (13%)	171 (29%)	142 (15%)	8 (20%)	64 (43%)
Ascites	30 (2%)	13 (9%)	254 (44%)	11 (1%)	4 (10%)	52 (35%)
Irregularwall	419 (32%)	105 (76%)	400 (69%)	260 (28%)	31 (76%)	112 (76%)
Hormtherapy	198 (15%)	15 (11%)	56 (10%)	211 (23%)	6 (15%)	23(16%)
Shadows	210 (16%)	4 (3%)	20 (3%)	148 (16%)	1 (2%)	7 (5%)
Testicular cancer	Development set (n=544)			Validation set (n=550)		
	Benign	Teratoma	Viable Cancer	Benign	Teratoma	Viable Cancer
Outcome, n(%)	245 (45%)	231 (42%)	68 (13%)	180 (33%)	304 (55%)	66 (12%)
Variable, n(%) or mean (SD)						
Teratoma	151 (62%)	74 (32%)	27 (40%)	128 (71%)	96 (32%)	27(41%)
AFP	115 (47%)	55 (24%)	16 (24%)	85 (47%)	57 (19%)	11 (19%)
hCG	117 (48%)	67 (29%)	21 (31%)	67 (37%)	87 (29%)	19 (29%)
Square root (residual mass size)	3.94 (2.10)	5.89 (2.57)	6.85 (2.70)	5.36 (2.53)	6.92 (2.89)	7.68 (3.64)
Change in diameter	6.43 (2.62)	2.88 (4.23)	2.88 (3.77)	4.65 (3.85)	0.54 (6.08)	1.14 (7.00)

Table 4.2: Assessment of calibration-in-the-large in the validation data.

	Observed event rate	Expected event rate	Calibration-in-the-large
Ovarian tumor			
- Benign	0.83	0.76	0.07
- Borderline	0.04	0.06	-0.02
- Invasive	0.13	0.18	-0.05
Testicular cancer			
- Benign	0.33	0.33	-0.01
- Teratoma	0.55	0.51	0.04
- Viable cancer	0.12	0.16	-0.04

plots the nominal plots show scatter, indicating that there is no one-to-one relationship between predicted probability and observed proportion. A one-to-one relationship exists between the multidimensional set of k estimated risks and the multidimensional set of k corresponding observed proportions, but this is lost when presenting calibration for each category separately in a two-dimensional plot. Only if there is perfect calibration, the one-to-one relationship still holds in the nominal calibration plots. We return to this issue in Section 4.5. The relation between predicted probability and observed proportion can be smoothed (e.g. using splines) to evaluate whether the relationship on average follows the ideal diagonal line. The overall non-parametric calibration plot for the ovarian case study, including smoothed relationships, is shown in Figure 4.2. For the testicular case study, the parametric plot was highly similar to the non-parametric calibration plot (Figure 4.3). For the ovarian tumor study, the probability of a benign tumor is overestimated whereas the probabilities of a borderline or invasive tumor are underestimated. Due to the low event rate of borderline tumors (4%) the largest predicted probability for this category in the validation data is around 20%. For the testicular cancer study the probability of viable cancer was overestimated. Given similar event rate in the development and validation data, the difference must be attributed to differences in predictor distributions. The probability of teratoma was underestimated, and some overfitting was observed for all categories (slopes < 1).

The non-parametric calibration plot is based on vector cubic smoothing splines, where the level of smoothing can be adapted by varying the degrees of freedom of the splines. A larger *df* allows more flexibility and thus less smoothing. We propose 2 *df* after inspection of the plots based on 2 to 4 *df* (Figure 4.4).



(a) benign (b) borderline

(c) invasive (d) overall

Figure 4.1: Parametric nominal calibration plot for the validation data of the ovarian tumor case study for each outcome category separately (panel (a), (b) and (c)) and overall (panel (d)).

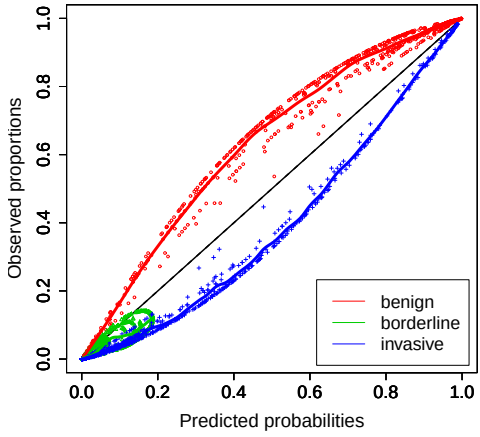
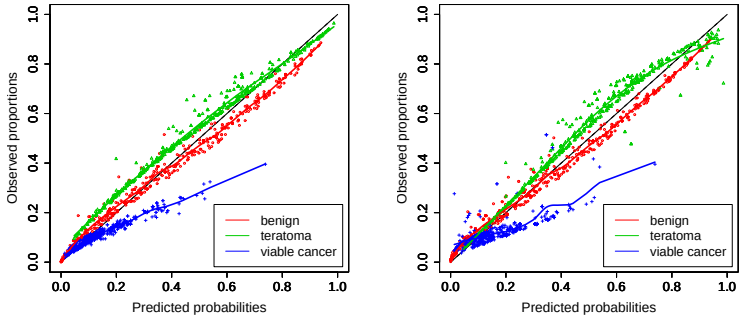
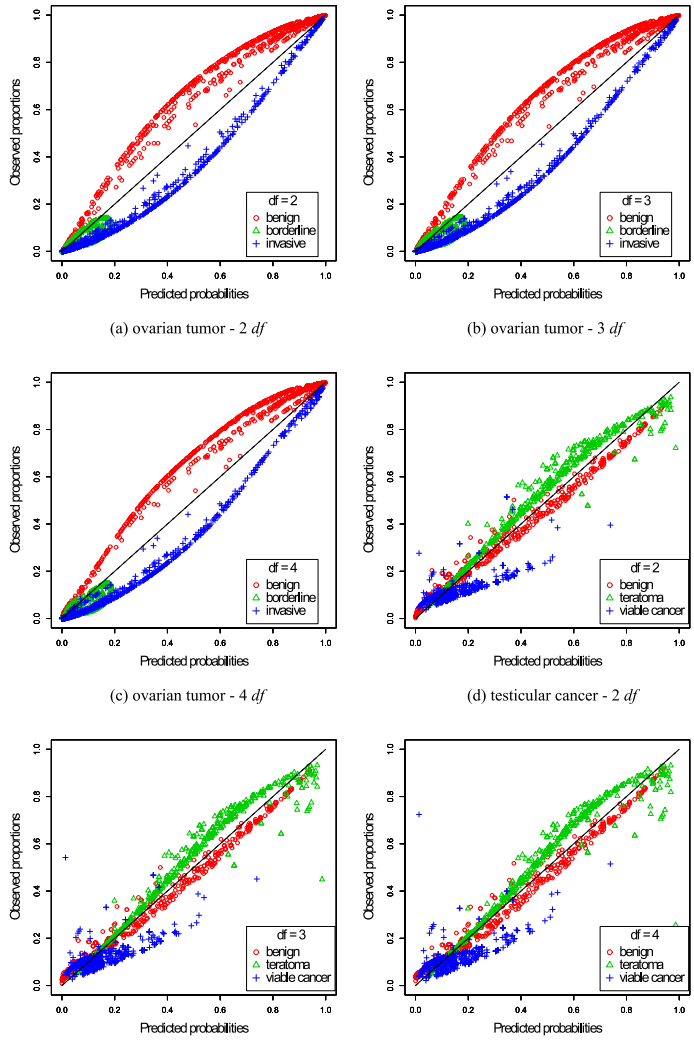


Figure 4.2: Non-parametric nominal calibration plot considering vector splines with 2 *df* for the validation data of the ovarian tumor case study. Cubic spline smoothers are added for each outcome category.



(a) (b)

Figure 4.3: Parametric (panel (a)) and non-parametric (panel (b)) nominal calibration plot considering vector splines with 2 *df* for the validation data of the testicular cancer case study. Cubic spline smoothers are added for each outcome category.



(a) ovarian tumor - 2 *df*

(b) ovarian tumor - 3 *df*

(c) ovarian tumor - 4 *df*

(d) testicular cancer - 2 *df*

(e) testicular cancer - 3 *df*

(f) testicular cancer - 4 *df*

Figure 4.4: Non-parametric calibration plots for the ovarian tumor (a-c) and the testicular cancer (d-f) case study for different degrees of smoothing (2 *df* in (a) and (d), 3 *df* in (b) and (e), 4 *df* in (c) and (f)). The reference categories were respectively benign tumors and benign tissue for the ovarian tumor case study and the testicular cancer case study.

4.4.2 Calibration intercepts and slopes

For the ovarian tumor case study both calibration intercepts are negative (i.e. -0.79 and -0.86), suggesting overestimation of the risks of a borderline malignancy and invasive tumor relative to the risk of a benign tumor (Table 4.3). For the testicular cancer case study, the risk of mature teratoma versus benign tissue is underestimated (i.e. calibration intercept of 0.12) while the risk of viable cancer versus benign tissue is overestimated (i.e. calibration intercept of -0.27). The calibration slopes are close to 1 for the ovarian cancer study (i.e. respectively 1.12 and 1.07), but suggest overfitting for the testicular cancer study (i.e. 0.78 and 0.78). The overall calibration test rejects the null hypothesis of perfect calibration (ovarian tumor study, $p < .0001$; testicular cancer study, $p = 0.0034$). For the ovarian tumor data the separate test for the calibration intercepts was statistically significant ($p < .0001$) but the separate test for the calibration slopes was not ($p = 0.55$), suggesting only a problem with calibration-in-the-large. For the testicular cancer study, both separate tests were significant (p-values 0.01 and 0.03), further demonstrating that we are confronted with a combination of over- and underestimation as well as overfitting with respect to the reference category. These observations are in line with the calibration plots.

Table 4.3: Calibration intercepts α_i when fixing β_{ii} at 1 and $\beta_{ij,j\neq i}$ at 0 and calibration slopes β_{ii} when fixing $\beta_{ij,j\neq i}$ at zero for the validation data with their 95% Wald confidence intervals.

	Calibration intercept	Calibration slope
Ovarian tumor		
- Borderline vs benign	$-0.79\,[-1.11; -0.46]$	$1.12\,[0.81; 1.43]$
- Invasive vs benign	$-0.86\,[-1.10; -0.61]$	$1.07\,[0.91; 1.23]$
Testicular cancer		
- Teratoma vs benign	$0.12\,[-0.10; 0.34]$	$0.78\,[0.62; 0.94]$
- Viable cancer vs benign	$-0.27\,[-0.58; 0.04]$	$0.78\,[0.59; 0.97]$

4.5 Discussion

The importance of evaluating calibration of prediction models has been repeatedly described [7,33,66,97,139]. We proposed tools to evaluate calibration of prediction models for multinomial outcomes, as extensions of existing tools for dichotomous outcomes [36, 39, 117, 167, 174, 220]. We focused on prediction models based on multinomial logistic regression [4], and addressed calibration-in-the-large, logistic recalibration, and parametric and non-parametric calibration

plots. Our main aim was the extension of dichotomous calibration plots [32, 64] to nominal outcomes in order to obtain a visual summary of the calibration of the risk prediction model.

4.5.1 Nominal versus dichotomous calibration

Nominal calibration is more complicated than dichotomous calibration. First, for dichotomous outcomes it is sufficient to evaluate only one of the two categories, whereas for polytomous outcomes we require more output. In the calibration plots, for example, we present results for every k categories. Second, the prediction model based on multinomial logistic regression involves k-1 linear predictors instead of one. Therefore, the logistic recalibration model, where the outcome is regressed on the $k-1$ linear predictors using MLR, has k-1 intercept terms and $(k-1) \cdot 2$ regression coefficients. Each submodel in the recalibration model, contrasting category i $(i = 2, \ldots, k)$ with the reference category 1, has a regression coefficient for the corresponding linear predictor lp_{i1} but also for k-2 so-called non-corresponding linear predictors. This has consequences for the calibration plots as well as for the estimation of calibration slopes.

4.5.2 Calibration plots

Regarding the calibration plots there is the lack of a one-to-one relationship between estimated probabilities and observed proportions, leading to scatter (Figure 4.5). The reason is that predicted risks are k-dimensional meanwhile the calibration is visualized for each category separately. Only in case of perfect calibration, scatter is absent and perfect diagonal relationships are observed (Figure 4.5a). With perfect calibration, we refer to what we label calibration in the 'strong' sense: among patients with x% risk for category 1, y% for category 2, and z% for category 3, it is true that x% belong to category 1, y% to category 2, and z% to category 3. Calibration in the strong sense is likely to be better if the sample size in the development data is sufficiently large and if the validation data is a representative sample from the same population compared to the development data. The presence of scatter still allows calibration in a 'weak' sense: among patients with w% risk of category i, it is true that w% belongs to category i. This averages out the estimated risks of categories 2 and 3. 'Weak' calibration is assessed by considering the trend of the scatter plot for category i, for example using spline smoothers, and by the estimation of calibration slopes as described (Figure 4.5b).

Furthermore, a low event rate of one of the outcome categories compared to the other categories, i.e. imbalanced data, influences the range of the predicted

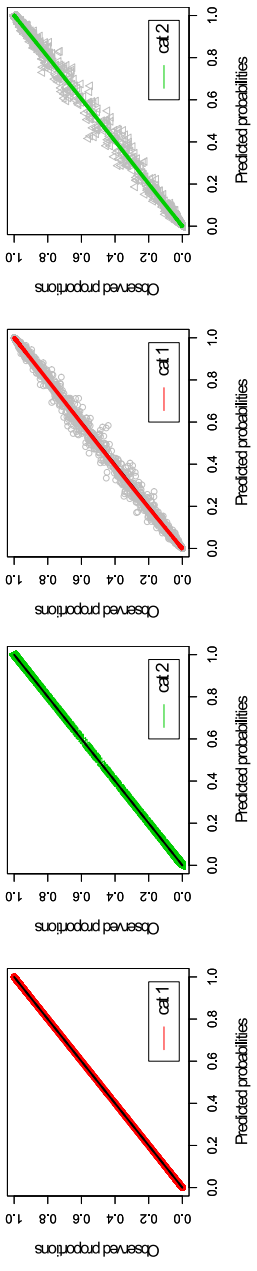
probabilities. The largest predicted probability of such an outcome category will be far below one (Figure 4.3) or even be below 0.5 (Figure 4.1).

4.5.3 Calibration intercepts and slopes

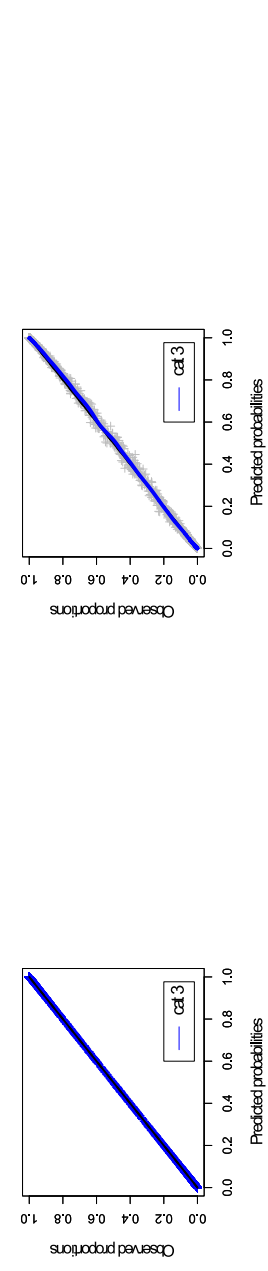
The estimation of calibration intercepts and slopes has limitations. To estimate the slopes (i.e. the regression coefficients for the corresponding linear predictors), as well as for statistical tests for miscalibration, the regression coefficients for the non-corresponding slopes are set to 0. This leads to results that depend on the choice of reference category in the recalibration model. However, for the case studies that we considered results were similar when changing the reference category (Table 4.4). To investigate the potential influence of this dependency, we performed a small simulation study (see Appendix C). Based on these simulations we can conclude that, although the results are slightly different depending on the chosen reference category, similar conclusions can be drawn in various settings where predictor distributions are different between development and validation populations. Note that calibration-in-the-large, calibration intercepts, and calibration plots are independent of the chosen reference category. The non-parametric calibration plots, however, also depend on the reference category because the vector spline gives slightly different fits (Figure 4.6 for the case studies). The simulation study demonstrated that differences are minimal and simply the result of small computational differences. In addition to the dependency on reference category, the calibration intercept and slope for category i have to be interpreted relative to the reference category. In contrast, an advantage of the calibration plots is that they visualize calibration for each outcome category individually. Therefore the plots are more attractive and interpretable.

4.5.4 Guidelines

Presenting guidelines or reference values for interpreting estimated calibration intercepts and slopes is largely arbitrary. Statistical significance, or lack thereof, only tells a part of the story due to the strong dependence on sample size. Following typical bio-equivalence limits, one could argue for a -20% to +20% margin such that the calibration intercept should fall between -0.2 and 0.2 and the calibration slope between 0.8 and 1.2. However, these margins may not simply translate to the current situation. Also, reference values will depend on the specific clinical setting, and it may not be optimal to set independent margins for intercept and slope. Therefore we suggest that the estimated values should be interpreted in combination with a calibration plot. Such plots support

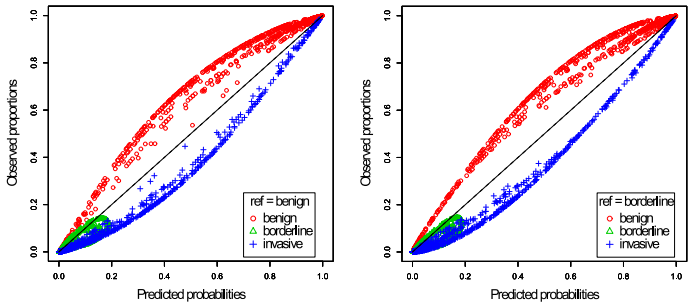


(a) 'Strong' calibration



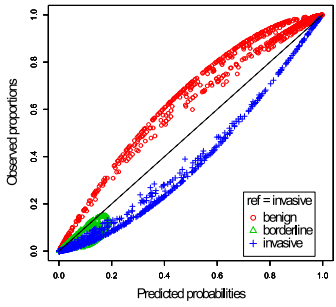
(b) 'Weak' calibration

Figure 4.5: 'Strong' and 'weak' calibration.

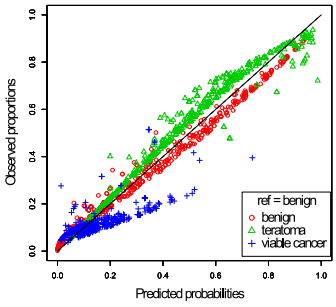


(a) ovarian tumor - benign

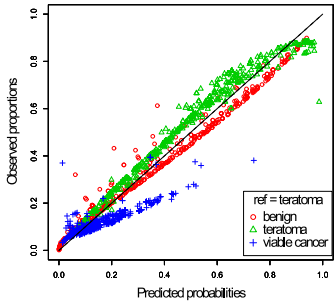
(b) ovarian tumor - borderline



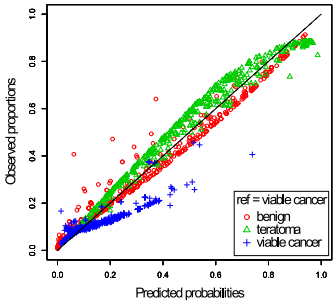
(c) ovarian tumor - invasive



(d) testicular cancer - benign



(e) testicular cancer - teratoma



(f) testicular cancer - viable cancer

Figure 4.6: Non-parametric calibration plots for different choices of reference category. For the ovarian tumor study the reference categories are benign (a), borderline (b) and invasive (c). For the testicular cancer study the reference categories are benign tissue (d), mature teratoma (e) and viable cancer (f).

Table 4.4: Calibration slopes β_{ii} when fixing $\beta_{ij,j\neq i}$ at zero (with 95% Wald confidence intervals) and p-values obtained on the validation data for different choices of reference category.

Reference category	Submodel	Calibration slope	p-values: Joint test of intercept and slopes	p-values: Test of slopes
Ovarian tumor				
Benign	Bo vs Be	1.12 [0.81; 1.43]	<.0001	0.549
	In vs Be	1.07 [0.91; 1.23]		
Borderline	Be vs Bo	1.00 [0.79; 1.21]	<.0001	0.350
	In vs Bo	1.02 [0.93; 1.44]		
Invasive	Be vs In	1.07 [0.93; 1.23]	<.0001	0.301
	Bo vs In	0.91 [0.59; 1.22]		
Testicular cancer				
Benign	Te vs Be	0.78 [0.62; 0.94]	0.0034	0.027
	Vi vs Be	0.78 [0.59; 0.97]		
Teratoma	Be vs Te	0.82 [0.65; 0.98]	0.0018	0.013
	Vi vs Te	0.56 [0.16; 0.95]		
Viable cancer	Be vs Vi	0.79 [0.62; 0.95]	0.0033	0.026
	Te vs Vi	0.76 [0.43; 1.08]		

Abbreviations: Be, benign; Bo, borderline; In, invasive; Te, teratoma; Vi, viable cancer.

the explanation of the calibration intercepts and slopes, and allow judgments of calibration when using specific risk thresholds for classification and decision making.

4.6 What is new?

Key findings

- We extended calibration tools for dichotomous outcomes to polytomous outcomes modeled via a multinomial logistic regression model. The multinomial calibration measures and plots were illustrated on a case study of ovarian tumors and testicular cancer.
- The calibration plots visualize the agreement between predicted probabilities and observed proportions for each outcome category.

What this research adds to what was known

- Calibration is an important aspect in model validation which gains attention. Existing dichotomous calibration tools are extended towards multinomial calibration tools.
- The calibration plots adequately summarize calibration performance of a multinomial risk model in a given cohort of patients and may suggest directions for updating of the model.

Implications of this research

- Before a risk prediction model is used in clinical practice, calibration should be assessed. Calibration plots can be used in order to visualize the reliability of the predicted risks.

4.7 Application: multinomial risk prediction model for ovarian tumors

In this section the use of some of the presented calibration tools is illustrated in an application on the case study of ovarian tumors. This section is based on a manuscript submitted to *The BMJ* [197]. The IOTA data (see Section 1.4.1) were used for the development and validation of a multinomial prediction model. The data from phase 1, 1b and 2 (n=3506)[1] were used for development and the women from phase 3 (n=2403) for validation. The final diagnosis separated five 'tumor groups': benign, borderline, stage I invasive, stage II-IV invasive and metastatic cancer.

Variable selection was performed in two phases. First the number of potential predictors was reduced to ten, based on prior domain knowledge and the stability of the predictors over centers [63, 167, 226]. Four clinical variables (i.e. age, serum CA125, famhistovca, oncology referral center) and six ultrasound variables (i.e. lesdmax, propsol, locules10, papnr, shadows, ascites) were retained (see Section 1.4.1 and Table 1.3). Second, data-driven variable selection using multivariable fractional polynomials (MFP) for each dichotomous risk prediction submodel was considered [155, 158]. This technique simultaneously selects variables and their optimal transformation. The considered dichotomous submodels are borderline versus benign, stage I invasive versus benign, stage

[1] From the original 3511 women, 5 cases were omitted based on logistic regression influence diagnostics [79] and further data review.

II-IV invasive versus benign and secondary metastatic cancer versus benign. The elimination criteria in backward variable selection as well as selection of functional form was set to the p-value that mimics the Bayesian Information Criterion (BIC) [167]. This criterion penalizes for the complexity of the model, with the penalization depending on the number of events: the p-value cut-off decreases when the number of events increases [167]. In our case, the number of events relates to the number of malignancies in each submodel. The p-value cut-off ranges between 0.01 and 0.03 depending on the submodel. E.g. for the submodel benign versus borderline, we have 186 borderline tumors and the p-value associated with BIC is 0.02. Age and oncology referral center were forced into the model, therefore the p-value criterion of variable selection was set to 1. Since some values were missing for the serum CA125 marker (i.e. in 31% of the women), multiple imputation was used in order to create 100 completed data sets [163].

For the risk prediction model a multinomial logistic regression with random center intercept, to acknowledge inter-center variability, was developed [102]. A random intercept MLR model estimates for each center a center-specific intercept instead of one overall intercept. The random intercepts extension of Equation 4.1 with $i = 2, ..., k$ is:

$$\log\left(\frac{\pi_{iml}}{\pi_{1ml}}\right) = a_i + \zeta_{il} + b_i^T \mathbf{x}_{ml} = \mathrm{lp}_{i1}^{\mathrm{RI}} \tag{4.11}$$

with π_{iml} the prediction of outcome i for individual m in center l, $\mathbf{x}_{ml}$ the predictor vector for individual m in center l and ζ_{il} the center specific intercept for center l. The predictor coefficients were shrunk with uniform shrinkage factors to avoid overfitting [167,210]. The model development was done for each of the 100 completed data sets and averaging the obtained linear predictors, without the random intercepts, led to the multinomial probabilities.

Several tools were considered for the evaluation of discrimination: going from dichotomous measures (i.e. AUC, sensitivity and specificity for benign versus malignant), to polytomous measures (i.e. conditional pairwise c-statistics and polytomous discrimination index) and even polytomous visualization (i.e. discrimination plot) [196, 199]. For the assessment of calibration, multinomial calibration plots showing the relationship between predicted probabilities and observed proportions for each tumor group were produced. The predicted probabilities were obtained via a multinomial logistic regression model with random intercepts (Equation 4.11). Therefore, a nominal logistic recalibration framework with random intercepts was used to estimate the observed probabilities. Similarly as above, the random intercept extension of

Equation 4.3 is:

$$\log\left(\frac{P(\mathrm{Y}=i)}{P(\mathrm{Y}=1)}\right) = \alpha_i + \zeta_{il} + \sum_{j=2}^{k} \beta_{ij} \mathrm{lp}_{j1}^{\mathrm{RI}}, \tag{4.12}$$

where $i = 2, ..., k$, $\mathrm{lp}_{j1}^{\mathrm{RI}}$ the linear predictor for category j versus 1 of Equation 4.11 and ζ_{il} a center-specific intercept to account for the inter-center variability [102].

Due to a successful model validation, the model was retrained on the pooled data (n=5909) in order to make use of all available information.

Results

Nine of the ten considered variables were included in the Assessment of Different NEoplasias in the adneXa (ADNEX) model, only family history of ovarian cancer was excluded based on the MFP variable selection analysis. The descriptive statistics for all ten a priori selected variables are given in Table 4.5.

The AUC of the ADNEX model for the discrimination between benign and malignant (1-vs-rest) tumors was 0.954 (95% CI 0.947-0.961) on the development data and 0.943 (0.934-0.952) on the validation data (Table 4.6). The discriminative ability of the different centers was consistent as can be seen in the forest plot in Figure 4.7. The prediction of malignancy was considered for different cut-offs (Table 4.6). A 10% cut-off on the overall risk of malignancy, resulted in a sensitivity of 96.5% and a specificity of 71.3% on the validation data. The conditional pairwise c-statistics on the validation data varied between 0.71 (stage I invasive vs metastatic cancer) and 0.99 (benign vs stage II-IV invasive) (Table 4.7). The outcome groups that were most difficult to distinguish were stage I invasive versus borderline tumors and stage I invasive versus metastatic cancer. The PDI of the developed model was 0.56 (0.54-0.59) on the validation data.

The calibration plots showed a lot of scatter, only allowing 'weak' calibration (Figure 4.8, see Section 4.5.2). Spline smoothers were added and indicated acceptable 'weak' calibration of the predicted probabilities. Below the plots, histograms were plotted to indicate the distribution of the estimated risks. We can for example conclude that the high risks for metastatic cancer were overestimated, but such high risks were rather uncommon.

The discriminative ability of the ADNEX model after retraining on the pooled data (N=5909) is shown in Tables 4.6-4.7. A visualization of the discriminative potential is given in Figure 4.9. The gray horizontal lines are the baseline

Table 4.5: Descriptive statistics of the ten a priori considered predictors stratified for outcome category in the pooled data set (N=5909).

	Benign	BOT	Stage I	Stage II-IV	Metastatic
Outcome, N	N=3980	N=339	N=356	N=988	N=246
Variable, N(%) or median (IQR)					
Age (years)	42 (32-54)	49 (36-62)	54 (44-64)	59 (50-67)	57 (47-68)
CA125⋆	18 (11-39)	30 (16-86)	51 (20-195)	442 (145-1238)	91 (29-271)
famhistovca	79 (2.0)	10 (3.0)	13 (3.7)	57 (5.8)	5 (2.0)
lesdmax	63 (45-87)	86 (51-150)	106 (71-153)	85 (56-123)	86 (56-124)
Solid tissue					
- Presence of solid tissue	1322 (33.2)	267 (78.8)	328 (92.1)	968 (98.0)	234 (95.1)
- Propsol if present	42 (20-100)	37 (24-59)	61 (38-100)	100 (56-100)	100 (64-100)
Papnr					
- 0	3424 (86.0)	135 (39.8)	227 (63.8)	772 (78.1)	213 (86.6)
- 1	333 (8.4)	69 (20.4)	25 (7.0)	56 (5.7)	12 (4.9)
- 2	80 (2.0)	21 (6.2)	17 (4.8)	30 (3.0)	0 (0)
- 3	66 (1.7)	24 (7.1)	17 (4.8)	28 (2.8)	2 (0.8)
- >3	77 (1.9)	90 (26.5)	70 (19.7)	102 (10.3)	19 (7.7)
Locules10	199 (5.0)	74 (21.8)	69 (19.4)	93 (9.4)	36 (14.6)
Shadows	676 (17.0)	8 (2.4)	18 (5.1)	30 (3.0)	10 (4.1)
Ascites	64 (1.6)	28 (8.3)	65 (18.3)	473 (47.9)	90 (36.6)
Oncology referral center	2030 (51)	242 (71)	258 (72)	801 (81)	205 (83)

⋆ The results for CA125 are based on multiple imputation of missing values. The CA125 level is missing for 1447 (36.4%) benign tumors, 62 (18.3%) borderline tumors, 71 (19.9%) stage I, 163 (16.5%) stage II-IV invasive tumors and 62 (25.2%) metastatic tumors.
Abbreviations. BOT, borderline tumor; IQR, interquartile range.

Table 4.6: The AUC and sensitivity and specificity of the ADNEX model for distinguishing benign and malignant tumors on the development, the validation and the pooled data (after retraining). For sensitivity and specificity different cut-offs are considered (i.e. a probability greater than or equal to the cut-off is classified as malignant). The exact Binomial 95% CI are reported.

Threshold for the probability of malignancy	Measure	Development data (N=3056)	Validation data (N=2403)	After retraining on pooled data (N=5909)
-	AUC	0.954 [0.947;0.961]	0.943 [0.934;0.952]	0.950 [0.944;0.955]
3%	Sensitivity	98.8 [97.9;99.4]	98.9 [98.0;99.4]	99.1 [98.6;99.5]
5%		97.9 [96.8;98.7]	98.4 [97.4;99.1]	98.0 [97.3;98.6]
10%		95.9 [94.4;97.1]	96.5 [95.2;97.6]	96.4 [95.4;97.2]
15%		94.4 [92.8;95.8]	94.2 [92.5;95.6]	94.5 [93.4;95.5]
3%	Specificity	52.3 [50.4;54.3]	46.6 [44.0;49.2]	43.4 [41.8;45.0]
5%		65.4 [63.6;67.3]	59.4 [56.8;62.0]	61.1 [59.5;62.6]
10%		75.5 [73.8;77.2]	71.3 [68.9;73.7]	73.2 [71.8;74.6]
15%		81.0 [79.4;82.5]	77.2 [74.9;79.3]	78.7 [77.4;79.9]

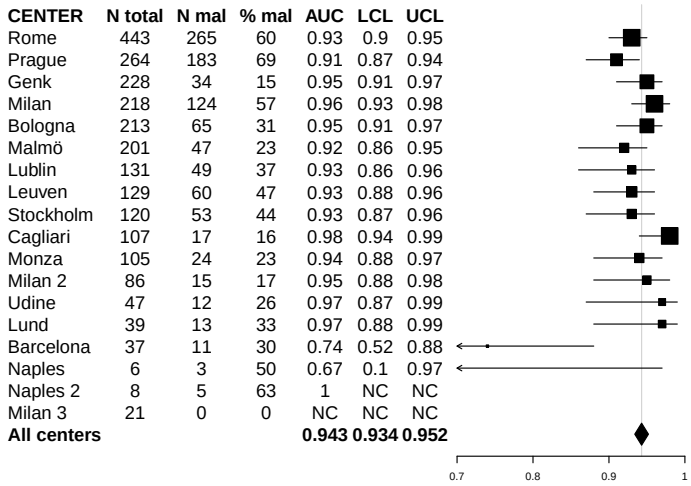


Figure 4.7: Forest plot with AUCs for the different centers of the validation data for the discrimination between benign and malignant tumors. (NC means not computed)

Table 4.7: Polytomous discrimination measures (i.e. conditional pairwise c-statistics and polytomous discrimination index (PDI)) with 95% CI of the ADNEX model on the development, the validation data and the pooled data (after retraining). The random performance corresponds with a c-statistic of 0.5 and a PDI of 0.2 (i.e. $1/k$).

Performance measures	Development data (N=3056)	Validation data (N=2403)	After retraining on pooled data (N=5909)
AUC			
- Benign vs BOT	0.91 [0.88;0.93]	0.85 [0.82;0.88]	0.88 [0.87;0.90]
- Benign vs Stage I	0.94 [0.92;0.96]	0.92 [0.90;0.93]	0.93 [0.92;0.94]
- Benign vs Stage II-IV	0.99 [0.98;0.99]	0.99 [0.98;0.99]	0.99 [0.98;0.99]
- Benign vs Metastatic	0.96 [0.95;0.98]	0.95 [0.93;0.97]	0.96 [0.95;0.97]
- BOT vs Stage I	0.71 [0.65;0.76]	0.75 [0.69;0.79]	0.75 [0.71;0.79]
- BOT vs Stage II-IV	0.91 [0.88;0.93]	0.95 [0.93;0.96]	0.93 [0.91;0.95]
- BOT vs Metastatic	0.86 [0.81;0.90]	0.87 [0.82;0.91]	0.88 [0.85;0.91]
- Stage I vs Stage II-IV	0.83 [0.79;0.86]	0.87 [0.83;0.90]	0.85 [0.82;0.87]
- Stage I vs Metastatic	0.77 [0.71;0.82]	0.71 [0.65;0.76]	0.75 [0.70;0.78]
- Stage II-IV vs Metastatic	0.76 [0.71;0.81]	0.82 [0.78;0.86]	0.80 [0.76;0.83]
PDI	0.554 [0.530;0.579]	0.567 [0.540;0.591]	0.569 [0.553;0.586]

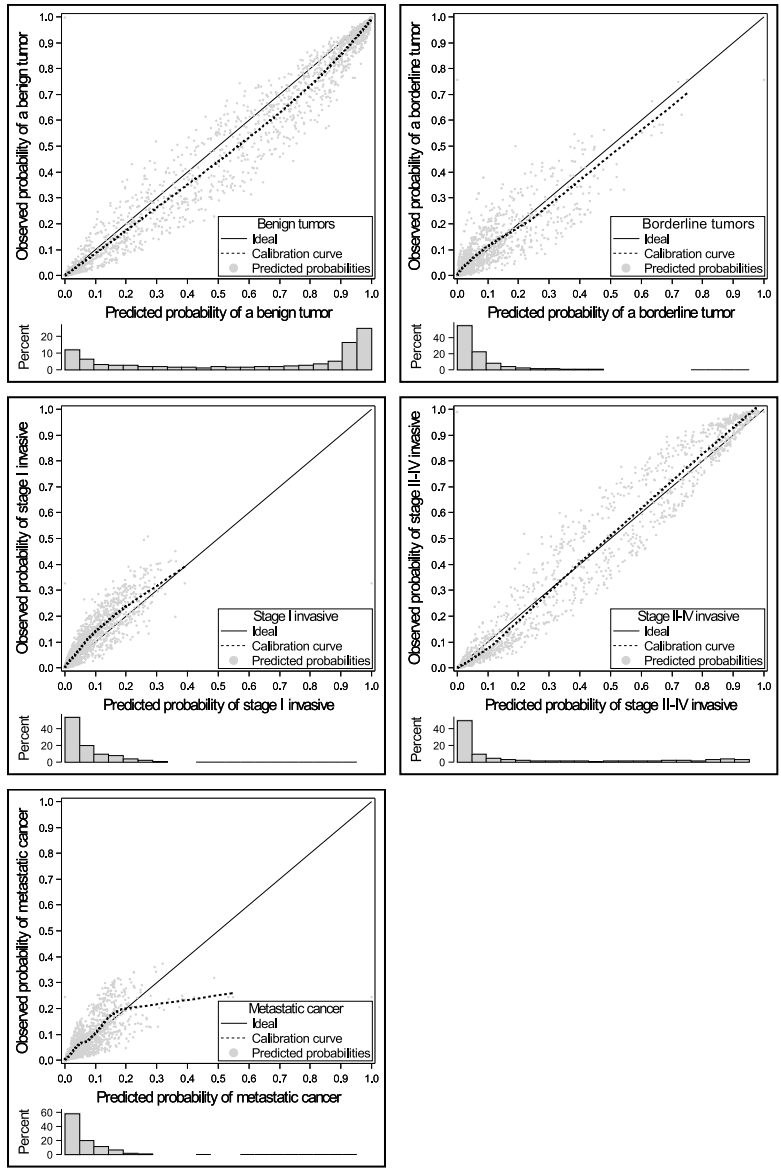
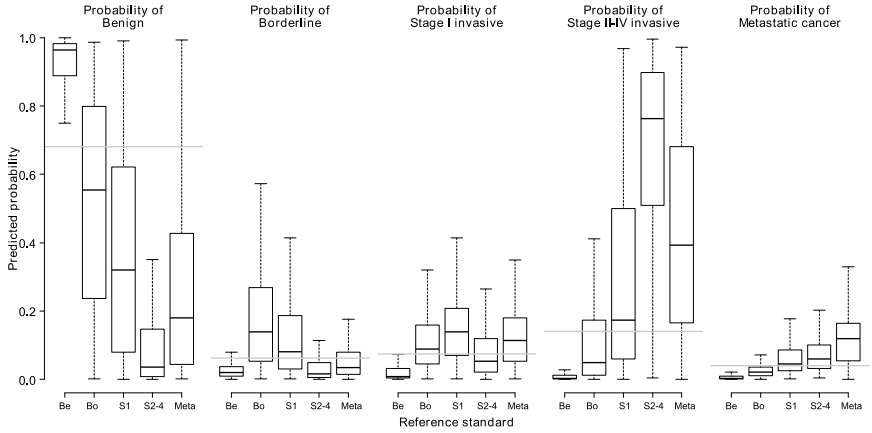


Figure 4.8: Calibration plot of the ADNEX model for each outcome category on the validation data. Perfect agreement between predicted probabilities (x-axis) and observed probabilities (y-axis) corresponds with the diagonal line. Below the calibration plot the distribution of the predicted probabilities is shown in a histogram.

probabilities obtained by fitting a random intercepts multinomial LR model containing only intercepts. We can conclude that the predicted probability of a specific outcome category is highest for cases with that specific outcome category (i.e. reference standard).



Abbreviations: Be, benign tumor; Bo, borderline tumor; S1, Stage I invasive tumor; S2-4, Stage II-IV tumor; Meta, Metastatic cancer.

Figure 4.9: Discrimination plot of the ADNEX model on the pooled data (N=5909) after retraining. Boxplots of the predicted probabilities for each outcome category stratified for the actual observed outcome category are presented. The gray horizontal lines show the baseline probability of each outcome category.

Conclusion

The developed multinomial risk prediction model was validated on data recruited at a later time. Due to promising results, the model was retrained on all data and resulted in the current ADNEX model. This model discriminates very well benign and malignant tumors but can also reliably distinguish four types of malignant tumors (i.e. borderline, stage I invasive, stage II-IV invasive and metastatic cancer). The ADNEX model is available online as well as through mobile applications.

4.8 Conclusion

This Chapter started with an overview of some multinomial discrimination measures. A distinction can be made between set measures, 1-vs-rest and pairwise, and overall measures. Further, extensions of calibration tools for dichotomous outcomes towards multinomial outcomes are given. The (non-) parametric calibration plots are of particular interest since these visualize the agreement between predicted probabilities and observed proportions for each outcome category. These plots summarize the performance of a multinomial risk prediction model with respect to calibration and may indicate possible directions for model updating. To end, the use of the proposed calibration plots was illustrated on an application of the ovarian tumor case study.

Chapter 5

Multinomial updating

In this Chapter, the topic on updating risk models for multinomial outcomes is covered. First, in Section 5.2, an introduction is made by modeling the multinomial outcome via sequential dichotomous models (cf. Chapter 3). In this way, existing dichotomous updating techniques can be used (cf. Chapter 2). Afterwards, in Section 5.3, direct multinomial updating techniques for multinomial logistic regression are proposed and compared. For the assessment of discrimination and calibration, measures proposed in Chapter 4 are considered. The results of this Chapter are described in a paper in Journal of Clinical Epidemiology [205] and a submitted manuscript for Statistics in Medicine.

5.1 Introduction

Risk prediction models are important for personalized healthcare and to assist with clinical decision making. Therefore these models have to be sufficiently accurate and need to be validated with respect to their performance. The core elements of this performance include discrimination and calibration [167, 174, 186]. It is of particular importance to perform validation of the model using data collected later in time (temporal validation) and/or in different locations or hospitals (external or geographical validation) [5, 87, 97]. Disappointing validation results do not necessarily imply that the prediction model should be discarded. Model updating, which combines information from the existing model with new data, might be a sensible alternative [83,167,168,186,205,207]. Despite a model's disappointing performance, it may still contain useful information such

as relevant predictors or predictor effects. It may be more efficient to update an existing prediction model than to start the development anew, especially in relatively small validation sets [168]. This approach has clear practical relevance, because it is often not realistic to expect that a single model will work in various settings due to differences in patient management protocols across centers and regions. In addition, models may often loose performance over time, for example due to changes in management protocols or referral patterns. Furthermore it might be that new promising markers or predictive variables are found.

5.2 Dichotomous updating methods

The aim of this section is assessing whether and how previously suggested updating techniques for dichotomous models (cf. Chapter 2) translate to the setting of multinomial outcome prediction [83, 167, 168]. Therefore, the multinomial outcome is modeled via sequential dichotomous modeling (cf. Chapter 3) [52, 154]. Recalibration, revision and redevelopment methods are illustrated on two case studies dealing with testicular cancer and ovarian tumors.

5.2.1 Patients and methods

Study set-up

The case studies deal with predicting ovarian pathology and predicting residual mass histology after chemotherapy for testicular cancer. We derived a prediction model on the development data, and applied it to validation data from a different setting (Figure 5.1). The model was updated with a number of updating methods. To investigate the performance of the updating methods, the validation data were divided at random into an updating set and a test set on which the updated model was evaluated. This division was repeated 500 times and the average over test sets calculated.

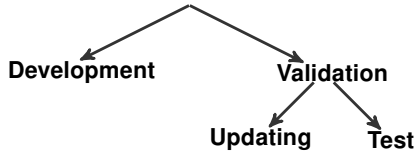


Figure 5.1: Data structure: development and validation data the latter subdivided in updating and test set.

Case studies

Ovarian tumor. When an ovarian tumor is detected on ultrasound examination, an accurate diagnosis of the tumor as benign, borderline malignant, or invasive is crucial for selecting the optimal management strategy for each patient [193]. More information concerning the case study and the data is given in Section 1.4.1. Development data were collected at tertiary referral centers for oncology in seven countries within the framework of the International Ovarian Tumor Analysis (IOTA) study [181,183,204]. The data set contained 2037 women including 1315 (65%) with a benign tumor, 139 (7%) with a borderline malignancy and 583 (29%) with invasive cancer (Table 5.1). The prediction model was updated on data from regional or public hospitals in five countries, also as part of the IOTA study. The validation data contained 1107 patients (918 (83%) with a benign tumor, 41 (4%) with a borderline malignancy and 148 (13%) with invasive cancer) (Table 5.1).

Testicular cancer. After patients have been treated with cisplatin-based chemotherapy for advanced nonseminomatous testicular germ cell cancer, retroperitoneal lymph nodes may still contain tumor cells [15,219]. An accurate prediction of the residual mass containing necrosis or fibrosis (benign tissue), teratoma, or viable cancer is important to decide whether or not the mass should be surgically resected (teratoma or viable cancer), and whether additional chemotherapy is indicated (viable cancer). This dataset and case study is further clarified in Section 1.4.2. The development data consisted of resected masses of patients recruited from different hospitals in three countries [219]. Among 502 patients, 237 (47%) had benign masses, 213 (42%) a teratoma and 52 (10%) viable cancer (Table 5.1). The prediction model was updated on data from a tertiary referral center from Indiana University. The validation data consist of 273 resected masses (Table 5.1), of which 76 (28%) were benign, 162 (59%) were teratomas and 35 (13%) were viable cancers [219].

Polytomous prediction models

Polytomous prediction models were obtained using a sequential dichotomous modeling strategy [52, 154]. Roukema and colleagues observed similar performance for polytomous logistic regression and sequential dichotomous logistic regression [154]. For outcomes with three categories as in our case studies, this involves the development of two models: one for category A versus B+C and one for category B vs C (Section 3.2.2). The risk estimates for categories B or C were obtained by multiplying the risk for B+C from the first model with the risk for category B or category C from the second model. Logistic regression was used to develop the dichotomous models. In general

Table 5.1: Description and descriptive statistics for the case studies (i.e. ovarian tumor and testicular cancer).

Ovarian tumor	Development set (n=2037)	Validation set (n=1107)
Setting	Tertiary referral centers for oncology in seven countries	Regional or public hospitals in five countries
Outcome, n(%)		
Benign	1315 (65%)	918 (83%)
Borderline	139 (7%)	41 (4%)
Invasive	583 (29%)	148 (13%)
Variable, n(%) or mean (SD)		
Age	48 (16)	46 (16)
Square root (propsol)*	0.41 (0.42)	0.31 (0.37)
Bilateral	358 (17.6%)	214 (19.3%)
Ascites	297 (14.6%)	67 (6.1%)
Irregularwall	923 (45.3%)	403 (36.4%)
Hormtherapy	269 (13.2%)	240 (21.7%)
Shadows	234 (11.5%)	156 (14.1%)
Testicular cancer	**Development set (n=502)**	**Validation set (n=273)**
Setting	Different hospitals in three countries	Tertiary referral center from Indiana University
Outcome, n(%)		
Benign	237 (47%)	76 (28%)
Teratoma	213 (42%)	162 (59%)
Viable cancer	52 (10%)	35 (13%)
Variable, n(%) or mean (SD)		
Teratoma	235 (46.8%)	104 (38.1%)
AFP	175 (34.9%)	68 (24.9%)
hCG	193 (38.4%)	75 (27.5%)
Square root (residual mass size)	4.75 (2.35)	7.79 (3.08)
Change in diameter (per 10%)	4.87 (5.73)	1.44 (5.73)

* The ratio of the maximum diameters of the largest solid component and the lesion, i.e. the proportion solid tissue if present.

for k categories there will be k-1 dichotomous models. There are different ways to combine categories (Section 3.2.2). Where possible, the choice of which categories to combine at each stage should depend on clinical judgment based on differences between categories concerning treatment consequences and clinical profile. Category prevalence may also play a role.

Ovarian tumor. The first model discriminated between benign and malignant (borderline or invasive) tumors, the second model between borderline and invasive tumors. The models estimated respectively the probability of a benign tumor and of a borderline tumor. Seven variables were considered: age, square root of the ratio of the maximum diameters of the largest solid component and the lesion, presence of masses on both ovaries (yes/no), presence of ascites (yes/no), irregular internal cyst wall (yes/no), current use of hormonal therapy (yes/no) and presence of acoustic shadows (yes/no) (Table 5.1 and Section 1.4.1). After backward variable selection with Akaike's Information Criterion (AIC) as stopping rule, the dichotomous logistic regression model for benign versus malignant contained all seven variables. The second dichotomous model contained five variables and excluded current use of hormonal therapy and presence of acoustic shadows (the model coefficients are shown in Table 5.2).

Testicular cancer. The first model discriminated between benign and malignant (teratoma or viable cancer) resected masses, the second model between teratomas and viable cancers [166]. The models estimated respectively the probability of a benign mass and of a teratoma. For both models five predictor variables were considered: presence of teratoma elements in the primary tumor (yes/no), pre-chemotherapy levels of the serum tumor markers α-fetoprotein (AFP) and human chorionic gonadotropin (hCG) (normal/elevated), the square root of the maximum diameter of the residual mass measured on computerized tomography after chemotherapy, and the change in diameter induced by chemotherapy (Table 5.1 and Section 1.4.2). Backward variable selection with AIC as stopping rule was used to develop the prediction models. The first model (benign vs malignant) included all five predictors whereas the second model (teratoma vs viable cancer) contained four predictors after dropping the prechemotherapy level of the serum tumor marker AFP (Table 5.2).

Updating methods

Six different updating methods were considered (METHOD 0-4 and R from Table 2.1) [168] and applied to the two sequential dichotomous models on the updating data. None of these methods used the original development data to update the prediction model. The methods varied in the number of parameters

Table 5.2: Model coefficients for the original dichotomous prediction models, and average coefficients for the updated prediction models obtained on the 500 updating sets. For model redevelopment (METHOD R), only coefficients that differed from zero were considered to compute the average, i.e. coefficients from selected predictors.

Ovarian tumor	Original coefficients*	EPV 5 METHOD 2	EPV 5 METHOD 3	EPV 5 METHOD 4	EPV 5 METHOD R	EPV 3 METHOD 2	EPV 3 METHOD 3	EPV 3 METHOD 4	EPV 3 METHOD R
Benign vs malignant									
Age	-0.03	-0.03	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.04
Square root (propsol)	-3.13	-3.44	-3.63	-3.47	-3.91	-3.47	-3.62	-3.53	-3.98
Bilateral	-0.59	-0.65	-0.63	-0.65	-0.64	-0.66	-0.64	-0.66	0.79
Ascites	-2.24	-2.46	-2.39	-2.46	-2.42	-2.48	-2.45	-2.49	-2.50
Irregularwall	-1.39	-1.53	-1.49	-1.52	-1.45	-1.54	-1.52	-1.54	-1.48
Hormtherapy	0.73	0.80	0.77	0.78	0.55	0.81	0.77	0.76	0.77
Shadows	2.13	2.34	2.18	2.29	1.71	2.36	2.21	2.29	1.81
Benign vs malignant									
Age	-0.03	-0.03	-0.03	-0.03	-0.03	-0.04	-0.03	-0.04	-0.04
Square root (propsol)	-2.89	-3.22	-3.68	-3.45	-4.92	-3.35	-3.86	-3.71	-5.29
Bilateral	-0.72	-0.80	-0.72	-0.78	-0.93	-0.83	-0.77	-0.80	-1.40
Ascites	-1.65	-1.84	-1.61	-1.70	-1.19	-1.91	-1.67	-1.65	-3.60
Irregularwall	0.45	0.50	0.45	0.43	1.04	0.52	0.45	0.42	-0.09
Hormtherapy	-◇	-◇	-◇	-◇	1.02	-◇	-◇	-◇	-1.99
Shadows	-◇	-◇	-◇	-◇	-16.36	-◇	-◇	-◇	-16.85

Continued on next page

Table 5.2 – *Continued from previous page*

Testicular cancer	Original coefficients⋆	EPV 5 METHOD 2	EPV 5 METHOD 3	EPV 5 METHOD 4	EPV 5 R	EPV 3 METHOD 2	EPV 3 METHOD 3	EPV 3 METHOD 4	EPV 3 R
Benign vs malignant									
Teratoma	0.91	0.64	1.37	1.37	1.69	0.66	1.28	1.27	1.77
AFP	0.88	0.62	1.00	1.15	1.43	0.64	0.92	1.10	1.56
hCG	0.50	0.35	0.41	0.27	0.78	0.36	0.42	0.29	0.85
Square root (residual mass size)	-0.11	-0.08	-0.05	0.00	0.13	-0.08	-0.05	-0.01	0.09
Change in diameter (per 10%)	0.26	0.18	0.13	0.15	0.12	0.19	0.16	0.17	0.15
Teratoma vs viable									
Teratoma	-0.65	-0.75	-0.71	-0.71	-0.91	-0.77	-0.70	-0.70	-0.91
AFP	-◇	-◇	-◇	-◇	-1.13	-◇	-◇	-◇	1.08
hCG	-0.53	-0.62	-0.62	-0.60	-0.92	-0.64	-0.62	-0.59	-0.77
Square root (residual mass size)	-0.20	-0.23	-0.24	-0.24	-0.31	-0.24	-0.25	-0.25	-0.34
Change in diameter (per 10%)	-0.12	-0.13	-0.13	-0.14	-0.15	-0.14	-0.14	-0.14	-0.18

⋆ The coefficients from the original dichotomous prediction models obtained on the development data. These coefficients are unchanged for METHOD 0 (no updating) and METHOD 1 (intercept adjustment).
◇ This variable was not considered in the original dichotomous prediction model.

that were (re-)estimated. A short description of the methods is given below (or Section 2.4).

The simplest method (METHOD 0) uses the original model without any adjustment, and therefore represents a reference method with which the other methods are compared. METHOD 1 and METHOD 2 are recalibration methods. METHOD 1 only updates the model's intercept to correct the calibration-in-the-large, i.e. to take care that the mean predicted probability equals the observed prevalences. METHOD 2 updates the intercept and uses the calibration slope as a single overall adjustment factor to update the regression coefficients. METHOD 3 and 4 are more extensive revision methods. METHOD 3 first uses the approach from METHOD 2 to update the model coefficients. Afterwards, likelihood ratio tests are used for each predictor to assess whether the effect in the updating data is statistically significantly different from the effect obtained using METHOD 2. If so, the effect of the predictor is re-estimated. A forward re-estimation procedure is followed: we first consider the predictor with the strongest difference, and continue until all tests fail to reach statistical significance. For METHOD 4 we re-estimate the intercept and individual model regression coefficients for all predictors on the updating data. For METHOD 3 and 4 parameter shrinkage is used to prevent overfitting. We used a simple heuristic shrinkage factor defined as (model χ^2 - df) / model χ^2 [168, 210]. Model χ^2 refers to the difference in -2 log likelihood between the model with re-estimated predictors and the re-calibrated model (METHOD 2), and df corresponds to the difference in degrees of freedom of these two models. By multiplying the difference in regression coefficients of METHOD 3 or METHOD 4 and METHOD 2 with the shrinkage factor (see Equation 2.18), the obtained coefficients are pulled towards their re-calibrated values [83, 168, 210]. The intercept of the shrunken model is re-estimated to ensure that the sum of predicted probabilities equals the sum of observed outcomes [168]. Finally, for METHOD R a completely new model with exactly the same procedure as used for developing the original model was built (i.e. including backward variable selection).

Performance evaluation, events per variable and simulations

We assessed the discrimination and calibration performance of the polytomous risk predictions for each updating method. Discrimination was assessed with the Polytomous Discrimination Index (PDI, Section 3.3) and the standard c-statistic for the two sequential dichotomous models [196]. PDI estimates the probability to correctly identify the patient from a randomly selected outcome category from a group of patients who all belong to a different category [196]. PDI corresponds to $1/k$ in case of random performance [196]. The c-statistic in

case of random performance corresponds to 0.5. The calibration is assessed using calibration-in-the-large: we determined for each outcome category the difference between the observed prevalence and the average predicted risk (i.e. the expected prevalence).

Given that one updating method involves the development of a new model on the updating data set, we decided to vary the number of events per variable (EPV) in the updating set to assess its influence. We calculated EPV by dividing the size of the smallest outcome category of the updating set by the number of considered variables in the model development process, thereby acknowledging that this ignores the fact that each variable is considered in two dichotomous models. To vary EPV, we considered different sample sizes for the updating data, while the size of the test set was kept the same. We focused on low EPV situations (EPV 5 or 3). For the ovarian tumor study we considered seven predictors, such that the smallest category in the updating set has 35 events for EPV 5 (7 predictors $\cdot 5$ events per predictor $= 35$) and 21 for EPV 3 ($7 \cdot 3 = 21$). For the testicular cancer case study we considered five predictors, such that the smallest category in the updating set has 25 events for EPV 5 ($5 \cdot 5 = 25$) and 15 for EPV 3 ($5 \cdot 3 = 15$). The random division in updating and test data was stratified by outcome category (Table 5.3). The division of the validation set into updating and test set was repeated 500 times, and average test set results were reported.

Table 5.3: Sizes of development, updating and test set

	Ovarian tumor: total sample size (separate for benign - borderline - invasive)		
EPV	Development set	Updating set	Test set
5	2037 (1315 - 139 - 583)	945 (784 - 35 - 126)	162 (134 - 6 - 22)
3	2037 (1315 - 139 - 583)	567 (470 - 21 - 76)	162 (134 - 6 - 22)
	Testicular cancer: total sample size (separate for benign - teratoma - viable cancer)		
EPV	Development set	Updating set	Test set
5	502 (237 - 213 - 52)	195 (54 - 116 - 25)	78 (22 - 46 - 10)
3	502 (237 - 213 - 52)	117 (33 - 69 - 15)	78 (22 - 46 - 10)

5.2.2 Results

Performance of original models in development data

For the ovarian tumor study the c-statistic was $0.91\,[95\%\text{CI} : 0.90 - 0.93]$ for modeling benign vs malignant tumors and $0.84\,[95\%\text{CI} : 0.81 - 0.88]$ for model-

Table 5.4: Polytomous discrimination as assessed by the PDI on the updating and test data. The values are averaged over the 500 divisions of the validation data into updating and test data. The 5th and 95th percentiles of these results are shown between square brackets.

Ovarian tumor		METHOD 0	METHOD 1	METHOD 2	METHOD 3	METHOD 4	METHOD R
EPV 5	Updating	0.74 [0.72; 0.75]	0.72 [0.71; 0.74]	0.73 [0.71; 0.74]	0.73 [0.71; 0.75]	0.73 [0.71; 0.75]	0.73 [0.71; 0.76]
	Test	0.73 [0.63; 0.82]	0.72 [0.62; 0.81]	0.72 [0.62; 0.81]	0.72 [0.62; 0.81]	0.72 [0.62; 0.81]	0.70 [0.61; 0.80]
EPV 3	Updating	0.74 [0.69; 0.78]	0.73 [0.68; 0.77]	0.73 [0.68; 0.78]	0.74 [0.69; 0.79]	0.73 [0.69; 0.78]	0.74 [0.70; 0.79]
	Test	0.73 [0.63; 0.83]	0.72 [0.62; 0.81]	0.72 [0.62; 0.82]	0.72 [0.61; 0.81]	0.72 [0.61; 0.81]	0.70 [0.60; 0.79]
Testicular cancer		METHOD 0	METHOD 1	METHOD 2	METHOD 3	METHOD 4	METHOD R
EPV 5	Updating	0.56 [0.52; 0.60]	0.56 [0.52; 0.60]	0.55 [0.51; 0.59]	0.59 [0.55; 0.63]	0.59 [0.55; 0.64]	0.61 [0.57; 0.65]
	Test	0.57 [0.47; 0.66]	0.56 [0.48; 0.65]	0.55 [0.46; 0.65]	0.57 [0.48; 0.67]	0.58 [0.48; 0.67]	0.58 [0.49; 0.68]
EPV 3	Updating	0.56 [0.50; 0.63]	0.56 [0.50; 0.63]	0.56 [0.49; 0.63]	0.59 [0.52; 0.67]	0.60 [0.52; 0.68]	0.61 [0.52; 0.70]
	Test	0.56 [0.47; 0.65]	0.56 [0.47; 0.65]	0.55 [0.46; 0.64]	0.56 [0.47; 0.65]	0.57 [0.47; 0.66]	0.55 [0.44; 0.65]

ing borderline vs invasive malignancies. The PDI was $0.73\,[95\%\text{CI} : 0.70 - 0.76]$. For the testicular cancer study the c-statistic for modeling benign vs malignant masses was $0.81\,[95\%\text{CI} : 0.77 - 0.85]$, and for modeling teratoma vs viable cancer $0.65\,[95\%\text{CI} : 0.56 - 0.74]$. The PDI of the polytomous model was $0.57\,[95\%\text{CI} : 0.52 - 0.62]$.

Discrimination after updating

The average multinomial discrimination performance of the different updating methods on the 500 test sets is shown in Table 5.4. The c-statistic for the dichotomous models is identical for METHODS 0, 1 and 2 because the ranking of the predicted risks is maintained for recalibration (Table 5.5). However, when combining the dichotomous models to obtain a polytomous risk model, small decreases for the PDI were observed for METHODS 0 to 2. For the ovarian tumor case study, redevelopment showed similar discrimination as the original model on the updating set but lower discrimination on the test set. For the testicular cancer case study, revision (METHODS 3 and 4) or redevelopment resulted in a higher PDI on the updating set compared to no updating. This increase was not maintained in the test set, especially with an EPV of 3.

Table 5.5: C-statistics for the sequential dichotomous models averaged over the 500 updating and test data. The 5th and 95th percentiles of these results are shown between square brackets.

Ovarian tumor		METHOD 0-2	3	4	R
EPV 5	Benign vs malignant				
	Updating	0.92 [0.92; 0.93]	0.93 [0.92; 0.93]	0.92 [0.92; 0.93]	0.93 [0.92; 0.93]
	Test	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.95]
	Borderline vs invasive				
	Updating	0.83 [0.81; 0.86]	0.84 [0.82; 0.87]	0.84 [0.81; 0.87]	0.85 [0.83; 0.88]
	Test	0.83 [0.69; 0.95]	0.83 [0.69; 0.95]	0.83 [0.69; 0.95]	0.77 [0.35; 0.93]
EPV 3	Benign vs malignant				
	Updating	0.92 [0.91; 0.94]	0.93 [0.91; 0.94]	0.93 [0.91; 0.94]	0.93 [0.91; 0.94]
	Test	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.87; 0.95]
	Borderline vs invasive				
	Updating	0.84 [0.78; 0.89]	0.85 [0.79; 0.90]	0.84 [0.79; 0.90]	0.86 [0.81; 0.91]
	Test	0.83 [0.70; 0.95]	0.83 [0.70; 0.95]	0.83 [0.70; 0.95]	0.79 [0.57; 0.95]
Testicular cancer		METHOD 0-2	3	4	R
EPV 5	Benign vs malignant				
	Updating	0.78 [0.75; 0.81]	0.82 [0.78; 0.85]	0.82 [0.79; 0.85]	0.82 [0.79; 0.85]
	Test	0.78 [0.70; 0.86]	0.80 [0.72; 0.87]	0.81 [0.73; 0.88]	0.81 [0.73; 0.88]
	Teratoma vs viable cancer				
	Updating	0.67 0.67 [0.62; 0.72]	0.68 0.68 [0.62; 0.73]	0.68 0.68 [0.62; 0.74]	0.70 0.70 [0.64; 0.75]
	Test	0.68 [0.54; 0.80]	0.67 [0.54; 0.80]	0.67 [0.54; 0.80]	0.66 [0.52; 0.80]
EPV 3	Benign vs malignant				
	Updating	0.78 [0.73; 0.84]	0.81 [0.75; 0.87]	0.82 [0.77; 0.88]	0.82 [0.76; 0.88]
	Test	0.78 [0.70; 0.86]	0.79 [0.70; 0.87]	0.79 [0.72; 0.87]	0.79 [0.70; 0.88]
	Teratoma vs viable cancer				
	Updating	0.67 [0.58; 0.76]	0.69 [0.58; 0.78]	0.69 [0.59; 0.79]	0.71 [0.50; 0.81]
	Test	0.68] [0.54; 0.79]	0.67 [0.53; 0.79]	0.67 [0.53; 0.79]	0.63 [0.50; 0.77]

Table 5.6: Calibration-in-the-large results on the test data, expressed as the difference in the observed prevalence (in %) and the expected prevalence (in %). The values are averaged over the 500 divisions of the validation data into updating and test data. The 5th and 95th percentiles of these results are shown between square brackets.

Ovarian tumor		METHOD 0	METHOD 1	METHOD 2	METHOD 3	METHOD 4	METHOD R
EPV 5	Benign	7.2 [4.7; 10.0]	−0.2 [−2.7; 2.3]	−0.2 [−2.8; 2.4]	−0.2 [−2.8; 2.3]	−0.2 [−2.8; 2.4]	−0.2 − [−2.9; 2.4]
	Borderline	−2.5 [−3.2; −1.8]	−0.4 [−1.1; 0.2]	−0.3 − [−1.0; 0.3]	−0.2 [−0.9; 0.4]	−0.3 [−1.0; 0.4]	0.0 [−0.7; 0.6]
	Invasive	−4.7 [−7.0; −2.4]	0.7 [−1.7; 2.9]	0.5 [−1.9; 2.8]	0.4 [−2.0; 2.8]	0.5 [−1.9; 2.8]	0.2 [−2.2; 2.6]
EPV 3	Benign	7.4 [5.0; 10.1]	0.0 [−2.7; 2.6]	0.0 [−2.7; 2.8]	0.1 [−2.6; 2.9]	0.0 [−2.7; 2.8]	0.1 [−2.8; 2.8]
	Borderline	−2.5 [−3.2; −1.8]	−0.5 [−1.2; 0.2]	−0.4 [−1.2; 0.3]	−0.3 [−1.1; 0.4]	−0.3 [−1.1; 0.4]	−0.1 [−0.8; 0.6]
	Invasive	−4.9 [−7.1; −2.7]	0.5 [−1.9; 2.7]	0.3 [−2.2; 2.8]	0.2 [−2.3; 2.6]	0.3 [−2.2; 2.6]	0.0 [−2.5; 2.6]
Testicular cancer		METHOD 0	METHOD 1	METHOD 2	METHOD 3	METHOD 4	METHOD R
EPV 5	Benign	1.8 [−1.6; 5.3]	0.1 [−4.9; 5.0]	0.2 [−3.6; 4.0]	0.2 [−3.9; 4.3]	0.2 [−3.8; 4.3]	0.2 [−4.3; 4.5]
	Teratoma	1.8 [−1.5; 4.5]	0.3 [−4.4; 4.4]	0.4 [−3.7; 3.9]	0.1 [−4.2; 4.0]	0.0 [−4.4; 3.9]	−0.2 [−4.3; 3.8]
	Viable cancer	−3.6 [−5.1; −2.3]	−0.4 [−2.0; 1.2]	−0.6 [−2.3; 1.1]	−0.3 [−2.2; 1.4]	−0.2 [−1.9; 1.5]	0.0 [−2.3; 2.0]
EPV 3	Benign	2.1 [−1.4; 5.6]	0.0 [−5.6; 5.5]	0.0 [−4.7; 4.6]	0.0 [−5.0; 4.8]	0.0 [−5.2; 4.8]	0.1 [−5.5; 5.2]
	Teratoma	1.6 [−1.8; 4.6]	0.3 [−4.8; 5.3]	0.5 [−4.2; 4.5]	0.2 [−4.4; 4.9]	0.2 [−4.7; 4.9]	0.0 [−4.7; 5.1]
	Viable cancer	−3.7 [−5.1; −2.2]	−0.3 [−2.0; 1.4]	−0.5 [−2.5; 1.6]	−0.3 [−2.4; 1.9]	−0.2 [−2.1; 2.0]	−0.1 [−2.9; 2.5]

Calibration after updating

Average calibration-in-the-large over the 500 test sets was poor when using the original sequential dichotomous prediction model, especially for the ovarian tumor case study (Table 5.6). Calibration was greatly improved by all other methods.

Updating parameters, variable selection and estimated effects

The positive intercept adjustments (METHOD 1) of the ovarian tumor case study indicate a systematical underestimating of the probability of a benign tumor and of a borderline malignancy in the original dichotomous models (Table 5.7). This may not be surprising as women presenting to public hospitals have much

less aggressive tumors compared to women attending an oncology referral center. The slope adjustments (METHOD 2) were close to 1.

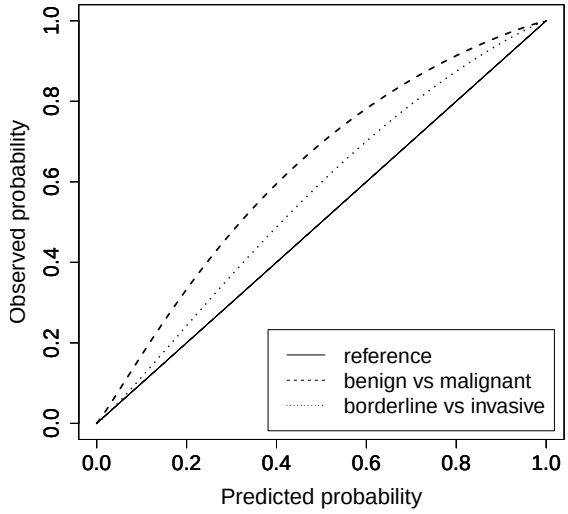
For the testicular cancer models, the positive intercepts indicate that the original dichotomous models on average underestimated the probability of benign masses and the risk of teratomas (Table 5.7). The calibration slope showed that the original model coefficients were too high in the benign vs malignant model (slope smaller than 1). For the teratoma vs viable cancer model the original coefficients were too low (slope larger than 1). The average recalibration results from METHOD 2 were used to generate calibration plots of the original dichotomous models (Figure 5.2).

Table 5.7: Average of the estimated updating parameters for the different considered updating methods.

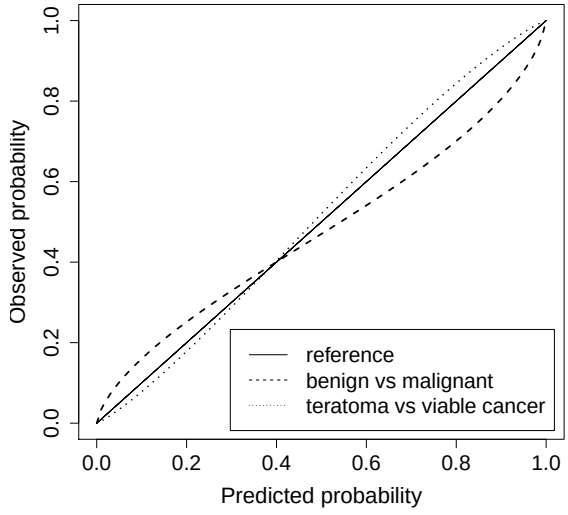
Ovarian tumor	Benign vs malignant		Borderline vs invasive	
Method - parameter	EPV 5	EPV 3	EPV 5	EPV 3
METHOD 1 - intercept adjustment	0.84	0.83	0.27	0.26
METHOD 2 - intercept	0.83	0.82	0.40	0.44
METHOD 2 - slope	1.10	1.11	1.11	1.16
METHOD 3 - shrinkage	0.35	0.24	0.29	0.23
METHOD 4 - shrinkage	0.06	0.08	0.12	0.13
Testicular cancer	Benign vs malignant		Teratoma vs viable cancer	
Method - parameter	EPV 5	EPV 3	EPV 5	EPV 3
METHOD 1 - intercept adjustment	0.12	0.14	0.26	0.26
METHOD 2 - intercept	-0.12	-0.08	0.08	0.06
METHOD 2 - slope	0.70	0.73	1.16	1.19
METHOD 3 - shrinkage	0.78	0.62	0.08	0.11
METHOD 4 - shrinkage	0.66	0.48	0.06	0.08

METHOD 3 selectively re-estimated individual coefficients. For the ovarian tumor study, re-estimation was rare, except for one predictor (Table 5.8). For the first model of the testicular cancer study all but one variable were frequently re-estimated despite the low number of events per variable. For the second model re-estimation was less frequent, related to similar effects of predictors and lower sample size. Generally, effects that were often re-estimated with EPV 5 were less often re-estimated with EPV 3 due to a decrease in power. The reverse was seen for effects that were rarely re-estimated with EPV 5, due to an increase in variability.

METHOD 3 and 4 are revision methods that make use of shrinkage. The average



(a) Ovarian tumors



(b) Testicular cancer

Figure 5.2: Logistic calibration curves for the original sequential dichotomous models. The plots use the average intercept and slope adjustments for METHOD 2 with EPV 5 over the 500 updating sets. Panel (a) gives the curves for ovarian tumor study, panel (b) for the testicular cancer study.

Table 5.8: Frequency of effect re-estimation for METHOD 3 for the 500 updating sets.

Ovarian tumor	Benign vs malignant		Borderline vs invasive	
	EPV 5	EPV 3	EPV 5	EPV 3
Age	0	3	0	4
Square root (propsol)	216	101	214	131
Bilateral	0	4	0	5
Ascites	0	0	29	35
Irregularwall	0	8	1	16
Hormtherapy	8	18	-$^{\diamond}$	-$^{\diamond}$
Shadows	74	69	-$^{\diamond}$	-$^{\diamond}$
Testicular cancer	**Benign vs malignant**		**Teratoma vs viable cancer**	
	EPV 5	EPV 3	EPV 5	EPV 3
Teratoma	284	214	30	29
AFP	155	68	-$^{\diamond}$	-$^{\diamond}$
hCG	8	14	0	5
Square root (residual mass size)	124	90	36	52
Change in diameter (per 10%)	217	165	0	5

$^{\diamond}$ This variable was not considered in the original prediction model and is therefore not considered for re-estimation.

shrinkage factors are shown in Table 5.7. As expected EPV 3 led to more shrinkage.

Model redevelopment (METHOD R), led to variable selection in agreement with the original models for the ovarian tumor study, and less so for the testicular cancer study (Table 5.9). This is consistent with the stronger slope adjustments by METHOD 2 for the testicular cancer study, and the more frequent effect re-estimations by METHOD 3, and the observation that the predictor effects differed more strongly between the original and the new setting for the testicular cancer study (Table 5.2, mainly the comparison of original coefficients with average coefficients for METHOD 4). This explains why discrimination in the updating set could be increased by methods 3 to 5 for the testicular cancer study but not for the ovarian tumor study. Table 5.9 further demonstrated the instability of the variable selection results. The standard deviation of the number of selected variables increased when EPV decreased. Also, predictors with strong effects were selected less often and predictors with weak effects more often when EPV decreased. This is similar to the findings for METHOD 2. A

lower EPV results in a lower power of selecting important covariates, furthermore the risk exists of selecting one or more random covariates due to the multiple comparisons that are made [42,164]. Finally, the overfitting of model coefficients for model redevelopment can be seen in Table 5.2.

Table 5.9: Frequency of variable selection and mean number of selected variables for model redevelopment (METHOD R) for the 500 updating sets.

Ovarian tumor	Benign vs malignant		Borderline vs invasive	
	EPV 5	EPV 3	EPV 5	EPV 3
Age	500	500	500	500
Square root (propsol)	500	500	500	500
Bilateral	490	350	225	137
Ascites	500	500	195	163
Irregularwall	500	499	1	20
Hormtherapy	316	208	1	31
Shadows	500	500	61	107
Mean number of selected variables (SD)	*6.6 (0.49)*	*6.1 (0.65)*	*3.0 (0.63)*	*2.9 (0.78)*
Testicular cancer	Benign vs malignant		Teratoma vs viable cancer	
	EPV 5	EPV 3	EPV 5	EPV 3
Teratoma	500	499	28	56
AFP	500	456	98	108
hCG	30	79	44	62
Square root (residual mass size)	16	48	499	448
Change in diameter (per 10%)	498	448	498	390
Mean number of selected variables (SD)	*3.1 (0.30)*	*3.1 (0.52)*	*2.3 (0.51)*	*2.1 (0.92)*

5.2.3 Discussion

Prediction models for polytomous outcomes can be very relevant for appropriate decision making, yet such outcomes often suffer from categories with low prevalence. Previously suggested methods to update existing dichotomous outcomes were translated to the polytomous setting. The polytomous models were based on sequential dichotomous modeling. A model to predict category A vs category B+C was combined with a model to predict category B vs category C. Both dichotomous models were updated separately to obtain updated polytomous predictions. This approach was demonstrated on two

case studies for which the updating setting involved other (types of) hospitals than the hospitals used for model development. Using the original model ('no updating') resulted in poor calibration-in-the-large whereas every other updating method greatly improved the calibration-in-the-large in a similar way. Regarding discrimination, we observed that none of the updating methods could improve performance compared to no updating. This is consistent with previous studies on model updating, and may be partly related to the insensitivity of c-statistics to detect differences in performance [221]. Overall, the findings corroborate those obtained when updating dichotomous models [66, 82, 83, 167, 168, 186].

Sample size

Improvement of model discrimination through updating is more likely when predictor effects are strongly different in the new setting, and sample size of the updating data is sufficient. Further, again if sample size permits, model extension by adding one or more additional predictors may result in better discrimination depending on the situation. For example, a model can be extended with a newly available and promising marker, or with potentially predictive variables not considered when developing the original model [171]. Model extension was not considered due to limited EPV. As a consequence of the low EPV, known perils such as model instability and overfitting also became visible in a polytomous setting when extensive updating methods were used. This tended to reduce performance in new patients. Not surprisingly, these problems were most pronounced for model redevelopment, with highly unstable variable selection results and overfitted model coefficients.

Sample size is typically a more severe problem for polytomous outcomes than for dichotomous outcomes, since one or more of the outcome categories often have very low prevalence. This necessitates the availability of large data sets to have a sufficient number of EPV for such categories, as clearly illustrated in our study. For the ovarian tumor case study the validation data contained 1107 patients, yet only 41 of them had a borderline malignant tumor. Therefore we varied the number of EPV even though this has already been studied extensively for dichotomous models. We simply divided the number of cases in the smallest category by the number of considered predictors, however for polytomous models more than one coefficient can be estimated per predictor. For example, we considered in the ovarian tumor case study seven predictors to predict a three-category outcome, such that two coefficients per predictor were estimated. If we adopt the common rule of thumb of EPV$\geq$10, we would need at least $7 \cdot 2 \cdot 10 = 140$ borderline tumors. With a prevalence of 41 in 1107 (3.7%), a total sample size of at least 3800 patients is then recommended. For the testicular cancer study, a similar calculation yields a recommended sample size

of at least 780 patients. Probably, one reason why polytomous outcomes are often dichotomized is to reduce sample size requirements.

Model validation

Model validation is crucial before deploying a prediction model in routine clinical practice. This also includes an assessment of the differences in prevalence and patient characteristics between the patients from the setting where a model was developed and patients from the setting in which a model is planned to be used. This may reveal that updating would be advisable in order to ascertain that accurate probabilities are obtained.

Remark concerning model calibration

In this Section we considered the overall calibration-in-the-large for the assessment of the reliability of the predicted risks. In addition, calibration intercepts and slopes of the dichotomous prediction models could have been calculated, although this does not evaluate the multinomial risk predictions on itself. Calibration plots of the multinomial risk predictions are not given since chronologically, this research topic was considered first, before the calibration plots were extended towards polytomous outcomes modeled via multinomial logistic regression (cf. Chapter 4) and their extension towards multinomial risk prediction models in general (cf. Chapter 6).

5.2.4 What is new?

Key findings

- Updating methods for prediction models with dichotomous outcomes show similar behavior when applied to polytomous prediction models based on sequential dichotomous modeling. Simple recalibration methods were preferred over no updating or redevelopment.
- Sample size issues are common for polytomous outcomes, such that revision methods can often not be reliably implemented even when predictor effects are different in the updating population.

What this research adds to what was known

- Existing updating methods can be used for polytomous prediction models. Similar to the updating of dichotomous models, simple methods are recommended, especially when the amount of data available for updating is limited. When sample size is large, more complex methods such as revision or even redevelopment can be a valid alternative.

Implications of this research

- To obtain more accurate predictions, polytomous prediction models should be updated when performance in a new setting is poor. The preferred updating method depends on the amount of data available. Because of these findings, further research will focus on updating methods for multinomial logistic regression, which is the most common approach for polytomous outcomes (Section 5.3).

5.3 Nominal updating methods

The aim of this section is introducing methods to directly update risk models for polytomous outcomes based on multinomial logistic regression, which is the most commonly used method for nominal outcomes. We present recalibration, revision, and extension methods and illustrate these on a case study dealing with pregnancy of unknown location (PUL). This is a condition where a woman has a positive pregnancy test although the pregnancy cannot be visualized using transvaginal ultrasound [12, 93]. In order to guide further follow-up, we want to predict how likely it is that the PUL is a failed PUL (FPUL), an ectopic pregnancy (EP) or an intra-uterine pregnancy (IUP) (Section 1.4.3) [12]. M4 is a multinomial risk model developed for this purpose, and is based on the serum human chorionic gonadotropin (hCG) levels at presentation (hCG0) and around 48 hours later (hCG48) [29]. In this Section, the M4 model is updated in time and in an external setting.

5.3.1 Patients and methods

Case study

The M4 model was developed on data from women with a PUL recruited at St George's Hospital (SGH) in London between March and November 2002 [29].

Among the 197 patients, 109 (55%) had a FPUL, 76 (39%) an IUP and 12 (6%) an EP (Table 5.10). M4 has the following predictors: the logarithm of the average of hCG0 and hCG48 (hCGm), the ratio of hCG48 and hCG0 (hCGr), and hCGr squared. A centered version of hCGr was used by subtracting the median ratio of 1.17 from all values (hCGrc). M4 is a multinomial logistic regression model with IUP as the reference category. The linear predictor for failed PUL versus IUP is

$$\text{lp}_{FI} = \log\left(\frac{\pi_{\text{FPUL}}}{\pi_{\text{IUP}}}\right) = 5.88 - 1.18 \cdot \log(\text{hCGm}) - 5.56 \cdot \text{hCGrc} + 2.05 \cdot \text{hCGrc}^2, \quad (5.1)$$

while the linear predictor for EP versus IUP is

$$\text{lp}_{EI} = \log\left(\frac{\pi_{\text{EP}}}{\pi_{\text{IUP}}}\right) = 0.39 - 0.06 \cdot \log(\text{hCGm}) - 0.26 \cdot \text{hCGrc} - 3.93 \cdot \text{hCGrc}^2. \quad (5.2)$$

The predicted risks for each outcome category are obtained as

$$\begin{cases} \pi_{\text{FPUL}} = \dfrac{e^{\text{lp}_{FI}}}{1 + e^{\text{lp}_{FI}} + e^{\text{lp}_{EI}}}, \\ \pi_{\text{IUP}} = \dfrac{1}{1 + e^{\text{lp}_{FI}} + e^{\text{lp}_{EI}}}, \\ \pi_{\text{EP}} = \dfrac{e^{\text{lp}_{EI}}}{1 + e^{\text{lp}_{FI}} + e^{\text{lp}_{EI}}}. \end{cases} \quad (5.3)$$

In this paper we will update the M4 model in time, using data from St George's collected after November 2002, and externally using data from Queen Charlotte and Chelsea's Hospital (QCCH) in London.

Patients

The recruitment of women with a pregnancy of unknown location took place between November 2002 and February 2007 at SGH, and between April 2009 and December 2013 at Queen Charlotte and Chelsea's Hospital (QCCH) [29, 94, 189]. Of the 2708 patients, 2295 (85%) are available for analysis after applying the exclusion criteria (see Section 1.4.3). For SGH, 83 of 1610 (5.2%) were lost to follow-up, 12 (0.7%) had no hCG48 level, and 93 (5.8%) women had an hCG0 level $\leq$25 IU/l. For QCCH, 60 of 1098 (5.5%) patients were lost to follow-up, 1 (<0.1%) woman had no hCG0 level, 125 (11.4%) no hCG48 level, 22 (2.0%) had hCG48 measured more than 3 days after hCG0, and 17 (1.9%) had an initial hCG level $\leq$25 IU/l (see Section 1.4.3). Therefore, 1422 patients are available

from SGH for temporal updating, and 873 from QCCH for external updating. The patients are randomly divided into updating and validation data in a 2:1 ratio, stratified according to outcome. M4 will be updated on 949 and validated on 473 pregnancies from SGH, while for QCCH respectively 583 and 290 pregnancies will be considered (Table 5.10). At SGH there were 50% FPUL, 41% IUP and 9% EP. The QCCH data contained 58% FPUL, 28% IUP and 14% EP.

Updating methods

Eight updating methods were implemented (Table 5.11). The updating methods consisted of two recalibration, three revision and three extension approaches.

Reference method. METHOD 0 applies the original prediction model without any adjustments.

Recalibration methods. The simplest recalibration method (METHOD 1) updates the intercepts of the two linear predictors lp_{FI} and lp_{EI} of M4 :

$$\begin{cases} \text{lp}_{1,FI} = \alpha_{FI} + \text{lp}_{FI} = \alpha_{FI} + 1 \cdot \text{lp}_{FI} + 0 \cdot \text{lp}_{EI} \\ \text{lp}_{1,EI} = \alpha_{EI} + \text{lp}_{EI} = \alpha_{EI} + 0 \cdot \text{lp}_{FI} + 1 \cdot \text{lp}_{EI} \end{cases} \tag{5.4}$$

This adaptation leads to a correction of the calibration-in-the-large to ensure that the observed event rates and the mean predicted probabilities of the three outcome categories are the same. For METHOD 2 a multinomial logistic recalibration framework with lp_{FI} and lp_{EI} is applied [206]:

$$\begin{cases} \text{lp}_{2,FI} = \alpha_{FI} + \beta_{FF} \cdot \text{lp}_{FI} + \beta_{FE} \cdot \text{lp}_{EI} \\ \text{lp}_{2,EI} = \alpha_{EI} + \beta_{EF} \cdot \text{lp}_{FI} + \beta_{EE} \cdot \text{lp}_{EI} \end{cases} \tag{5.5}$$

In this way, intercepts (α), coefficients for 'corresponding' linear predictors (β_{ii}) and coefficients for 'non-corresponding' linear predictors ($\beta_{ij,j\neq i}$) are estimated in order to update the regression coefficients [206]. This method allows adjustments to correct for miscalibration of the predicted probabilities from M4 (over- or underestimation and over- or underfitting).

Revision methods. METHOD 4 and 4P re-estimate the intercepts and the coefficients of the individual predictors using the updating data. Method

Table 5.10: Description and descriptive statistics for development data M4, updating and validation data of SGH (i.e. temporal updating) and QCCH (i.e. external updating).

		SGH		QCCH	
	Development	Updating	Validation	Updating	Validation
Predictors					
- hCGm (IU/L)	273 (633)	465 (1083)	463.5 (894)	514.5 (1320)	616 (1342.25)
- hCGr	0.80 (1.66)	1.06 (1.70)	1.03 (1.73)	0.60 (1.15)	0.70 (1.09)
- Age (years)	30 (8)	30 (9.75)[1]	31 (8)[3]	32 (10)	32 (10)
- *prog* (nmol/L)	16.5 (61.25)	20 (55)[2]	23 (62)[4]	10 (30)[5]	8 (31.25)[6]
Outcome, n (%)					
- FPUL	109 (55)	478 (50)	239 (51)	335 (57)	167 (58)
- IUP	76 (39)	385 (41)	192 (41)	164 (28)	81 (28)
- EP	12 (6)	86 (9)	42 (9)	84 (14)	42 (14)

Data are expressed as median (interquartile range). Some data were missing. [1] The maternal age is recorded for 476 FPUL, 384 IUP and 86 EP. [2] The initial progesterone is recorded for 463 FPUL, 369 IUP and 86 EP. [3] The maternal age is recorded for 232 FPUL, 191 IUP and 42 EP. [4] The initial progesterone is recorded for 229 FPUL, 188 IUP and 40 EP. [5] The initial progesterone is recorded for 292 FPUL, 150 IUP and 70 EP. [6] The initial progesterone is recorded for 148 FPUL, 73 IUP and 31 EP.

Table 5.11: Updating methods.

METHOD		Description
0	Reference	The original prediction model (i.e. M4) *without* any adjustments.
1	Recalibration	The original prediction where the *intercepts* are adjusted.
2		The original prediction model where *intercepts* and *(non-)corresponding slopes* for the linear predictors are estimated.
4	Revision	The *intercept* and *coefficients* of *the individual predictors* of M4 are re-estimated.
4P		METHOD 4 where the recalibrated linear predictors (METHOD 2) are considered as *offset* and *Ridge penalization* is used.
4L		The *intercept* and *coefficients* for *logarithm of average* hCG and *logarithm of* hCG *ratio* are (re-)estimated.
7	Extension	METHOD 4 where also *coefficients* for *age* and *logarithm of initial progesterone* are estimated.
7P		METHOD 7 where the recalibrated linear predictors (METHOD 2) are considered as *offset* and *Ridge penalization* is used.
7L		The *intercept* and *coefficients* for *logarithm of average* hCG, *logarithm of* hCG *ratio*, *age* and *logarithm of initial progesterone* are (re-)estimated.

METHOD 4 applies a straightforward refit using multinomial logistic regression:

$$\begin{cases} \text{lp}_{4,FI} = a_{FI} + b_{F1}\log(\text{hCGm}) + b_{F2}\ \text{hCGrc} + b_{F3}\ \text{hCGrc}^2 \\ \text{lp}_{4,EI} = a_{EI} + b_{E1}\log(\text{hCGm}) + b_{E2}\ \text{hCGrc} + b_{E3}\ \text{hCGrc}^2 \end{cases} \tag{5.6}$$

METHOD 4P uses the recalibrated linear predictors from METHOD 2 (i.e. $\text{lp}_{2,FI}$ and $\text{lp}_{2,EI}$) as an offset:

$$\begin{cases} \text{lp}_{4P,FI} = a_{FI} + \text{lp}_{2,FI} + b_{F1}\log(\text{hCGm}) + b_{F2}\ \text{hCGrc} + b_{F3}\ \text{hCGrc}^2 \\ \text{lp}_{4P,EI} = a_{EI} + \text{lp}_{2,EI} + b_{E1}\log(\text{hCGm}) + b_{E2}\ \text{hCGrc} + b_{E3}\ \text{hCGrc}^2 \end{cases} \tag{5.7}$$

Adding the linear predictors as offset implies that the changes in the intercepts and predictor coefficients with respect to METHOD 2 are modeled. We used Ridge penalization on these changes in order to prevent overfitting [53]. Without such penalization, METHOD 4 and 4P would yield identical results. METHOD 4L adapts the original prediction model by refitting using the logarithm of hCGr instead of hCGrc and the square of hCGrc:

$$\begin{cases} \text{lp}_{4L,FI} = a_{FI} + b_{F1}\log(\text{hCGm}) + b_{F2}\log(\text{hCGr}) \\ \text{lp}_{4L,EI} = a_{EI} + b_{E1}\log(\text{hCGm}) + b_{E2}\log(\text{hCGr}) \end{cases} \tag{5.8}$$

Given the small dataset on which M4 was originally developed, the quadratic effect may be exaggerated and better replaced by the log-transformation that is commonly used to model biomarkers [17].

Extension methods. The last three methods consider two additional markers, age and logarithm of the progesterone level at presentation in the hospital ($\log(prog)$), but are otherwise similar to METHOD 4, 4P and 4L. METHOD 7 fits a standard model with log(hCGm), hCGrc, hCGrc squared, age, and $\log(prog)$:

$$\begin{cases} \text{lp}_{7,FI} = a_{FI} + b_{F1}\log(\text{hCGm}) + b_{F2}\ \text{hCGrc} + b_{F3}\ \text{hCGrc}^2 + b_{F4}\ \text{age} \\ \qquad\qquad +b_{F5}\log(prog) \\ \text{lp}_{7,EI} = a_{EI} + b_{E1}\log(\text{hCGm}) + b_{E2}\ \text{hCGrc} + b_{E3}\ \text{hCGrc}^2 + b_{E4}\ \text{age} \\ \qquad\qquad +b_{E5}\log(prog) \end{cases} \tag{5.9}$$

METHOD 7P enters the linear predictors from METHOD 2 as offset and uses Ridge penalization to estimate (1) the changes in the intercept and the coefficients of

log(hCGm), hCGrc, and hCGrc squared, and (2) the coefficients for the two new predictors:

$$\begin{cases} \text{lp}_{7P,FI} & = a_{FI} + \text{lp}_{2,FI} + b_{F1}\log(\text{hCGm}) + b_{F2}\ \text{hCGrc} + b_{F3}\ \text{hCGrc}^2 \\ & \quad + b_{F4}\ \text{age} + b_{F5}\log(\textit{prog}) \\ \text{lp}_{7P,EI} & = a_{EI} + \text{lp}_{2,EI} + b_{E1}\log(\text{hCGm}) + b_{E2}\ \text{hCGrc} + b_{E3}\ \text{hCGrc}^2 \\ & \quad + b_{E4}\ \text{age} + b_{E5}\log(\textit{prog}) \end{cases} \tag{5.10}$$

METHOD 7L fits a model similar to METHOD 7 but using logarithm of hCGr instead of hCGrc and its square:

$$\begin{cases} \text{lp}_{7L,FI} & = a_{FI} + b_{F1}\log(\text{hCGm}) + b_{F2}\log(\text{hCGr}) + b_{F3}\ \text{age} \\ & \quad + b_{F4}\log(\textit{prog}) \\ \text{lp}_{7L,EI} & = a_{EI} + b_{E1}\log(\text{hCGm}) + b_{E2}\log(\text{hCGr}) + b_{E3}\ \text{age} \\ & \quad + b_{E4}\log(\textit{prog}) \end{cases} \tag{5.11}$$

Ridge penalization

METHOD 4P and 7P use Ridge (L2) penalization which shrinks coefficients of correlated predictors towards each other [53]. In these two methods the coefficients of the predictors are shrunken towards the offset, the linear predictors from METHOD 2. Instead of maximizing the log-likelihood function $l(\theta)$ with $\theta = [a\ b_1\ b_2\ ...\ b_n]^T$ from Equation 2.4, the following problem is solved [103]:

$$\max_\theta \quad l(\theta) - \lambda \sum_{j=1}^{n} b_j^2, \tag{5.12}$$

with λ the ridge parameter controlling the amount of shrinkage. In case of $\lambda = 0$, solving Equation 5.12 corresponds with ordinary maximum likelihood estimation. This regularization parameter λ is estimated using 10 fold cross-validation. Ridge penalization is implemented with the `glmnet` package in `R` [53].

Missingness

METHOD 7, 7P and 7L take two new predictors into account: age and the logarithm of initial progesterone. For `SGH` respectively 11 (0.8%) and 47 (3.3%)

women had a missing value for age and initial progesterone, while for QCCH 109 (12.5%) patients lacked an initial progesterone level. We dealt with this missingness by using single imputation [163]. Age and log(*prog*) are imputed via predictive mean matching regression that also includes the logarithm of hCG0, the logarithm of hCG48, and outcome [159, 188].

Performance evaluation: discrimination and calibration

The overall discrimination is evaluated using the polytomous discrimination index (PDI) [195, 196, 199]. In addition, conditional pairwise c-statistics are calculated [199]. For calibration assessment different measures are considered. Firstly, calibration-in-the-large, the difference between observed proportion and expected prevalence (i.e. average predicted risk), is evaluated [167]. Secondly, calibration intercepts and calibration slopes (with 95% CI) are assessed using a nominal logistic recalibration framework (see Section 4.2.1) [167, 206]. The calibration intercepts indicate whether the risks are systematically too high or too low (i.e. over- or underestimation), and their results should correspond to the calibration-in-the-large results. The calibration slopes indicate the presence of over- or underfitting. Ideally we expect calibration intercepts of zero and a calibration slope of 1. Thirdly, calibration is visualized via multinomial non-parametric calibration plots based on vector splines (see Section 4.2) [206, 229]. In these plots the predicted probabilities are plotted versus the observed probabilities.

The methods to assess calibration are based on the common definition of calibrated risks that states that estimated risks correspond to observed risks/proportions: for patients with an estimated risk of 0.3 for a particular event we expect 30% to have/develop this event. A stronger definition for calibration states that estimated risks correspond to observed risks per covariate profile, i.e. per combination of values for the predictors [144, 187] . It may occur that refitting has no apparent benefit on the common definition of calibration when compared to recalibration methods, but has benefit on "strong calibration". We evaluated this possibility through a likelihood ratio test of straightforward refitting (METHOD 4) vs recalibration (METHOD 2) using a chi-squared distribution. To find the number of degrees of freedom we have to compute the difference in degrees of freedom for METHOD 2 and 4. For METHOD 2, 2 intercepts, 2 corresponding and 2 non-corresponding slopes are estimated (i.e. 6 *df*), while for METHOD 4, 2 intercepts and $2n$ parameters are estimated, with n the number of considered predictors (i.e. $2n$+2 *df*). Comparing model revision, METHOD 4, with model recalibration, METHOD 2, therefore results in $2n$-4 *df*or $2 \cdot 3 - 4$ or 2 *df*. So we considered a likelihood ratio test using a chi-squared distribution with 2 degrees of freedom.

5.3.2 Results

Discrimination

Temporal updating. A decrease in PDI on the validation data compared to the updating data was seen for all methods, including the reference method (Table 5.12). All recalibration and revision methods except METHOD 1 (intercept adjustment) had a higher PDI than the reference method on the updating data, but could not preserve this advantage on the validation data. Only the extension methods resulted in a higher PDI on the validation data relative to METHOD 0. METHOD 4L and 7L, which use a log-transformation of hCG ratio, had a higher PDI than similar methods that use a quadratic effect for hCG ratio (METHOD 4, 4P, 7 and 7P). Overall, the best PDI was obtained for METHOD 7L. The added markers were useful to increase discrimination between IUP and EP, but less so for the discrimination between FPUL and IUP and between FPUL and EP (Table 5.13).

Table 5.12: Polytomous discrimination index on the updating and validation data.

	SGH		QCCH	
	Updating	Validation	Updating	Validation
METHOD 0	0.879	0.865	0.819	0.766
METHOD 1	0.871	0.858	0.818	0.768
METHOD 2	0.889	0.859	0.817	0.766
METHOD 4	0.888	0.859	0.824	0.764
METHOD 4P	0.889	0.859	0.821	0.767
METHOD 4L	0.892	0.866	0.798	0.735
METHOD 7	0.909	0.874	0.832	0.773
METHOD 7P	0.905	0.870	0.827	0.771
METHOD 7L	0.914	0.882	0.810	0.740

External updating. Again, recalibration and revision methods did not result in an increased PDI on the validation data relative to METHOD 0 (Table 5.12). In line with temporal updating, extension methods (METHOD 7 and 7P) had a higher PDI, although this increase was less pronounced. The two methods which consider the logarithm of hCG ratio (METHOD 4L and 7L) resulted in the poorest discrimination, worse than METHOD 0, suggesting that the log-transformation is not a good alternative. These methods were especially poor for discriminating between IUP and EP (Table 5.13). Overall, METHOD 7 obtained the best PDI. This method had the highest c-statistic for all three pairwise comparisons (Table 5.13).

Table 5.13: Conditional pairwise c-statistics on the updating and validation data.

		FPUL vs IUP		FPUL vs EP		IUP vs EP	
	Meth	Updating	Validation	Updating	Validation	Updating	Validation
Temporal updating (SGH)	0	0.98	0.99	0.92	0.91	0.90	0.86
	1	0.98	0.99	0.92	0.91	0.90	0.86
	2	0.99	0.99	0.92	0.93	0.93	0.87
	4	0.99	0.99	0.92	0.94	0.93	0.88
	4P	0.99	0.99	0.92	0.94	0.93	0.87
	4L	0.99	0.99	0.93	0.94	0.94	0.88
	7	0.99	1.00	0.93	0.94	0.95	0.90
	7P	0.99	1.00	0.93	0.94	0.94	0.90
	7L	0.99	0.99	0.93	0.94	0.96	0.92
		FPUL vs IUP		FPUL vs EP		IUP vs EP	
	Meth	Updating	Validation	Updating	Validation	Updating	Validation
External updating (QCCH)	0	0.96	0.93	0.92	0.90	0.84	0.82
	1	0.96	0.93	0.92	0.90	0.84	0.82
	2	0.97	0.94	0.94	0.92	0.85	0.79
	4	0.97	0.94	0.95	0.93	0.86	0.80
	4P	0.97	0.94	0.94	0.92	0.85	0.79
	4L	0.96	0.93	0.95	0.92	0.78	0.71
	7	0.96	0.95	0.95	0.93	0.86	0.81
	7P	0.97	0.94	0.95	0.93	0.85	0.79
	7L	0.96	0.94	0.95	0.92	0.79	0.73

Calibration

Temporal updating. The original model (Reference method) overestimated the risk of FPUL and low risks of IUP, as indicated by the calibration plot (Figure 5.4), calibration-in-the-large results (Table 5.14), and calibration intercepts that strongly deviated from zero (Figure 5.3, Table 5.15). The calibration slopes were below 1, suggesting overfitting as expected due to the small sample size on which M4 was developed. Hence intercept adjustments (METHOD 1) were insufficient as this does not improve the calibration slopes (Figure 5.3, Table 5.15). Simple recalibration (METHOD 2) already improved calibration drastically, and re-estimating the individual coefficients (METHOD 4 and 4P) did not yield further improvements (Figures 5.3 and 5.4, Table 5.15). Model extension (METHOD 7 and 7P) slightly improved intercepts and slopes, but also resulted in more variability (Figure 5.4). In the calibration plots of methods that used the quadratic effect of hCG ratio, we observed for some cases low predicted probabilities but high observed proportions of FPUL, and high predicted probabilities but low observed proportions of IUP (Figure 5.4). Using

a log-transformation for hCG ratio (METHOD 7L and 7L) resolved this issue (Figure 5.4).

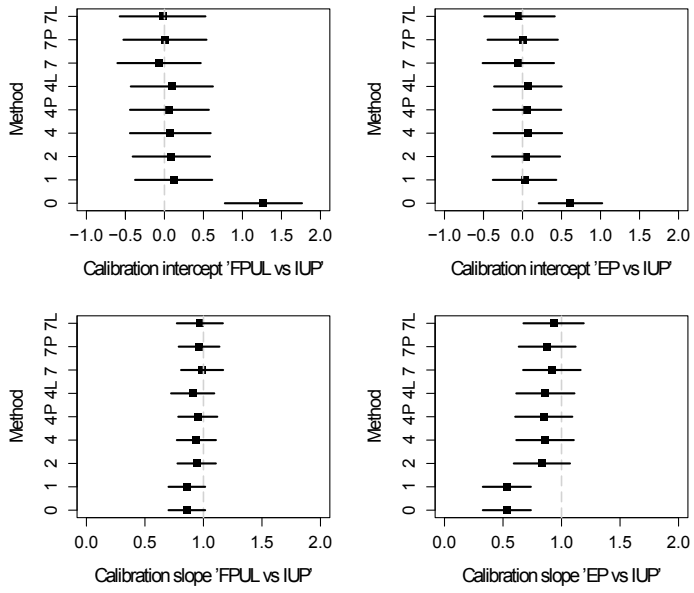
Table 5.14: Calibration-in-the-large (i.e. difference in observed prevalence and expected prevalence in percentage) on the validation data.

	Temporal updating (SGH)			External updating (QCCH)		
	FPUL	IUP	EP	FPUL	IUP	EP
METHOD 0	3.61	-5.20	1.59	0.67	-5.14	-4.47
METHOD 1	0.40	-0.36	-0.04	1.60	0.45	-2.06
METHOD 2	0.29	-0.29	-0.00	1.69	0.01	-1.70
METHOD 4	0.13	-0.32	0.20	1.59	0.12	-1.70
METHOD 4P	0.11	-0.29	0.18	1.66	0.02	-1.68
METHOD 4L	0.20	-0.35	0.15	1.56	-0.89	-0.67
METHOD 7	-0.14	0.25	-0.11	1.23	0.31	-1.54
METHOD 7P	0.02	-0.02	-0.01	1.51	0.08	-1.59
METHOD 7L	0.03	0.15	-0.18	1.10	-0.60	-0.50

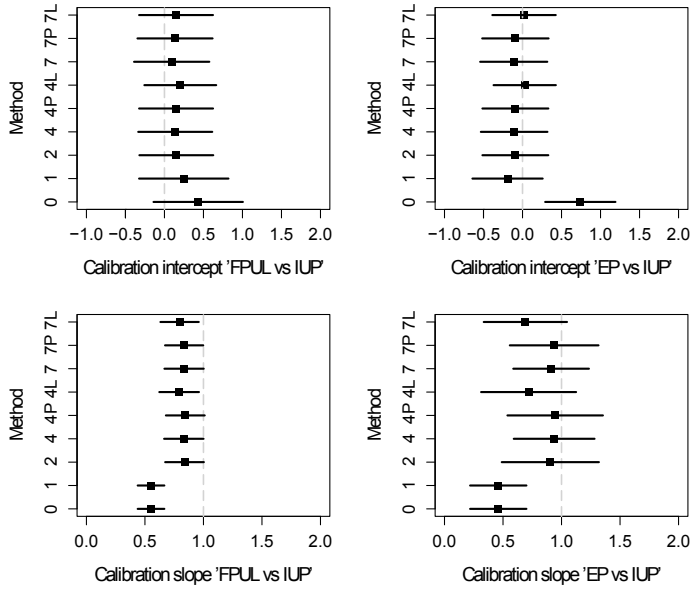
External updating. As with discrimination, results for calibration showed similarities between temporal and external updating. The original model overestimated the risk of IUP and low risks of FPUL on external validation, but now also underestimation of the risk of EP was observed (Figure 5.5, calibration-in-the-large in Table 5.14). Calibration intercepts differed clearly from 0, and calibration slopes were further from 1 on external validation compared to temporal validation (Figure 5.3, Table 5.15). Similarly to temporal updating, recalibration (METHOD 2) strongly improved calibration (Figures 5.3, 5.5 and Table 5.15). Model revision and extension did not provide further improvements.

Recalibration versus revision: strong calibration

The evaluation of calibration showed similar performance of recalibration through adjustments of intercepts and slopes (METHOD 2) and straightforward refitting (METHOD 4). For temporal updating, the log-likelihood was -286.83 for METHOD 2 and -281.75 for METHOD 4. A likelihood ratio test with 2 degrees of freedom lead to a p-value of 0.006. For external updating, log-likelihoods were -270.74 and -264.14, respectively, resulting in a p-value of 0.001. These results suggest that refitting on average improved predicted risks per covariate profile.

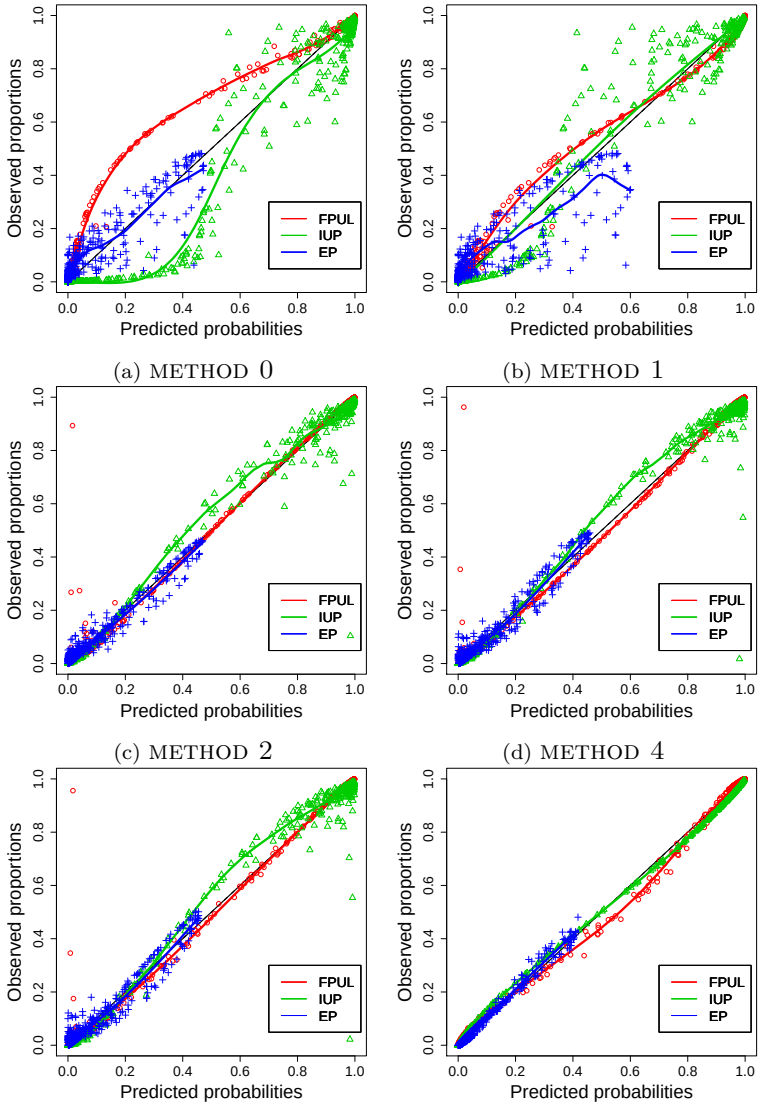


(a) SGH



(b) QCCH

Figure 5.3: Calibration intercept and slope (with 95% CI) on the validation data.

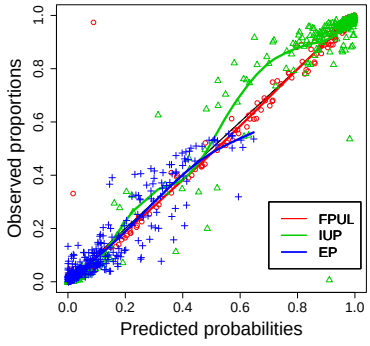


(a) METHOD 0 (b) METHOD 1

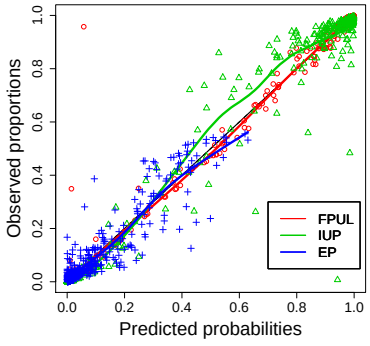
(c) METHOD 2 (d) METHOD 4

(e) METHOD 4P (f) METHOD 4L

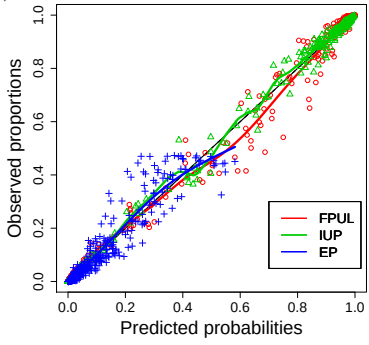
Figure 5.4: Non-parametric calibration plot for the different updating methods considering vector splines with 2 *df* for the validation data of the `SGH` case study. Cubic spline smoothers are added for each outcome category.



(g) METHOD 7



(h) METHOD 7P

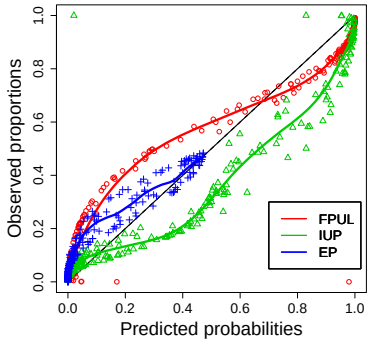


(i) METHOD 7L

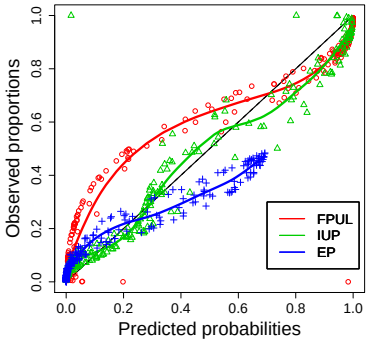
Continued Figure 5.4: Non-parametric calibration plot for the different updating methods considering vector splines with 2 *df* for the validation data of the `SGH` case study. Cubic spline smoothers are added for each outcome category.

Table 5.15: Calibration intercept and slope (with 95% CI) on the validation data.

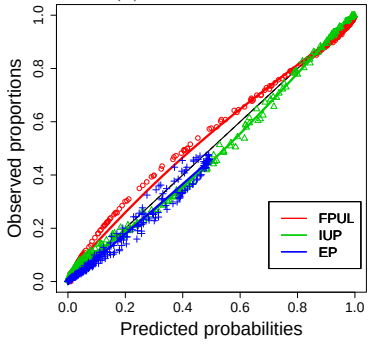
Temporal updating (SGH)	Calibration intercept (with 95% CI)		Calibration slope (with 95% CI)	
	FPUL vs IUP	EP vs IUP	FPUL vs IUP	EP vs IUP
METHOD 0	1.27 [0.77;1.76]	0.61 [0.21;1.02]	0.86 [0.70;1.01]	0.53 [0.33;0.74]
METHOD 1	0.12 [-0.38;0.61]	0.03 [-0.38;0.43]	0.86 [0.70;1.01]	0.53 [0.33;0.74]
METHOD 2	0.09 [-0.41;0.58]	0.05 [-0.39;0.48]	0.94 [0.78;1.11]	0.83 [0.59;1.07]
METHOD 4	0.08 [-0.44;0.59]	0.07 [-0.37;0.51]	0.94 [0.77;1.11]	0.86 [0.61;1.10]
METHOD 4P	0.06 [-0.44;0.57]	0.06 [-0.38;0.50]	0.95 [0.78;1.12]	0.85 [0.60;1.09]
METHOD 4L	0.09 [-0.43;0.62]	0.07 [-0.36;0.50]	0.91 [0.72;1.09]	0.86 [0.61;1.11]
METHOD 7	-0.07 [-0.60;0.46]	-0.05 [-0.51;0.40]	0.99 [0.81;1.17]	0.92 [0.67;1.16]
METHOD 7P	0.01 [-0.53;0.54]	0.00 [-0.45;0.45]	0.96 [0.79;1.14]	0.88 [0.63;1.12]
METHOD 7L	-0.02 [-0.57;0.53]	-0.04 [-0.49;0.35]	0.97 [0.77;1.17]	0.93 [0.67;1.19]
External updating (QCCH)	Calibration intercept (with 95% CI)		Calibration slope (with 95% CI)	
	FPUL vs IUP	EP vs IUP	FPUL vs IUP	EP vs IUP
METHOD 0	0.43 [-0.14;1.01]	0.74 [0.29;1.19]	0.55 [0.44;0.67]	0.46 [0.22;0.70]
METHOD 1	0.25 [-0.33;0.82]	-0.19 [-0.64;0.26]	0.55 [0.44;0.67]	0.46 [0.22;0.70]
METHOD 2	0.15 [-0.32;0.63]	-0.09 [-0.51;0.33]	0.84 [0.67;1.00]	0.90 [0.49;1.32]
METHOD 4	0.14 [-0.34;0.61]	-0.11 [-0.53;0.32]	0.83 [0.66;1.00]	0.94 [0.59;1.28]
METHOD 4P	0.15 [-0.32;0.62]	-0.10 [-0.52;0.33]	0.84 [0.68;1.01]	0.95 [0.54;1.30]
METHOD 4L	0.20 [-0.26;0.66]	0.03 [-0.37;0.43]	0.79 [0.62;0.96]	0.72 [0.31;1.12]
METHOD 7	0.09 [-0.39;0.58]	-0.11 [-0.54;0.32]	0.83 [0.67;1.00]	0.91 [0.59;1.23]
METHOD 7P	0.14 [-0.35;0.62]	-0.09 [-0.52;0.33]	0.84 [0.67;1.00]	0.94 [0.56;1.26]
METHOD 7L	0.15 [-0.32;0.62]	0.02 [-0.39;0.43]	0.80 [0.63;0.96]	0.69 [0.34;1.05]



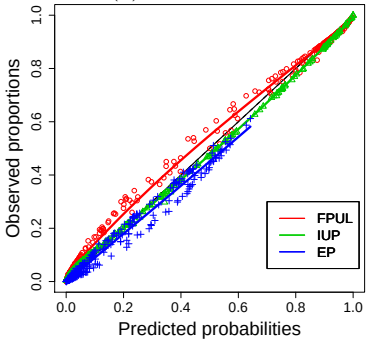
(a) METHOD 0



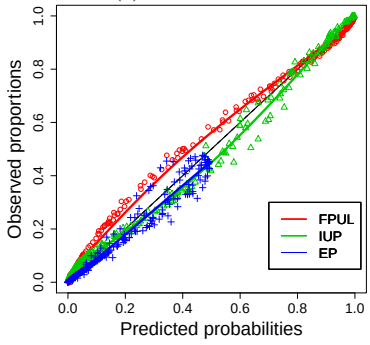
(b) METHOD 1



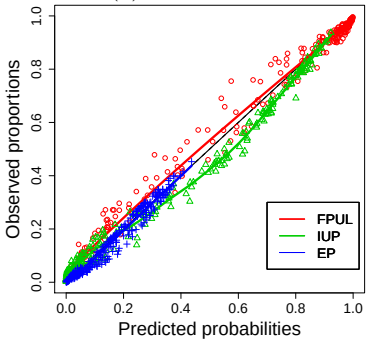
(c) METHOD 2



(d) METHOD 4

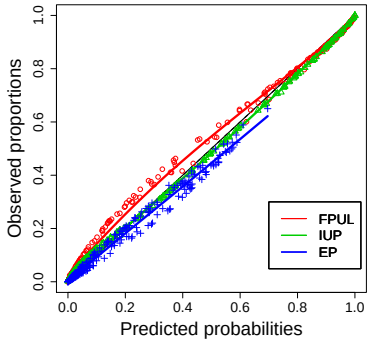


(e) METHOD 4P

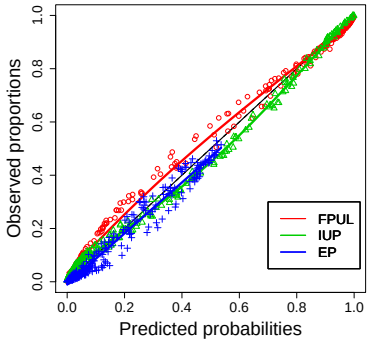


(f) METHOD 4L

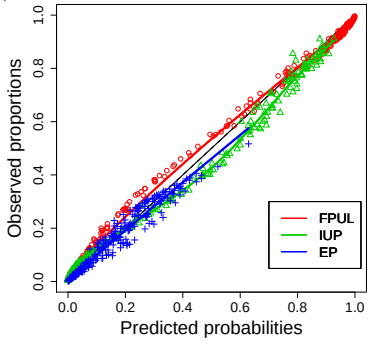
Figure 5.5: Non-parametric calibration plot for the different updating methods considering vector splines with 2 *df* for the validation data of the `QCCH` case study. Cubic spline smoothers are added for each outcome category.



(g) METHOD 7



(h) METHOD 7P



(i) METHOD 7L

Continued Figure 5.5: Non-parametric calibration plot for the different updating methods considering vector splines with 2 *df* for the validation data of the `QCCH` case study. Cubic spline smoothers are added for each outcome category.

5.3.3 Discussion

We adapted methods to update dichotomous risk prediction models towards polytomous risk models based on multinomial logistic regression. As a case study, the M4 model to predict the outcome of pregnancies of unknown location [29] was updated in time and in an external setting. M4 was developed on a small dataset, and uses the mean and ratio of hCG levels measured two days apart as predictors, and includes a quadratic effect of the hCG ratio. Eight updating techniques were considered and compared to the use of the original model without adjustments. The recalibration methods involve adjustment of the intercepts (METHOD 0) or of the intercepts and slopes (METHOD 2). The revision methods involve a straightforward refit of the individual coefficients (METHOD 4), a refit where changes with respect to the recalibrated coefficients from METHOD 2 are penalized (METHOD 4P), and a straightforward refit where the quadratic effect of the hCG ratio is replaced with a log-transformation (METHOD 4L) as the quadratic effect may be the result of overfitting. The three extension methods are similar to the revision methods but add two markers (age and the log-transformed progesterone level).

Conclusions for the case study on the M4 model

Revision methods were unable to improve discrimination compared to the original M4 model, but extension methods did. Considering a log-transformation for the hCG ratio instead of using a quadratic effect as in the original model improved discrimination for temporal updating but clearly decreased discrimination for external updating. The original M4 model was poorly calibrated in the temporal and external updating settings. The different updating methods improved the calibration of the original prediction model in a similar way, except for METHOD 1 in which only intercepts were adjusted. For the temporal updating setting, we noticed that changing the functional form of the hCG ratio resolved some clear discrepancies between predicted probabilities and observed proportions. Despite similar calibration curves for recalibration of intercepts and slopes compared to revision, likelihood ratio tests show that revision still improves predicted probabilities for individual patients over recalibration.

Differences between updating of dichotomous vs multinomial risk models

Some dissimilarities can be seen between dichotomous and nominal updating methods. For dichotomous outcomes, recalibration does not change the c-statistic. A linear transformation of the linear predictor does not change the

order of the predicted risks and consequently does not affect rank order statistics such as the c-statistic. A multinomial logistic model contains multiple linear predictors, one for each outcome category vs the reference category. As a result, recalibration methods can change the PDI (nominal c-statistic) and the conditional pairwise c-statistic. We would generally expect improvements when adjusting intercepts and slopes, which is indeed observed in our case study (Tables 5.12 and 5.13).

In addition, a polytomous calibration plot is more complex than a dichotomous one [206]. For the former we have one curve for each outcome category while for the latter a single curve is sufficient. Further, when assessing calibration for dichotomous outcomes using a parametric or non-parametric recalibration model such as loess or splines [7], there is a one-to-one relationship between predicted probabilities and observed proportions. This one-to-one relationship is lost in multinomial calibration assessment of risk models for nominal outcomes [206]. Therefore additional cubic smoothing splines are used in the calibration plots in order to visualize the overall relationship between predicted risks and observed proportions.

In the logistic recalibration framework for nominal risk models with k outcome categories we have k-1 calibration intercepts and k-1 calibration slopes. Recalibration by intercept and slope adjustments aim to obtain calibration intercepts that equal 0 and calibration slopes that equal one. This complicates recalibration methods compared to dichotomous risk models where there is one calibration intercept and slope. In our case study we observe that recalibration or more advanced updating methods yield calibration slopes that lie between 0.8 and 1 on the validation data. This may suggest that larger sample sizes are needed to obtain calibration slopes that are very close to one for nominal versus dichotomous outcomes. For clinical applications, the calibration plot is even more pivotal for nominal compared to dichotomous outcomes. The plot allows optimal interpretation of calibration as opposed to a collection of calibration intercepts and slopes. [206].

Choice of updating method

Calibration, defined as the correspondence between predicted probabilities and observed proportions, can often be improved with a simple intercept and/or slope adjustment. This might often be the case when updating a prediction model that was originally based on a small sample. In these situations, for example our case study, intercepts-only adjustments will probably be insufficient as it is highly likely that the original model is overfitted. If strong calibration is desired, where predicted probabilities correspond to observed proportions per

covariate profile, revision methods should generally be preferred if the sample size available for updating is sufficiently large [187, 206].

In research on updating methods for risk models, the functional form or optimal transformation of continuous predictors [63] receives little attention. However, transformations used in the original model may not hold for every setting in which the model may be used. Settings will for example vary with respect to the homogeneity of the patient population, or the transformation used in the original model may be the result of overfitting. In the temporal updating setting of our case study the quadratic term for the hCG ratio used in the original model was inferior to a log-transformation, which is rather common for biomarkers [17–19].

Limitations

We were confronted with missing values in our data. We only included women where we could apply M4. This means women were excluded if the pregnancy outcome was unknown (lost to follow-up), or if one or both hCG measurements were unavailable. For the added markers used in the extension methods (age and progesterone), missing values were singly imputed with predictive mean matching. An alternative would be to use multiple imputation for all missing values in order to avoid bias and maximize sample size [216]. We chose pragmatic solutions to deal with missing values to keep a clear focus. Another potential limitation was that we randomly divided data into updating and validation data with stratification based on the outcome. This allowed us to construct calibration plots, as opposed to using repeated data splitting, cross-validation, or bootstrapping. Results may depend on the obtained split due to the modest sample sizes, but we expect that this dependency is limited. Nevertheless, lower sample sizes may more easily result in lower calibration slopes for nominal as opposed to dichotomous outcomes.

5.3.4 What is new?

Key findings

- We adapted updating methods to recalibrate, revise, or extend risk models for dichotomous outcomes to models for polytomous outcomes based on multinomial logistic regression. We successfully applied these to a case study about the M4 model to predict pregnancy location.
- The M4 model was developed on a small dataset, hence we did not expect to observe that recalibration did very well compared to revision. Only

some extension methods were able to improve discrimination, whereas recalibration, revision, and extension methods improved risk calibration. Revision performed better than recalibration to obtain calibrated risks per covariate profile.

- Altering the functional form of a predictor improved performance in some settings but not in others.

What this research adds to what was known?

- There is growing attention for methods to update prediction models for use in new settings or situations, however only methods for dichotomous outcomes existed. This study shows how models for multinomial outcomes can be updated using similar techniques.
- Recalibration can work well even when the original model was based on a small dataset.
- Recalibration may perform as well as revision for calibration on the general level, but still be inferior for calibration per covariate profile.

Implications of this research

- When a multinomial risk prediction model has suboptimal performance when validated on a different population, multinomial updating techniques should be considered. The functional form of continuous predictors deserves more attention during the updating process if sample size permits.

5.4 Conclusion

From the first part of this chapter we can conclude that simple dichotomous updating methods work well for models that predict multinomial outcomes and that are made up of a sequence of dichotomous models. Recalibration can easily be done for sequential dichotomous models and will be the recommended approach when overall sample size for the updating data is small or when at least one of the outcome categories has a relatively low prevalence. However, if sufficient data are available, more complex updating methods such as revision, extension, or model redevelopment may become valid alternatives.

Due to the promising results we conducted research concerning the extension of dichotomous updating methods to prediction models based on multinomial

logistic regression. From this research, in the second part of this chapter, we can conclude that updating methods for dichotomous risk models can be successfully adapted to multinomial risk models. Simple recalibration methods may still work very well even if the original prediction model was based on a small sample. If the available sample size is large, revision methods may still be preferred to better tailor predictions to individual patients or covariate profiles.

Chapter 6

Calibration of multinomial or multiclass risk predictions

In the previous Chapters we covered dichotomous logistic regression (cf. Chapter 2), multinomial logistic regression (cf. Chapter 3) and multinomial calibration tools (cf. Chapter 4). In this Chapter we present tools to assess the calibration of dichotomous or binary and multinomial or multiclass risk models irrespective of the adopted algorithm. The terminology dichotomous and multinomial is common for logistic regression, while in the Machine Learning world mostly binary and multiclass are used. In this Chapter we will mainly use binary and multiclass. We propose the general calibration framework based on a vector spline multinomial logit model (cf. Chapter 4). This framework can be used to generate calibration plots as well as a measure that quantifies the lack of calibration. This Chapter is based on a manuscript in preparation.

6.1 Introduction

In medical applications, prediction models that provide probabilistic (risk) estimates of an event of interest are highly relevant for clinical decision support, personalized healthcare, and shared decision making. As mentioned earlier in this dissertation, prior to the implementation of such tools in clinical practice, validation with respect to discrimination and calibration is required [6, 97, 135, 167, 186]. A model needs to be able to distinguish the outcome categories, a performance issue that can be evaluated by the c-statistic or multiclass

extensions thereof (cf. Chapter 2 and 3). In addition, however, the reported risk estimates need to be reliable. For binary outcomes, the relationship between predicted and observed probabilities can be visualized by means of a calibration plot [167, 206].

In Chapter 4 binary calibration plots are extended towards multiclass models based on multinomial logistic regression (MLR) [206]. A parametric and non-parametric approach were proposed. Logistic regression is a common algorithm to build binary and multiclass clinical prediction models in the medical literature, which works with risk estimates. However, machine learning algorithms are also used in this context [29, 44, 92, 96, 107, 202] and are very prominent in high dimensional and/or "large p, small n" prediction studies (i.e. large number of predictors and small number of patients) [38, 101, 157]. Moreover, although machine learning approaches for classification problems are often less suited for or concerned with probability estimation, methods do exist [85,99,100,111,191,234]. Calibration performance of risk models is all too often neglected, and it is not a surprise that this holds in particular for models based on machine learning algorithms notwithstanding a few notable exceptions [84, 85, 99, 100, 112, 176].

The aim of this Chapter is to propose the non-parametric calibration framework to evaluate calibration of multiclass risk prediction models irrespective of the used modeling technique. Based on this framework we also derive an ad hoc calibration index to easily quantify and compare calibration performance between models. We use risk predictions based on logistic regression, support vector machines, `k`-nearest neighbors, random forest, naive Bayes and nearest shrunken centroids [4, 21, 60, 178, 179, 212, 213]. Multiclass probabilities can be based on direct multiclass (all-at-once) methods or on combinations of binary risk models such as nested dichotomies [52] and pairwise coupling [68] (cf. Chapter 3). The case study on the diagnosis of ovarian tumors is used to illustrate the methods.

6.2 Patients and methods

6.2.1 Study set-up

We captured and visualized the calibration of various approaches of dichotomous and multinomial logistic regression and binary and multiclass machine learning methods on a case study on ovarian tumors. As a starter, binary models distinguishing benign and malignant tumors are developed and validated. Afterwards, we discuss the extension towards multiclass problems and distinguish benign, borderline, stage I invasive, stage II-IV invasive and metastatic tumors.

As explained in Chapter 4, the assessment of calibration for multiclass problems is more complex. The prediction models were derived on a development set and afterwards validated on a separate validation set. The data-split, with 1:1 ratio, was random with stratification according to the multiclass outcome.

6.2.2 Case study

For this research, the same data as for the application in Chapter 4 were considered. The accurate diagnosis of ovarian tumors prior to surgery is crucial for the choice of appropriate management and correct referral. Given that different types of malignancies require different management, we aimed to develop a model to predict the risk that an ovarian tumour is benign, borderline malignant, stage I cancer, stage II-IV cancer, or secondary metastatic disease. The 5909[1] patients from the IOTA consortium (see Section 1.4.1) were randomly divided into a development and validation set stratified for the outcome [197]. The following predictors were considered without further selection: age (years), serum CA125 (U/ml), oncology referral center (yes/no), maximum diameter of the lesion (mm), proportion of solid tissue (between 0 and 1), presence of more than 10 cyst locules (yes/no), number of papillary projections (0, 1, 2, 3 and more than 3), presence of acoustic shadows (yes/no) and presence of ascites (yes/no) (Table 6.1) [197]. The proportion of solid tissue equals the ratio of the maximum diameter of the largest solid component and the maximum diameter of the lesion. These nine variables are the predictors from the multiclass risk prediction model ADNEX (see Section 4.7) [197]. Some values for serum CA125 were missing and we dealt with this for the present study by using a single imputation [163]. We used predictive mean matching regression using variables that were either related to the level of CA125 itself, or the unavailability of CA125 via an indicator whether the CA125 was missing or not [159, 163].

6.2.3 Performance evaluation

The ability to distinguish between different outcome categories (discrimination) as well as the correspondence between predicted risks and observed proportions (calibration) is assessed.

[1]From the original 5914 women, 5 cases were omitted based on logistic regression influence diagnostics [79] and further data review.

Table 6.1: Descriptive statistics of the ovarian tumor case study (N=5909).

	Benign	BOT	Stage I	Stage II-IV	Metastatic
Outcome, N	N=3980	N=339	N=356	N=988	N=246
Variable, N(%) or median (IQR)					
Age (years)	42 (32-54)	49 (36-62)	54 (44-64)	59 (50-67)	57 (47-68)
CA125*	19 (11-39)	31 (16-100)	52 (21-190)	447 (147-1215)	81 (30-271)
famhistovca	79 (2.0)	10 (3.0)	13 (3.7)	57 (5.8)	5 (2.0)
lesdmax	63 (45-87)	86 (51-150)	106 (71-153)	85 (56-123)	86 (56-124)
Solid tissue					
- Presence of solid tissue	1322 (33.2)	267 (78.8)	328 (92.1)	968 (98.0)	234 (95.1)
- Propsol if present	42 (20-100)	37 (24-59)	61 (38-100)	100 (56-100)	100 (64-100)
Papnr					
- 0	3424 (86.0)	135 (39.8)	227 (63.8)	772 (78.1)	213 (86.6)
- 1	333 (8.4)	69 (20.4)	25 (7.0)	56 (5.7)	12 (4.9)
- 2	80 (2.0)	21 (6.2)	17 (4.8)	30 (3.0)	0 (0)
- 3	66 (1.7)	24 (7.1)	17 (4.8)	28 (2.8)	2 (0.8)
- >3	77 (1.9)	90 (26.5)	70 (19.7)	102 (10.3)	19 (7.7)
Locules10	199 (5.0)	74 (21.8)	69 (19.4)	93 (9.4)	36 (14.6)
Shadows	676 (17.0)	8 (2.4)	18 (5.1)	30 (3.0)	10 (4.1)
Ascites	64 (1.6)	28 (8.3)	65 (18.3)	473 (47.9)	90 (36.6)
Oncology referral center	2030 (51)	242 (71)	258 (72)	801 (81)	205 (83)

* The results for CA125 are based on single imputation of missing values. The CA125 level is missing for 1447 (36.4%) benign tumors, 62 (18.3%) borderline tumors, 71 (19.9%) stage I, 163 (16.5%) stage II-IV invasive tumors and 62 (25.2%) metastatic tumors.

Abbreviations. BOT, borderline tumor; IQR, interquartile range.

Discrimination

The discrimination of a binary risk prediction model is evaluated via the c-statistic (see Section 2.3.1). For the multiclass risk prediction models, multiple measures are used. First of all, c-statistics based on the conditional risk for pairs of categories or conditional pairwise c-statistics are calculated; furthermore, the multiclass measures M-index and polytomous discrimination index (PDI) are evaluated (see Section 3.3) [61, 196, 199]. The former corresponds with the average over all pairwise c-statistics and is often used in machine learning [61]. The pairwise c-statistic for categories A and B is obtained by averaging the c-statistic based on π_A and the c-statistic based on π_B which are different since $\pi_A + \pi_B \neq 1$ [61]. The PDI is a nominal version of the c-statistic and estimates the average proportion of correctly identified cases for a group of cases each belonging to a different category [196]. The M-index and PDI in case of random performance for k outcome categories are respectively 0.5 and $1/k$ (i.e. 0.2 if k=5) [196].

Calibration

As an overall measure the calibration-in-the-large or mean calibration, i.e. the difference between the observed prevalence and the average predicted probability, was computed. For the assessment of the calibration of the individual risk predictions we used the non-parametric recalibration framework explained in Section 4.2.2. The non-parametric calibration plots are obtained by plotting the predicted probabilities $\widehat{\pi}_i$ versus the observed probabilities $\widehat{o}_i$ based on this non-parametric framework for each outcome category i $(i = 1, \ldots, k)$ [206]. Since the multiclass calibration plots lack a one-to-one relationship when plotted for each outcome category (see Section 4.5), spline smoothers are plotted to assess the trend of the scatter plot for each category [206]. Finally, the average squared difference of the predicted probabilities $\widehat{\pi}_i$ and the observed probabilities $\widehat{o}_i$ for each patient j and outcome category i:

$$\frac{\sum_{j=1}^{N}\sum_{i=1}^{k}\left(\widehat{\pi}_{ji}-\widehat{o}_{ji}\right)^2}{N\cdot k}, \tag{6.1}$$

with N the number of patients and k the number of outcome categories was calculated. If we multiply this average squared difference by $100 \cdot k/2$ we obtain an ad hoc measure that we hereafter refer to as the Estimated Calibration Index (ECI) because it is based on estimation of the observed proportions $\{\widehat{o}_1, ..., \widehat{o}_k\}$. The multiplication factor $100 \cdot k/2$ ensures that ECI has a theoretical range between 0 and 100, because $2/k$ is the theoretical maximum of the average squared difference. This is derived as follows. The (theoretically) worst case scenario for a patient occurs when $\widehat{\pi}_j = 1$ but $\widehat{o}_{i,i\neq j} = 1$. Hence $\sum_{i=1}^{k}\left(\widehat{\pi}_i - \widehat{o}_i\right)^2 = 2$ for this patient. If we have the worst case scenario for all patients, the average squared difference is $(2 \cdot N)/(N \cdot k) = 2/k$. This ECI is related to the Brier score (see Section 2.3.1), i.e. the squared difference between the actual dichotomous outcome y_i and the predicted probabilities $\widehat{\pi}_i$: $\sum_{i=1}^{N}\left(y_i - \widehat{\pi}_i\right)^2/N$ [54].

6.2.4 Logistic regression models

Dichotomous as well as multinomial logistic regression models were considered (cf. Chapter 2 and 3). The nine predictors, from Section 6.2.2, were used for model building without variable selection, variable transformation or interaction effects. For a real clinical application, model specification is crucial and a blind application of the logistic regression model $Y \sim \mathrm{x}_1 + ... + \mathrm{x}_n$ is probably not realistic [173].

6.2.5 Machine learning models

In this Section, an overview of the considered machine learning models is given. Similarly as for logistic regression, the nine considered predictors were included without any transformation or selection. The choice of hyperparameters, used in the different models, is discussed in Section 6.2.8.

Support vector machines (SVM)

The first considered machine learning method is support vector machines [1,98,160,212,213,227]. Suppose that we have a binary outcome y and predictors $\mathbf{x}$. The primal problem of separating both outcome categories is converted in an equivalent dual problem. The goal of the dual problem is finding a hyperplane that separates both outcome categories. The challenge is finding a trade-off between a maximal margin between both outcome categories and minimal misclassification. The former can result in a too simplistic model while the latter may result in an overfitted model. The regularization parameter C, with default value one, indicates the trade-off between maximization of margin and minimization of misclassification [160]. A low C means that misclassification is less important and results in a simple model (large margin), while a high C indicates that misclassification should be avoided as much as possible and implies a complex model (narrow margin). SVMs use a kernel trick to implicitly map the input space $\mathbf{x}$ to a high-dimensional feature space. Different kernels are possible and the choice influences the type of classifier (i.e. linear or non-linear) in the input space. We considered the popular RBF kernel, i.e. the ("Gaussian") radial basis function:

$$K(\mathbf{x}_a, \mathbf{x}_b) = exp\left(\frac{-\|\mathbf{x}_a - \mathbf{x}_b\|^2}{2\sigma^2}\right) \tag{6.2}$$

with σ a hyperparameter [160, 213]. This kernel results in a non-linear classifier. Probabilities for the SVMs were obtained via a parametric sigmoid function which is suggested by Platt [149].

`k` **Nearest Neighbors** (kNN)

This non-parametric algorithm starts from the principle that the `k` nearest neighbors (i.e. cases with similar predictor values $\mathbf{x}$) contain useful information concerning the outcome category for a new case with predictor values $\mathbf{x}_a$, with

k a hyperparameter [43, 162]. The Euclidean distance

$$d(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{\sum_{l=1}^{n} (\mathrm{x}_{al} - \mathrm{x}_{bl})^2}, \tag{6.3}$$

with n the number of considered predictors, is used to find the k closest cases. In case of ties all cases with the kth largest distance are included. The outcome category of each of these k nearest neighbors is considered and results in a probability. The probabilities correspond with the proportion of votes for each outcome category among the neighbors.

Random Forest (RF)

Breiman defined a random forest as "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [21]. In other words, a random forest is a set of classification or regression trees where at each split a certain fixed number, i.e. hyperparameter *mtry*, of randomly sampled candidate variables were considered [21, 26, 69, 99, 161, 175]. Five hundred different classification trees were grown using bootstrap samples of the development data. Each tree was grown until purity, i.e. until no additional splits were possible [111].The class probabilities based on this random forest, i.e. probabilities averaged over all trees, were computed. Note that since the trees are grown until purity the probability of each tree is either zero or one for a single observation.

Naive Bayes (NaB)

With Naive Bayes the conditional posterior probabilities of each outcome category Y_i, given a set of predictor variables $\mathbf{x}$, are obtained via Bayes' rule:

$$P(Y_i|\mathbf{x}) = \frac{P(Y_i)P(\mathbf{x}|Y_i)}{\sum_{j=1}^{k} P(Y_j)P(\mathbf{x}|Y_j)}, \tag{6.4}$$

with k the number of outcome categories and $i = 1, ..., k$. This modeling technique is a supervised learning algorithm in which the predictors are considered conditionally independent, i.e. $P(\mathbf{x}|Y_i) = \prod_{j=1}^{n} P(\mathrm{x}_j|Y_i)$, although this is not required according to the literature [72]. Rish stated that despite the unrealistic independence assumption, the NaB models may often correctly classify patients even if the probability estimates are inaccurate [153]. This NaB approach has a hyperparameter that identifies whether a kernel or a normal density estimate, $useKernel$, should be considered for $P(\mathbf{x}|Y_i)$.

Nearest Shrunken Centroids (NSC)

The last considered machine learning method is nearest shrunken centroids [179]. Suppose we have N cases with each a predictor vector $\mathbf{x} = [x_1 \ ... \ x_n]^T$. Each case belongs to one of the k outcome categories. For each class the centroid, i.e. the vector with standardized average predictor values for cases belonging to that class can be computed. In the same way the overall centroid-components, i.e. the standardized average predictor value for all cases, can be obtained. With nearest shrunken centroids, each of the class centroids is shrunk towards the overall centroid by a certain threshold Δ, i.e. a hyperparameter. Suppose we consider only one predictor x_1 and have an overall centroid of μ and the chosen threshold is Δ. A class centroid in the interval $[\mu - \Delta; \mu + \Delta]$ will be shrunk towards μ, a class centroid larger than $\mu + \Delta$ will be diminished by Δ, while for a class centroid smaller than $\mu - \Delta$, Δ will be added. The test cases are classified to the closest shrunken centroid based on the standardized squared distance corrected for the class prevalences [179]. Estimates of the class probabilities, based on the corrected distances to the k shrunken centroids, were computed [179].

6.2.6 Multiclass prediction models and combinations of binary prediction models

Multiclass prediction models can either be a combination of binary prediction models or an all-at-once model. Among the implemented algorithms, only the SVMs cannot give all-at-once multiclass risks. When binary models are combined, two types of combinations are used: a sequential dichotomous model or binary tree and pairwise coupling (see Section 3.2.2 and 3.2.3) [52, 68]. For the former a tree of binary prediction models is constructed. In case of 5 different outcome categories 105 different trees can be constructed of which we will consider three (see Section 3.2.2). In our case study, from a clinical point of view, the first two dichotomies of the binary tree are straightforward: first benign versus malignant, then the malignant tumors are divided into borderline and invasive tumors. The category of invasive tumors has three subgroups: stage I invasive, stage II-IV invasive and metastatic cancer. Since it is not obvious what the most clinically relevant binary tree is for these three categories we constructed the three possible sequences and considered the average (Figure 6.1). For pairwise coupling, the pairwise probabilities $\widehat{\pi}_{ij}$ of every pairwise prediction model are combined in order to obtain multiclass probabilities $\widehat{\pi}_i$ with $(i = 1, ..., k)$ by

solving the following system [224]:

$$\pi_i = \sum_{j=1, j\neq i}^{k} \frac{\pi_i + \pi_j}{k-1} \widehat{\pi}_{ij} \quad \forall j \text{ with } \sum_{i=1}^{k} \pi_i = 1, \pi_i \geq 0. \tag{6.5}$$

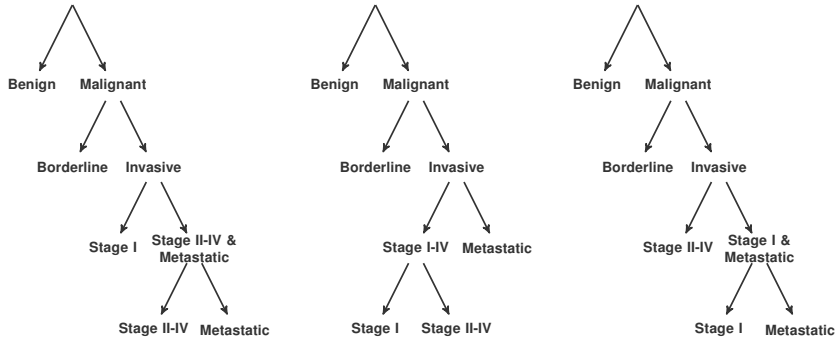


Figure 6.1: The three different considered sequential dichotomous models.

6.2.7 Overview of the implemented approaches

The binary risk models to predict the overall risk of malignancy were obtained via six different approaches: logistic regression and the five machine learning models introduced in Section 6.2.5. For the multiclass risk predictions nine different approaches were applied, of which three logistic regression methods and seven machine learning methods (Table 6.2). For logistic regression, all-at-once multinomial logistic regression (MLR), binary tree logistic regression (LR-BT) and pairwise coupling logistic regression (LR-PC) were used. For support vector machines, a binary tree (SVM-BT) and a pairwise coupling (SVM-PC) approach were applied. The remaining four approaches were `k` Nearest Neighbors, Random Forest, Naive Bayes and Nearest Shrunken Centroids, all of which directly provide multiclass risk estimates.

6.2.8 Tuning of hyperparameters

The optimal hyperparameter σ of the RBF kernel used in the support vector machine models was obtained via automatic sigma estimation [23, 77]. To

Table 6.2: Overview of the implemented multiclass approaches.

all-at-once	binary tree	pairwise coupling
MLR	LR-BT	LR-PC
-	SVM-BT	SVM-PC
kNN	-	-
RF	-	-
NaB	-	-
NSC	-	-

tune the remaining hyperparameters of the machine learning models, a grid of parameters was specified and each model was fitted (Table 6.3). Based on these models resampling based performance measures were calculated. Ten repeats of 10-fold cross-validation were considered. For 10-fold cross-validation the data are split into 10 folds and the model is fitted based on 90% of the cases and validated on the remaining 10% [69]. Each of the 10 folds is once considered as test sample and the other nine times part of the training sample. For the two approaches that were combinations of binary machine learning models (i.e. SVM-BT and SVM-PC), optimal tuning parameters were determined separately for each binary model. The choice of selection criterion for the hyperparameters can have a significant impact on the final performance of the model [112]. Two different optimal values for the tuning of hyperparameters were chosen based on either the accuracy (maximization) or the logarithmic loss (minimization). The latter will be abbreviated as logloss. Because logloss is related to both discrimination and calibration [112], we expected to see better calibration performance when logloss was used compared to accuracy.

Table 6.3: Machine learning models with their respectively considered grid for the different hyperparameters.

Method	Hyperparameter	Grid
Support Vector Machines*	C	$10^{-3:3\diamond}$
k Nearest Neighbors	k	1:50
Random Forest	$mtry$	1:9
Naive Bayes	$useKernel$	TRUE/FALSE
Nearest Shrunken Centroids	Δ	0:1:0.01 and 2:10

* For each binary support vector machine model the hyperparameter C is tuned.

$^\diamond$ The grid $10^{-3:3}$ corresponds with $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ or $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. For the binary prediction model stage II-IV versus metastatic the considered grid is $10^{-3:2} = \{0.001, 0.01, 0.1, 1, 10, 100\}$.

6.2.9 Software

For the logistic regression and machine learning models we made use of the following R-packages: `rms`, `vgam`, `klaR`, `pamr`, `randomForest`, `kernlab` and `caret`. The function `sigest()` of the R-package `kernlab` is used for automatic sigma estimation, while the functions `train()` and `trainControl()` of the R-package `caret` were used for the determination of optimal tuning parameters.

6.3 Results

6.3.1 Binary outcome

Hyperparameters

The optimal hyperparameter values based on the maximization of the accuracy and the minimization of the logloss were different for SVM regularization parameter C and for the number of neighbors for kNN. Concerning the cost function C, maximal accuracy was obtained for $C = 10$ and minimal logloss for $C = 1$. For kNN, 20 neighbors were used based on accuracy, whereas 50 were used based on logloss. Minimizing the logloss resulted in a simpler model, i.e. misclassification is less important resulting in a larger margin, and more neighbors, i.e. more cases are considered to obtain the probabilities, compared with maximizing accuracy. This implies less overfitting and smoother results.

Discrimination

The validation c-statistics indicated good discrimination for all approaches (Table 6.4). kNN and NSC had the lowest discriminative ability, whereas for logistic regression, random forest and support vector machines the best discriminative potential was observed. For SVM and kNN, with different optimal hyperparameter values based on accuracy and logloss, the best validation performance was obtained for logloss, even though the discriminative ability of the development data was better in case of accuracy.

Calibration

Mean calibration of the risk of malignancy was close to zero for logistic regression, overall it ranged between 0.25% (logistic regression) and 3.93% (naive bayes) on

Table 6.4: Discrimination (c-statistic) and calibration (mean calibration (in %) of the risk of malignancy and estimated calibration index (ECI)) for the binary prediction models for the validation data.

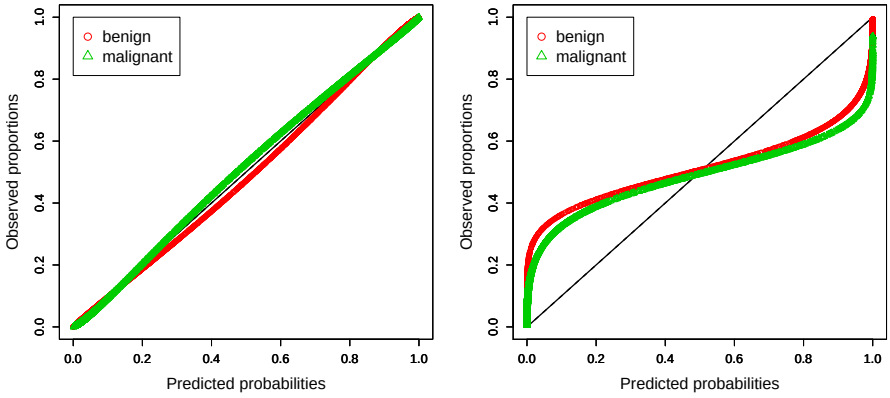
Model (optimization)	Discrimination	Calibration	
	c-statistic	mean calibration	ECI
LR	0.938	0.25	0.01
SVM (accuracy)	0.940	0.94	0.09
SVM (logloss)	0.945	0.40	0.04
kNN (accuracy)	0.860	1.80	0.61
kNN (logloss)	0.862	1.92	0.05
RF (accuracy)	0.948	1.64	0.05
RF (logloss)	0.948	1.64	0.05
NaB (accuracy)	0.910	3.93	1.67
NaB (logloss)	0.910	3.93	1.67
NSC (accuracy)	0.890	2.47	0.45
NSC (logloss)	0.890	2.47	0.45

the validation data suggesting slight underestimation of the risk of malignancy on average (Table 6.4). Most models showed an ECI close to zero, exceptions were the kNN model with accuracy as optimization method, the NaB and the NSC model. The two k nearest neighbor models had different ECI, 0.61 versus 0.05, although their mean calibration for the risk of malignancy was similar, 1.80% vs 1.92% (Table 6.4). The calibration plots of logistic regression (lowest ECI) and naive bayes (highest ECI) are shown in Figure 6.2. The optimization method logloss, resulted in the lowest ECI of 0.05 and the best calibration plot (Figure 6.3). Similarly, minimizing logloss for SVM resulted in a lower ECI compared with maximizing accuracy although the difference is less explicit.

6.3.2 Multiclass outcome

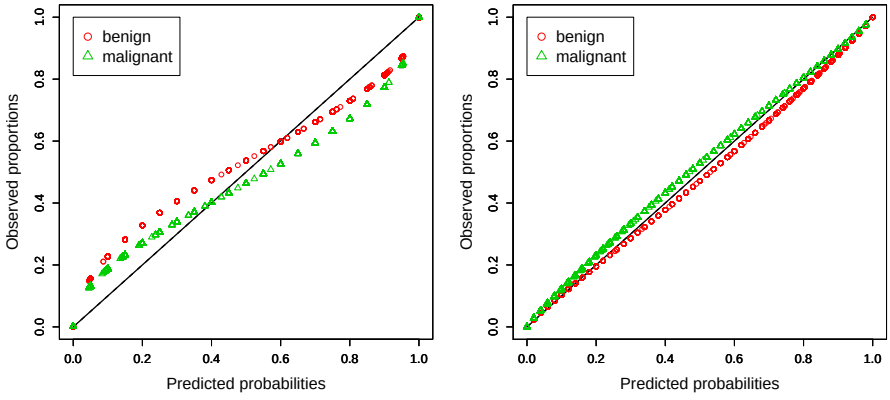
Hyperparameters

Only for one of the seven multinomial machine learning approaches, i.e. random forest, the hyperparameters obtained by maximizing the accuracy and minimizing the logloss were the same.



(a) LR (b) NaB

Figure 6.2: Calibration plot of the validation data for (a) dichotomous logistic regression and (b) binary naive bayes.



(a) accuracy (b) logloss

Figure 6.3: Calibration plot of the validation data for the binary `k` nearest neighbors with (a) maximizing accuracy as selection criterion and (b) minimizing logloss as selection criterion.

Discrimination

The discriminative ability to distinguish the multiclass categories as well as pairs of categories was assessed (Table 6.5 and 6.6). The best multiclass c-statistics were obtained for logistic regression models and random forest (Table 6.5). Combining dichotomous logistic regression models resulted in better discriminative potential than multinomial logistic regression (Table 6.5). The multiclass approach with the worst multiclass discrimination was kNN (Table 6.5). Regarding the conditional pairwise probabilities, we noted that the most difficult categories to distinguish were borderline and stage I invasive, and the different invasive categories (i.e. stage I invasive, stage II-IV invasive and metastatic cancer, Table 6.6).

Table 6.5: Discrimination (M-index and polytomous discrimination index PDI) and calibration (estimated calibration index ECI) for the multiclass prediction models for the validation data.

Model (optimization method)	Discrimination		Calibration
	M-index	PDI	ECI
MLR	0.790	0.529	0.27
LR-BT	0.798	0.542	0.30
LR-PC	0.795	0.537	0.26
SVM-BT (accuracy)	0.755	0.482	0.51
SVM-BT (logloss)	0.770	0.505	0.38
SVM-PC (accuracy)	0.764	0.467	0.76
SVM-PC (logloss)	0.764	0.469	0.81
kNN (accuracy)	0.684	0.403	0.76
kNN (logloss)	0.686	0.399	0.28
RF (accuracy)	0.792	0.539	0.40
RF (logloss)	0.792	0.539	0.40
NaB (accuracy)	0.742	0.451	8.25
NaB (logloss)	0.768	0.495	4.82
NSC (accuracy)	0.754	0.470	1.42
NSC (logloss)	0.753	0.471	1.19

Calibration

Overall the logistic regression approaches have the best mean calibration, while NaB is clearly worse specifically when accuracy was used to tune the hyperparameter (Table 6.7). For the other considered approaches the deviation is mild (up to 3%). No general tendency of underestimating or overestimating

Table 6.6: Conditional pairwise c-statistics for the multiclass prediction models for the validation data.

Model	Conditional pairwise probabilities				
(optimization method)	Be vs Bo	Be vs St1	Be vs St24	Be vs Met	Bo vs St1
MLR	0.887	0.917	0.964	0.950	0.702
LR-BT	0.887	0.920	0.967	0.949	0.724
LR-PC	0.883	0.919	0.966	0.950	0.710
SVM-BT (accuracy)	0.821	0.919	0.975	0.949	0.708
SVM-BT (logloss)	0.889	0.919	0.973	0.952	0.708
SVM-PC (accuracy)	0.710	0.887	0.973	0.937	0.733
SVM-PC (logloss)	0.701	0.896	0.974	0.928	0.730
kNN (accuracy)	0.683	0.784	0.945	0.797	0.589
kNN (logloss)	0.676	0.805	0.945	0.811	0.583
RF (accuracy)	0.880	0.919	0.983	0.929	0.748
RF (logloss)	0.880	0.919	0.983	0.929	0.748
NaB (accuracy)	0.867	0.887	0.973	0.932	0.720
NaB (logloss)	0.868	0.874	0.955	0.920	0.680
NSC (accuracy)	0.846	0.855	0.935	0.897	0.671
NSC (logloss)	0.844	0.859	0.937	0.898	0.679
Model	Conditional pairwise probabilities				
(optimization method)	Bo vs St24	Bo vs Met	St1 vs St24	St1 vs Met	St24 vs Met
MLR	0.881	0.893	0.773	0.722	0.684
LR-BT	0.881	0.883	0.782	0.724	0.704
LR-PC	0.881	0.886	0.780	0.730	0.698
SVM-BT (accuracy)	0.847	0.831	0.756	0.717	0.604
SVM-BT (logloss)	0.864	0.849	0.751	0.676	0.595
SVM-PC (accuracy)	0.895	0.898	0.746	0.744	0.594
SVM-PC (logloss)	0.895	0.888	0.746	0.724	0.608
kNN (accuracy)	0.799	0.622	0.750	0.517	0.669
kNN (logloss)	0.824	0.632	0.797	0.577	0.698
RF (accuracy)	0.912	0.864	0.793	0.676	0.695
RF (logloss)	0.912	0.864	0.793	0.676	0.695
NaB (accuracy)	0.895	0.892	0.785	0.742	0.693
NaB (logloss)	0.868	0.864	0.762	0.696	0.604
NSC (accuracy)	0.848	0.876	0.744	0.726	0.674
NSC (logloss)	0.850	0.880	0.743	0.728	0.674

Abbreviations: Be, benign; Bo, borderline; St1, Stage I; St24, Stage II-IV; Met, Metastatic.

Table 6.7: Mean calibration (in %) for the multinomial prediction models for the validation data.

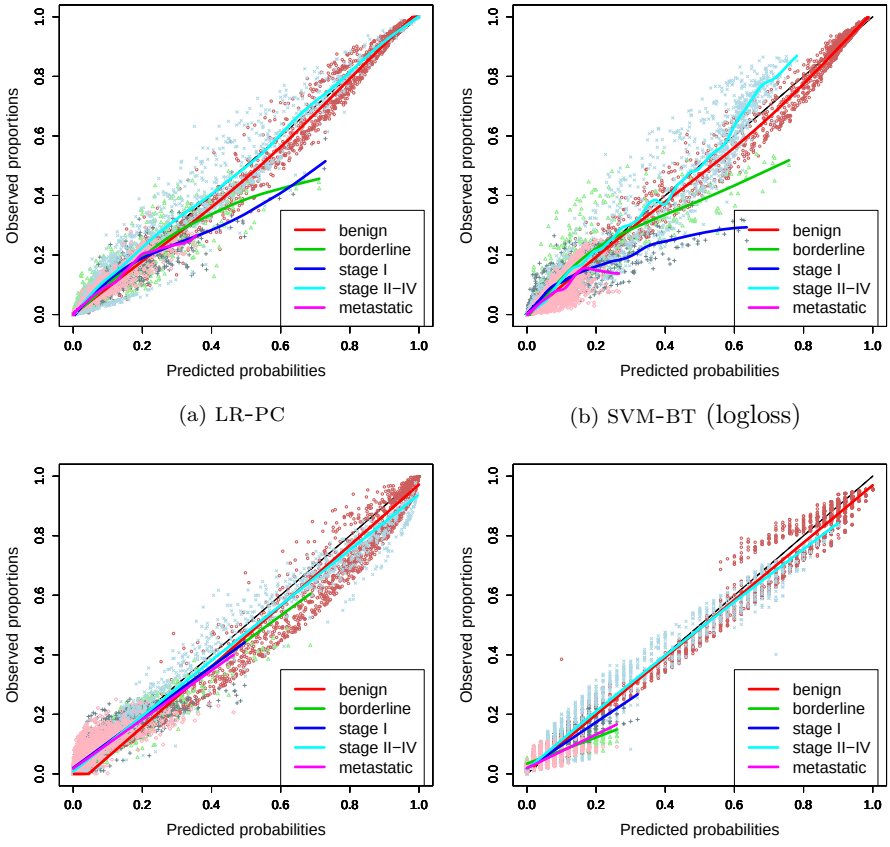
Model (optimization)	Be	Bo	St2	St24	Met
MLR	-0.13	-0.10	-0.16	0.37	0.03
LR-BT	-0.25	-0.31	-0.08	0.77	-0.13
LR-PC	-0.25	-0.40	-0.08	0.73	0.00
SVM-BT (accuracy)	-0.94	-0.80	0.14	1.51	-0.19
SVM-BT (logloss)	-0.40	0.01	-0.23	0.70	-0.40
SVM-PC (accuracy)	2.13	-1.63	-1.10	0.89	-0.29
SVM-PC (logloss)	2.10	-1.67	-1.01	0.86	-0.27
kNN (accuracy)	-1.74	0.58	0.45	0.54	0.17
kNN (logloss)	-1.92	0.52	0.57	0.61	0.21
RF (accuracy)	-3.35	1.10	1.31	-0.38	1.31
RF (logloss)	-3.35	1.10	1.31	-0.38	1.31
NaB (accuracy)	-22.60	4.90	5.49	8.44	3.77
NaB (logloss)	-0.48	-4.00	-0.45	5.81	-0.88
NSC (accuracy)	-2.80	0.10	-0.36	2.89	0.16
NSC (logloss)	-3.52	0.20	-0.27	3.28	0.31

Abbreviations: Be, benign; Bo, borderline; St1, Stage I; St24, Stage II-IV; Met, Metastatic.

the multiclass risks was observed. The sign depended on the considered model as well as the used optimization method (Table 6.7). The ECI was lowest for the logistic regression approaches (i.e. MLR, LR-BTand LR-PC) and kNN based on logloss, was slightly higher for SVM-BT based on logloss and RF, and was highest for NaB (Table 6.5). The multiclass calibration plots of the best calibrated models are shown in Figure 6.4. Recall from Section 6.2.3 that, contrary to the calibration plots for binary outcomes, there is no one-to-one relationship between predicted and observed probabilities. For that reason we added spline smoothers to summarize the trend. In Figure 6.5 simplified calibration plots with only the smoothed curves are shown.

The ECI was always lower when hyperparameters were based on logloss compared with accuracy, except for SVM-PC where results are nearly identical for both methods (Table 6.5). Hence minimization of logloss resulted in better calibration which is illustrated for kNN and NaB in Figure 6.6.

The multiclass calibration plots of the validation data for the different models based on accuracy and logloss are given in Appendix D.



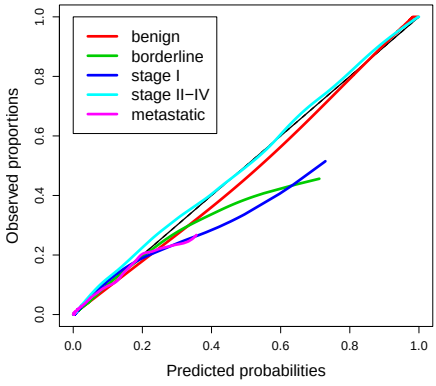
(a) LR-PC

(b) SVM-BT (logloss)

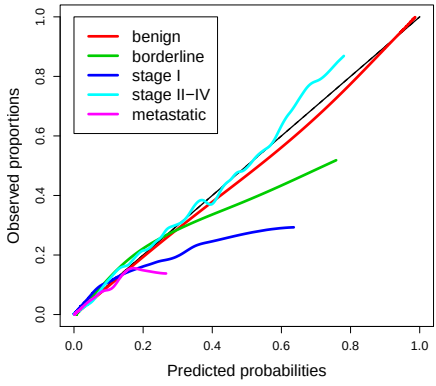
(c) RF

(d) kNN (logloss)

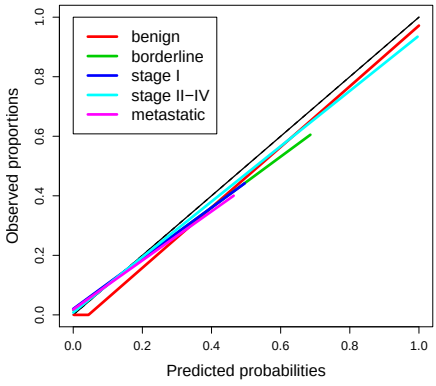
Figure 6.4: Multiclass calibration plot with smoother of the validation data for (a) pairwise coupling logistic regression model, (b) binary tree support vector machines with logloss as optimization method, (c) random forest and (d) k nearest neighbors with logloss as optimization method.



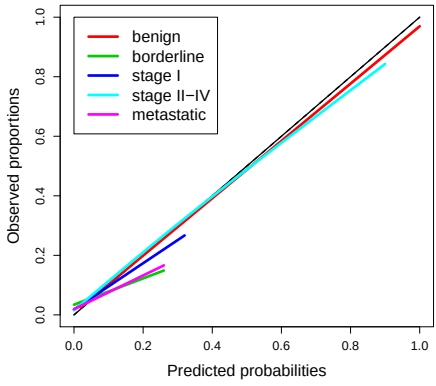
(a) LR-PC



(b) SVM-BT (logloss)

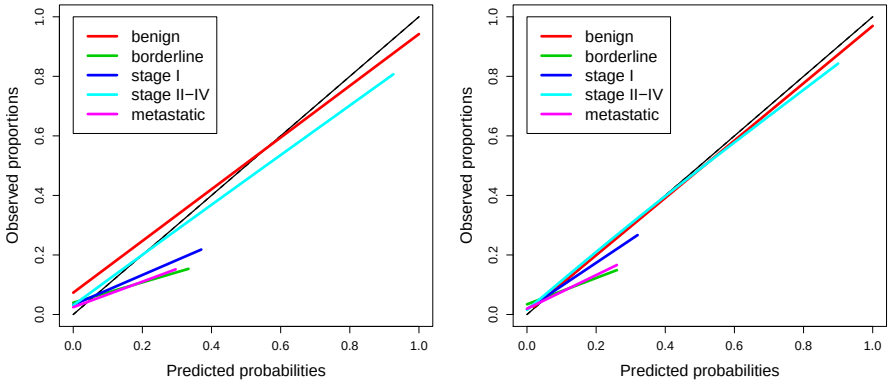


(c) RF



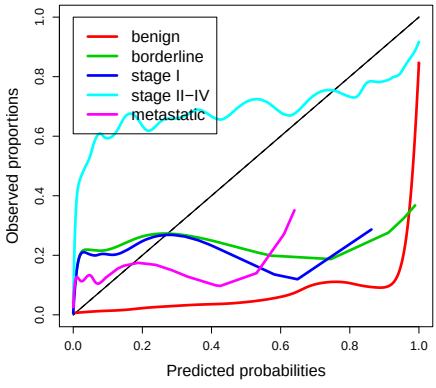
(d) kNN (logloss)

Figure 6.5: Smoothed calibration plot of the validation data for (a) nested dichotomous logistic regression models, (b) binary tree support vector machines with logloss as optimization method, (c) multiclass random forest and (d) multiclass k nearest neighbors with logloss as optimization method.

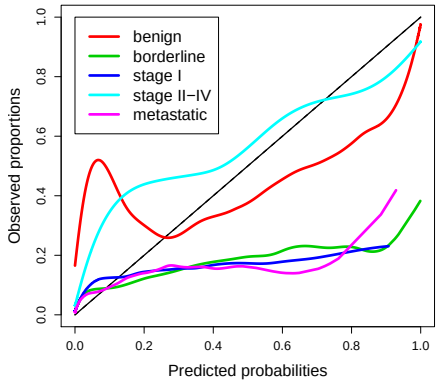


(a) kNN (accuracy)

(b) kNN (logloss)



(c) NaB (accuracy)



(d) NaB (logloss)

Figure 6.6: Smoothed calibration plot of the validation data for (a-b) multiclass k nearest neighbors and (c-d) multiclass naive bayes with accuracy and logloss as optimization method.

6.4 Discussion

When validating a risk model for medical applications, discrimination as well as calibration should be assessed [6, 97, 135, 167, 186]. The model needs not only to be able to distinguish the different outcome categories but also to reliably predict the risk of a certain outcome. In this Chapter we proposed the general calibration framework (cf. Chapter 4) to visualize and quantify the lack of calibration irrespective of the used modeling approach. A case study on the diagnosis of ovarian tumors was used for illustration. Different models were developed and validated with respect to discrimination and calibration: logistic regression, support vector machines, `k`-nearest neighbors, random forest, naive bayes and nearest shrunken centroids. Overall, logistic regression and random forest models showed the best calibration, whereas the calibration of naive bayes was disappointing. An explanation for the latter might be that the independence assumption is unrealistic for the data and therefore results in inaccurate estimates [153].

6.4.1 Visualization versus quantification of calibration

The general non-parametric recalibration framework from Chapter 4 was used to visualize the calibration by means of a calibration plot as well as to capture the lack of calibration through the estimated calibration index. The ECI takes the calibration of individual patient risks into account and has therefore an advantage over the general calibration-in-the-large or mean calibration. Both visualization and quantification have their advantages and disadvantages. The calibration plots are more informative and interpretable compared with ECI, whereas ECI is easier for comparing.

Since there is no one-to-one relationship between predicted and observed probabilities for the multiclass calibration plots (Figure 6.4) we use smoothed curves to visualize the general relationship. If only the smoothed curves are shown, as we did in Figures 6.5- 6.6 for clarity, in fact information is lost. Nevertheless, we believe the smoothed relationship is most inmportant in the assessment of calibration performance. Yet if this relationship is good, meaning that it is close to the diagonal line, it is still possible that predicted and observed probabilities are away from the diagonal line for many patients.

6.4.2 Tuning: logistic regression versus machine learning

The tuning of risk prediction models is an important aspect [99]. For logistic regression, aspects as variable selection, non-linear effects and interaction terms can be considered as tuning [99]. We have not investigated the inclusion of non-linear effects or interactions, although non-linear effects are most likely present [173]. In real clinical applications such tuning is advised, for example by modeling continuous predictors using spline functions, however the amount of tuning should depend on the available sample size [173]. Machine learning approaches have different hyperparameters that need tuning (e.g. number of neighbors for kNN and number of considered variables in each split for RF) [99]. The choice of optimization method for these hyperparameters may influence model performance [112]. As expected, we noticed in this Chapter that considering logloss as optimization method for tuning resulted in a better calibration compared with accuracy. We acknowledge that other optimization methods can be considered and may yield different model performance [112].

6.4.3 Choice of algorithms

In the literature, many machine learning models are available and the optimal hyperparameters of these models can be obtained via different optimization techniques. Different methods can be used to create probabilities, certainly for multiclass outcomes [191, 234]. This results in a plethora of possible algorithms to develop binary and multiclass risk models.

In this Chapter we considered some approaches, logistic regression as well as machine learning, more as an illustration of the multiclass calibration tools than as an attempt to compare algorithms with respect to the reliability of the estimated risks. The latter would be of interest, but then we believe a large benchmark study would be required to compare a large battery of algorithms on many different datasets.

6.4.4 Further practical limitations

The data were randomly divided into development and validation data, with stratification for the multiclass outcome. This random split allowed the construction of calibration plots, but the results may depend on the split. We expect that the dependency is rather limited certainly due to the large sample size. Also, in real applications it is more interesting to use an external validation dataset, for example by using new data collected later in time or data from different hospitals.

Another limitation is that we decided to use single imputation for CA125, because this ignores the fact that we are uncertain about whether the imputed values are correct [163]. An alternative to deal with the issue of uncertainty would be multiple imputation [163]. However, since the aim of this research was not the development of risk prediction models for clinical practice, we chose for the pragmatic solution of single imputation.

6.5 What is new?

Key findings

- The non-parametric recalibration framework proposed in Chapter 4 was proposed to visualize and quantify the lack of calibration for different binary and multiclass risk prediction models.
- The calibration plots visualize the agreement between predicted probabilities and observed proportions irrespective of the used modeling technique.

What this research adds to what was known

- Calibration is an important aspect because treatment decisions rely on the estimated risk of disease. The non-parametric recalibration framework proposed in Chapter 4 for multinomial logistic regression is used for other risk prediction models.
- Calibration plots as well as the estimated calibration index (ECI) can be used to assess calibration of multiclass probabilities from machine learning classifiers.
- The calibration plots visualize the reliability of the predicted risks while the estimated calibration index (ECI) quantifies the lack of calibration.

Implications of this research

- Prior to the use of a risk prediction model in clinical practice, calibration should be assessed. Calibration plots can be used as a visualization tool.
- Although research is conducted concerning the different possibilities to create probability estimates from classifiers, more attention for calibration in machine learning research is still required. During (external) validation of machine learning models, calibration (i.e. the reliability of the predicted

risks or the correspondence between predicted probabilities and observed proportions) is often overlooked.

6.6 Conclusion

The non-parametric recalibration framework suggested in Chapter 4 is an interesting framework to visualize the reliability of the predicted risks and to quantify lack of calibration. The estimated calibration index (ECI) is easy for comparing models, but is nevertheless less informative and interpretable than the calibration plots.

Chapter 7

Overview and future work

This final Chapter summarizes findings and conclusions of the presented work. In Section 7.1 an overview of my personal contributions is given. Afterwards, in Section 7.2, future research directions are suggested.

7.1 Overview personal contributions

The aim of this dissertation was twofold: extension of dichotomous calibration tools towards multinomial outcomes (cf. Chapter 4) and generalizing dichotomous updating techniques towards polytomous outcomes based on multinomial logistic regression (cf. Chapter 5).

In Chapter 3, different approaches to model a multinomial outcome via logistic regression are introduced and compared. The multinomial outcome can either be modeled all-at-once or via a combination of 1-vs-1 and 1-vs-rest dichotomous logistic regression models. The all-at-once multinomial logistic regression model is the common approach used in epidemiology and regarded as reference approach. As combinations we considered sequential dichotomous model or binary tree, dichotomous logistic regression models combined via pairwise coupling, individualized logistic regression and 1-vs-rest logistic regression. The methods differed with respect to the number of submodels as well as the complexity to obtain the multinomial risk predictions. The comparison of the different approaches on the case study of diagnosing ovarian tumors as benign, borderline or invasive indicated similar calibration-in-the-large values but differences in selected variables and discriminative ability.

In Chapter 4, multinomial calibration measures as well as multinomial calibration plots were proposed. We focused on the multinomial logistic regression model of an outcome Y with k categories. From this model we obtained multinomial risk predictions and $k-1$ linear predictors lp_{iR} with $i = 1, ..., k$, $i \neq R$ and R the chosen reference category (e.g. $R = 1$). First, the multinomial calibration-in-the-large was introduced. This overall measure calculates the difference between the observed prevalence and the average predicted probabilities for each outcome category. Afterwards a nominal logistic recalibration framework, modeling the multinomial outcome on the $k-1$ linear predictors lp_{iR} with $i = 1, ..., k$ and $i \neq R$, was introduced. This framework can be used to estimate the calibration intercepts and calibration slopes, which have to be interpreted with respect to the chosen reference category. The approach from Miller et al. was extended for statistical testing [117]. An alternative and more general non-parametric framework is obtained by applying vector spline smoothers on the different linear predictors. To visualize the correspondence between predicted probabilities and observed proportions, the latter can be estimated by the logistic or non-parametric recalibration framework. In this way, parametric and non-parametric calibration plots were obtained. The calibration tools were illustrated on the case studies of diagnosing ovarian tumors as benign, borderline and invasive and the prediction of residual mass after chemotherapy for testicular cancer as benign, teratoma or viable cancer. To end, as application, a multinomial risk prediction model was developed and evaluated on the case study of ovarian tumors distinguishing benign, borderline, stage I invasive and stage II-IV invasive and metastatic cancer. Multinomial calibration plots were made to visualize the reliability of the predicted risks.

In Chapter 5, the topic of multinomial updating was covered. As a first step, a multinomial outcome was modeled via a sequential dichotomous logistic regression model. Existing dichotomous updating techniques were applied on the different dichotomous submodels. The following types of updating approaches, which differ in complexity, were applied: recalibration, revision and redevelopment. The case studies concerning ovarian tumors and testicular cancer were considered. The dichotomous updating methods worked well for models that predict multinomial outcomes and that consist of a sequence of dichotomous logistic regression models. In case of a rather small sample size or a relatively low prevalence of one of the outcome categories, recalibration was recommended. Revision, extension and model redevelopment might be a valid alternative when a large amount of data are available. Afterwards, the dichotomous updating techniques were extended towards nominal updating techniques. We focused on risk prediction models based on multinomial logistic regression. Recalibration methods, based on the logistic recalibration framework introduced in Chapter 4, as well as revision and extension methods were proposed. The different suggested methods were illustrated and compared on the diagnosis

of pregnancies of unknown location as failed pregnancy (FPUL), intra-uterine pregnancy (IUP) and ectopic pregnancy (EP). The existing multinomial risk prediction model M4 was updated and validated. The obtained conclusions were similar to the ones in the first part of the Chapter. Recalibration methods worked well although the original risk prediction model was based on a small sample size (N=197: 109 FPUL, 76 IUP and 12 EP). If enough data are available revision and/or extension methods might still be preferred to improve the individual risk predictions.

Finally, in Chapter 6, the multinomial non-parametric calibration framework of Chapter 4, was proposed to evaluate calibration in the setting of machine learning models. This framework was used to generate calibration plots and to quantify lack of calibration by means of the estimated calibration index (ECI). The considered case study was the diagnosis of ovarian tumors as benign, borderline, stage I invasive, stage II-IV invasive and metastatic cancer. The calibration of risk predictions based on different machine learning approaches (i.e. support vector machines, k nearest neighbors, random forest, naive bayes and nearest shrunken centroids) was visualized and measured.

7.2 Future work

In this dissertation several interesting research topics were covered, but there are still other related areas that require further research. In this section we will cover these in more detail.

7.2.1 Sample size issues

A first topic concerns the influence of the sample size N in case of a polytomous outcome with k categories (i.e. $N = \sum_{i=1}^{k} N_i$). General guidelines for the development and validation of a multinomial logistic regression model may be formulated [64, 217]. For the development of a prediction model, a large enough dataset is required to obtain a robust and accurate prediction model [64]. A more complex research question results in more required data [167]. For validation, a too small dataset can result in not enough power to detect differences in model performance [217]. It might be an option to formulate guidelines in terms of a minimal total sample size N, although this might result in only a few cases in one or more of the outcome categories. An alternative is to specify a minimal number of required events per variable in the smallest outcome category j as for a dichotomous outcome. E.g. an EPV of 10 would imply that the number of cases in the smallest outcome category j (i.e. $\min_i(N_i) = N_j$) equals 10 times

the number of considered variables (i.e. $N_j = n \cdot 10$) or even $N_j = n \cdot 10 \cdot (k-1)$ since a coefficient is estimated for each predictor for each category i versus a certain reference category R with $i = 1, ...k$ and $i \neq R$. However, there are several possible situations for N_j cases in the smallest outcome category j: first of all, it might be that all outcome categories i have the same $N_i = N_j$ cases, an alternative is that only category j has only N_j cases, or it might be that several categories have only N_j patients. These three situations imply a different total sample size N. Does it make a difference if only one outcome category is small or all outcome categories are small? Should guidelines be formulated in terms of EPV and total sample size N [225]? Furthermore, is it enough to consider the number of considered variables or do we need to account for the number of times a coefficient is estimated (i.e. $k-1$) for a certain variable? Does it make a difference whether variable selection is performed or not? Finally, does the required sample size decrease when penalization techniques are used?

Furthermore, the influence of sample size on calibration and more specific calibration slopes should be investigated.

Finally, a sample size study for dichotomous and multinomial updating with varying sample size for the original prediction model as well as updating data should be conducted [168]. This could result in guidelines concerning the preferred updating method in different situations. A low sample size of the development data and a large new dataset may imply a preference towards more complex methods, while for the reverse simple recalibration methods may be suggested.

7.2.2 Multinomial risk prediction models

A multinomial logistic regression model consists of the simultaneous fitting of $k-1$ dichotomous logistic regression models, research concerning the functional form of the considered predictors as well as penalization techniques is desired [22, 32, 51, 53, 103, 168, 170, 208, 210]. Questions as whether all $k-1$ models need the same transformation for a certain predictor or whether each model requires the same shrinkage arise. Furthermore, the dichotomous procedure of multivariable fractional polynomials (MFP), which simultaneously selects variables and their optimal transformation, should be extended towards multinomial logistic regression [155, 158]. Cubic splines have already been extended towards vector responses [50, 230–232], but also restricted cubic splines should be extended [63].

7.2.3 Multinomial performance measures

For a dichotomous outcome there exist a plethora of measures concerning discrimination, calibration and clinical utility. Extensions for discrimination and calibration were already covered in this dissertation, yet there are still new multinomial measures needed. Clinical utility measures are gaining importance, implying that multinomial versions of the dichotomous measures are needed. Concerning the reclassification statistics to assess improvement, Li et al. suggested a multicategory NRI and IDI [106], whereas e.g. a multinomial net benefit and decision curve still have to be proposed.

Additionally, it might be of interest to use and consider benchmark values to distinguish case-mix effects and wrong coefficients for validation results of polytomous prediction models [214]. A low validation performance can either be the result of wrong parameter coefficients or due to differences in distribution of patient characteristics between development and validation data. The difference can appear in predictors considered in the prediction model as well as predictors related to the outcome that were not considered [214]. For example, a prediction model for the diagnoses of ovarian tumors might be developed on patients recruited in tertiary referral centers for oncology and evaluated on patients from regional or public hospitals. The benchmark values can help to distinguish case-mix effects and wrong coefficients and are therefore useful for the interpretation of the model performance in the validation data. For the evaluation of polytomous discrimination, is it enough to only consider a case-mix corrected PDI, or should the conditional pairwise c-statistics be taken into account as well? How can all this information be used to formulate a conclusion?

Moreover, another research question is whether the use of random intercepts, to capture e.g. center-specific effects, for multinomial logistic regression improves the calibration of the individual patient risks [20].

A last topic concerning multinomial performance measures is the development of an automatic selection procedure (e.g. generalized cross validation) for the level of smoothing for multinomial calibration plots [57].

7.2.4 Multinomial updating

For the topic of updating, research concerning altering the functional form of one or more predictors from the original prediction model might be conducted. Furthermore, the recently suggested closed test procedure to determine the

optimal updating method for dichotomous risk models should be extended towards multinomial models [215].

7.2.5 Ordinal outcome

In this dissertation we focused on nominal outcomes. Polytomous outcomes might be ordinal instead of nominal. The topics discussed in this dissertation, i.e. calibration tools and updating techniques, as well as the suggested research topics above could and should be investigated for ordinal outcomes as well [3, 113].

The most common used prediction model for ordinal outcomes with k levels is the proportional odds model with $j = 1, ..., k-1$:

$$\begin{aligned} \text{logit } P(\mathrm{Y} \leq j|\mathbf{x}) &= \log \frac{P(\mathrm{Y} \leq j|\mathbf{x})}{1 - P(\mathrm{Y} \leq j|\mathbf{x})} \\ &= \log \frac{\pi_1(\mathbf{x}) + \cdots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \cdots + \pi_k(\mathbf{x})} \\ &= a_j + b^T\mathbf{x}, \end{aligned} \tag{7.1}$$

where $b = [b_1\ b_2\ ...\ b_n]^T$, $\mathbf{x}^T = [\mathrm{x}_1\ \mathrm{x}_2\ \mathrm{x}_3\ ...\ \mathrm{x}_n]$ and $\pi_j(\mathbf{x})$ the probability of belonging to outcome level j. For the proportional odds model, $k-1$ intercepts a_j and one parameter vector b is estimated. The dichotomous calibration statistics, e.g. logistic recalibration framework with calibration intercept and slope, and calibration plots generalized towards multinomial outcomes should also be extended towards ordinal risk predictions based on the proportional odds model. Furthermore, updating techniques for the proportional odds model should be suggested.

Appendix A

R code

A.1 Multinomial calibration function

```
#################################################################
# Multinomial Calibration function R                            #
# Assessing calibration of multinomial risk prediction models   #
# Van Hoorde, K et al. Statistics in Medicine                   #
# libraries vgam and bayesm are needed                          #
#################################################################

Polcal <- function(outcome,k,p,LP,r=1,estimates=FALSE,dfr=2,
          plotoverall=TRUE,datapoints=TRUE,smoothing=TRUE,smoothpar=1,
          intercept=FALSE,slope=FALSE,test=FALSE){

# NOTE:
# This function is written for a multinomial outcome with three categories.
# If there are more than three categories, the code can be easily adapted.
# Comments are added to state where and how the code should be changed if
# needed.


#############################################################################
# outcome     = column vector containing the outcome for every case, with
#               values 1 to k (i.c. k=3)
# k           = number of outcome categories (i.c. 3)
# p           = matrix with the probabilities of the prediction model,
#               ordered from prob(cat. 1) to prob(cat. k)
# LP          = matrix with all the linear predictors with respect to the
#               chosen reference category, ordered (e.g. LP2vs1 and LP3vs1)
# r           = reference category (default: category 1)
# estimates   = indicates whether the coefficients of the parametric
#               recalibration framework are desired (default=FALSE)
# dfr         = degrees of freedom for the non-parametric calibration
#               default=2)
# plotoverall = indicates whether overall (non-)parametric calibration plots
#               are constructed (default=TRUE)
# datapoints  = indicates whether the individual datapoints are shown in the
#               overall (non-)parametric calibration plots (default = TRUE)
```

```
# smoothing   = indicates whether a smoothed line (using cubic splines) is
#               added to the calibration plots (default=TRUE)
# smoothpar   = smoothing parameter for the smoothed line (default=1)
# intercept   = indicates whether calibration intercepts are desired
#               (default=FALSE)
# slope       = indicates whether calibration slopes are desired
#               (default=FALSE)
# test        = indicates whether statistical tests for calibration are
#               desired (default=FALSE)
#############################################################################

# for this function you need to use library (VGAM)
 library(VGAM)
 library(bayesm) # necessary for testing

# checks
 if(k != length(table(outcome))){stop('The number of categories in outcome
     does not equal the specified number of categories.')}
 if(dim(p)[2]!=k){stop('The number of columns in p does not equal the
     specified number of categories.')}
 if(dim(LP)[2]!=k-1){stop('The number of columns in LP does not equal the
     specified number of categories minus 1.')}
 if(! r %in% 1:k){stop('The given reference category (r) is not possible.')}
 if(!is.matrix(p)){stop('p is not a matrix.')}
 if(!is.matrix(LP)){stop('LP is not a matrix.')}
 if(isTRUE(plotoverall) && !isTRUE(datapoints) && !isTRUE(smoothing)){stop('
     For the overall (non-)parametric calibration plots either datapoints or
      smoothed
 lines should be requested.')}

# if tests for perfect calibration are requested, automatically calibration
    intercepts and calibration slopes are given
  if(isTRUE(test)){intercept<-slope<-TRUE}

# probabilities
  probs <- split(p,col(p))

# linear predictors necessary for non-parametric calibration plot - give a
    name to each linear predictor seperately
  lps <- split(LP,col(LP))
  for(i in 1:(k-1)){assign(paste("lp", i, sep = ""),unlist(lps[[i]]))}


################################################
# parametric logistic recalibration framework #
# cf. section 2.2.1.                          #
################################################

# reference category r
# LP = matrix with linear predictors

  fitp<-vglm(outcome~LP,family=multinomial(refLevel=r))
  if(isTRUE(estimates)){est<-coefficients(fitp)
  names(est) <- paste('EST',names(est),sep='.')}


#########################################################
# non-parametric recalibration framework (using df=2) #
# cf. section 2.2.1.                                   #
#########################################################

# reference category r
# lpi = ith linear predictor
```

```
92 # for k outcome categories, there are k-1 linear predictors and the code
       should be adapted to:
93 # fitnp <- vgam(outcome~s(lp1,df=dfr)+s(lp2,df=dfr)+...+s(lpk-1,df=dfr),
       family=multinomial(refLevel=r))
94
95   fitnp<-vgam(outcome~s(lp1,df=dfr)+s(lp2,df=dfr),family=multinomial(
         refLevel=r))
96
97
98 ################################################
99 # Separate (non-)parametric calibration plots #
100 ################################################
101
102   windows()
103   par(mfrow=c(ceiling(k/2),2))
104
105 # parametric calibration plots #
106 # cf. section 2.2.2.           #
107 ################################
108
109   for(i in 1:k){p <- unlist(probs[[i]])
110   if(isTRUE(smoothing)){color<-'grey'}else{color<-1+i}
111   matplot(p,fitted(fitp)[,i],type="p",pch=i,col=color,lwd=1,ylab="",xlab="",
         xlim=0:1,ylim=0:1)
112   par(new=T)
113   ref <- rbind(c(0,0),c(1,1))
114   matplot(ref,ref,type="l",col=1,lwd=2,ylab="Observed proportions",xlab="
         Predicted probabilities",xlim=0:1,ylim=0:1)
115   # smoother for calibration plots #
116   ##################################
117   # a = smoothing parameter
118   if(isTRUE(smoothing)){
119   a = smoothpar
120   points(smooth.spline(p, fitted(fitp)[,i],spar=a), type="l", col=(1+i), lwd
          = 4)}
121   # legend
122   legende <- c(paste("cat ", i, sep = ""))
123   legend(x=0.6, y=(0.2),col=(1+i),lty =1,legend=legende)
124   title(main = "Parametric calibration plot")
125   par(new=F)}
126
127 # non-parametric calibration plot #
128 # cf. section 2.2.2.              #
129 ###################################
130   windows()
131   par(mfrow=c(ceiling(k/2),2))
132   for(i in 1:k){p <- unlist(probs[[i]])
133   if(isTRUE(smoothing)){color<-'grey'}else{color<-1+i}
134   matplot(p,fitted(fitnp)[,i],type="p",pch=i,col=color,lwd=1,ylab="",xlab=""
         ,xlim=0:1,ylim=0:1)
135   par(new=T)
136   ref <- rbind(c(0,0),c(1,1))
137   matplot(ref,ref,type="l",col=1,lwd=2,ylab="Observed proportions",xlab="
         Predicted probabilities",xlim=0:1,ylim=0:1)
138   # smoother for calibration plots #
139   ##################################
140   # a = smoothing parameter
141   if(isTRUE(smoothing)){
142   a = smoothpar
143   points(smooth.spline(p, fitted(fitnp)[,i],spar=a), type="l", col=(1+i),
         lwd = 4)}
144   # legend
145   legende <- c(paste("cat ", i, sep = ""))
```

```
146   legend(x=0.6, y=(0.2),col=(1+i),lty =1,legend=legende)
147   title(main = "Non-parametric calibration plot")
148   par(new=F)}
149
150
151 ##############################################
152 # Overall (non-)parametric calibration plot #
153 ##############################################
154
155   if(isTRUE(plotoverall)){
156   windows()
157
158 # parametric calibration plot #
159 # cf. section 2.2.2.          #
160 ###############################
161
162   if(isTRUE(datapoints)){for(i in 1:k){p <- unlist(probs[[i]])
163   matplot(p,fitted(fitp)[,i],type="p",pch=i,col=(1+i),lwd=1,ylab="",xlab="",
          xlim=0:1,ylim=0:1)
164   par(new=T)}}
165   ref <- rbind(c(0,0),c(1,1))
166   matplot(ref,ref,type="l",col=1,lwd=2,ylab="Observed proportions",xlab="
          Predicted probabilities",xlim=0:1,ylim=0:1)
167   # smoother for calibration plots #
168   ##################################
169   # a = smoothing parameter
170   if(isTRUE(smoothing)){
171   a = smoothpar
172   for(i in 1:k){p <- unlist(probs[[i]])
173   points(smooth.spline(p, fitted(fitp)[,i],spar=a), type="l", col=(1+i), lwd
            = 4)}}
174   # legend
175   for(i in 1:k){if(i <= 2){legende <- c("cat 1","cat 2")}
176   if(i > 2){legende <- c(legende,paste("cat ", i, sep = ""))}}
177   legend(x=0.7, y=(0.20+(k-3)*0.05),col=2:(k+1),lty =1,legend=legende)
178   title(main = "Parametric calibration plot")
179   par(new=F)
180
181 # non-parametric calibration plot #
182 # cf. section 2.2.2.              #
183 ###################################
184
185   windows()
186   if(isTRUE(datapoints)){for(i in 1:k){p <- unlist(probs[[i]])
187   matplot(p,fitted(fitnp)[,i],type="p",pch=i,col=(1+i),lwd=1,ylab="",xlab=""
          ,xlim=0:1,ylim=0:1)
188   par(new=T)}}
189   ref <- rbind(c(0,0),c(1,1))
190   matplot(ref,ref,type="l",col=1,lwd=2,ylab="Observed proportions",xlab="
          Predicted  probabilities",xlim=0:1,ylim=0:1)
191   # smoother for calibration plots #
192   ##################################
193   # a = smoothing parameter
194   if(isTRUE(smoothing)){a = smoothpar
195   for(i in 1:k){p <- unlist(probs[[i]])
196   points(smooth.spline(p, fitted(fitnp)[,i],spar=a), type="l", col=(1+i),
          lwd = 4)}}
197   # legend
198   for(i in 1:k){if(i <= 2){legende <- c("cat 1","cat 2")}
199   if(i > 2){legende <- c(legende,paste("cat ", i, sep = ""))}}
200   legend(x=0.7, y=(0.20+(k-3)*0.05),col=2:(k+1),lty =1,legend=legende)
201   title(main = "Non-parametric calibration plot")
202   par(new=F)}
```

```
203
204
205 ########################################
206 # estimation of calibration intercepts #
207 # cf. section 2.2.3. and 2.2.4.        #
208 ########################################
209
210   if(isTRUE(intercept)){int<-vgam(outcome~1,offset=LP,family=multinomial(
          refLevel=r))
211   coeffint<-coefficients(int)
212   se<-sqrt(diag(vcov(int)))
213   ci1i <- cbind(LL1 = coeffint[1] - qnorm(0.975) * se[1], UL1 = coeffint[1]
          + qnorm(0.975) * se[1])
214   ci2i <- cbind(LL2 = coeffint[2] - qnorm(0.975) * se[2], UL2 = coeffint[2]
          + qnorm(0.975) * se[2])
215   estint <- c(coeffint[1],ci1i,coeffint[2],ci2i)
216   names(estint) <- paste('CALINT',c('int1','LLint1','ULint1','int2','LLint2'
          ,'ULint2'),sep='.')}
217
218
219 ####################################
220 # estimation of calibration slopes #
221 # cf. section 2.2.3. and 2.2.4.    #
222 ####################################
223
224 # we used constraints to fix some coefficients to zero as appropriate
225 # for k outcome categories this code should be changed to:
226 # i <- diag(k-1)
227 # i2 <- cbind(c(1,rep(0,k-2)))
228 # i3 <- cbind(c(0,1,rep(0,k-1)))
229 # i4 <- cbind(c(0,0,1,rep(0,k-2)))
230 # ... (ij <- cbind(c(rep(0,j-2),1,rep(0,k-j)))
231 # ik <- cbind(c(rep(0,k-2),1))
232 # clist<-list("(Intercept)"=i,"lp1"=i2,"lp2"=i3,...,"lpk-1"=ik)
233 # slopes<-vgam(outcome~lp1+lp2+...+lpk-1,family=multinomial(refLevel=r),
        constraints=clist)
234
235   if(isTRUE(slope)){i<-diag(k-1)
236   i2<-rbind(1,0)
237   i3<-rbind(0,1)
238   clist<-list("(Intercept)"=i,"lp1"=i2,"lp2"=i3)
239   slopes<-vgam(outcome~lp1+lp2,family=multinomial(refLevel=r),constraints=
          clist)
240   coeffslopes<-coefficients(slopes)[k:length(coefficients(slopes))]
241   se<-sqrt(diag(vcov(slopes)))
242   ci1s <- cbind(LL1 = coeffslopes[1] - qnorm(0.975) * se[3], UL1 =
          coeffslopes[1] + qnorm(0.975) * se[3])
243   ci2s <- cbind(LL2 = coeffslopes[2] - qnorm(0.975) * se[4], UL2 =
          coeffslopes[2] + qnorm(0.975) * se[4])
244   estslopes <- c(coeffslopes[1],ci1s,coeffslopes[2],ci2s)
245   names(estslopes) <- paste('CALSLOPES',c('lp1','LLlp1','ULlp1','lp2','LLlp2
          ','ULlp2'),sep='.')}
246
247
248 ##################################
249 # calibration testing            #
250 # cf. section 2.2.3. and 2.2.4.  #
251 ##################################
252
253 # this code requires the bayesm library developed by Peter Rossi
254
255   if(isTRUE(test)){
256
```

```
# -2 log-likelihood of model without adaptations
# for k outcome categories this code should be changed to:
# alphas <- rep(0,k-1) #(i.e. all intercepts zero)
# beta1 <- c(1,rep(0,k-2)) #(i.e. first linear predictor for first equation)
# beta2 <- c(0,1,rep(0,k-3)) #(i.e. second linear predictor for second equation)
# betaj <- c(rep(0,j-1),1,rep(0,k-1-j)) #(i.e. jth linear predictor for jth equation)
# betak <- c(rep(0,k-2),1) #(i.e. kth linear predictor for kth equation)
# parametersk <- c(alphas, beta1, beta2, ..., betak)

  parametersk <- c(0,0,1,0,0,1) #c(alpha1,alpha2,b22,b23,b32,b33)
  Xdk=LP
  x <- createX(p=k,na=0,nd=k-1,Xa=NULL,Xd=Xdk,INT=TRUE,DIFF=FALSE,base=1)
  deviancewithout <- -2*llmnl(parametersk,outcome,x)
  names(deviancewithout)<-c('original deviance')

  devint <- deviance(int)
  names(devint)<-c('intercept deviance')
  devslopes <- deviance(slopes)
  names(devslopes)<-c('slopes deviance')

  # overall calibration (i.e. calibration intercepts and slopes) #
  ################################################################

  poverall<- pchisq(deviancewithout - devslopes, df = 2*(k-1), lower.tail =
      FALSE)

  # calibration intercepts #
  ##########################

  pint<- pchisq(deviancewithout - devint, df = k-1, lower.tail = FALSE)

  # calibration slopes #
  ######################

  pslopes<- pchisq(devint - devslopes, df = k-1, lower.tail = FALSE)
  names(poverall)<-c('p overall')
  names(pint)<-c('p int')
  names(pslopes)<-c('p slopes')}

# Printing of results
# The probabilities of calibration intercepts and slopes are only shown when
    the hypothesis of perfect calibration is rejected.

  results<-list(if(isTRUE(estimates)){est}else{'Not requested'},if(isTRUE(
      intercept)){estint}else{'Not requested'},if(isTRUE(slope)){estslopes}
      else{'Not requested'},if(isTRUE(test)){c(deviancewithout,devint,
      devslopes)}else{'Not requested'},if(isTRUE(test)){c(poverall,if(
      poverall<0.05){c(pint,pslopes)})}else{'Not requested'})
  names(results)<-c("Coefficients of parametric recalibration framework","
      Calibration Intercepts with 95% CI","Calibration Slopes with 95% CI","
      Deviances","P-values")
  n <- 1:5
  selection <- c(isTRUE(estimates),isTRUE(intercept),isTRUE(slope),isTRUE(
      test),isTRUE(test))
  results[n[selection]]

}
```

Appendix B

SAS code

B.1 Multinomial calibration macro

```
/*****************************************************************/
/* Multinomial Calibration Macro SAS                             */
/* Assessing calibration of multinomial risk prediction models */
/* Van Hoorde, K et al. Statistics in Medicine                   */
/*****************************************************************/

%macro Polcal(indata=,estimation=1);
/* This macro is written for an outcome with three different categories.
   If there are more than three categories, the code can be easily adapted.

   Indata: dataset with the outcome (variable name 'outcome') in the first
           column.
           Outcome has values 1, 2, and 3, with 1 the reference category.
           The second and third columns contain the linear predictors for
           category 2 vs 1 (variable name 'lp2') and for category 3 vs 1
           (variable name 'lp3').

   Estimation: indicates whether estimations of calibration intercepts and
               slopes (with statistical tests) are desired (this is default)
               or whether only calibration plots are needed.
               If only plots are needed, estimation should not be 1.
*/


/*********************************************************************/
/* (1) parametric logistic recalibration framework; cf section 2.2.1. */
/* reference category 1                                              */
/* lp2 = linear predictor for category 2 versus 1                    */
/* lp3 = linear predictor for category 3 versus 1                    */
/*********************************************************************/

title j=center font= 'Arial' c=blue bold "Parametric logistic recalibratrion
     framework";
```

```
33 title2 j=center font= 'Arial' c=green italic "Intercepts: alpha 2 and alpha
       3;  Corresponding slopes: b_22 and b_33;  Non-corresponding slopes: b_23
        and b_32";
34
35 proc nlmixed data=&indata;
36 parms   alpha2=0 b_22=1 b_23=0
37     alpha3=0 b_32=0 b_33=1;
38 eta1=0;
39 eta2= alpha2 + b_22*lp2 + b_23*lp3;
40 eta3= alpha3 + b_32*lp2 + b_33*lp3;
41 ARRAY exp_eta{3};
42 ARRAY estprobs{3};
43 exp_eta1=1;exp_eta2=exp(eta2);exp_eta3=exp(eta3);
44 bot=exp_eta1+exp_eta2+exp_eta3;
45 estprobs1=exp_eta{1}/bot;estprobs2=exp_eta{2}/bot;estprobs3=exp_eta{3}/bot;
46 p_outcome=exp_eta{outcome}/bot;
47 ll=log(p_outcome);
48 model outcome~GENERAL(ll);
49 id estprobs2 estprobs3;
50 predict estprobs1 out=mlrsep_logrecal;
51 run;
52
53
54 %if %sysevalf(&estimation)=1 %then %do;
55
56 /*************************************************************************/
57 /* (2) estimation of calibration intercepts; cf sections 2.2.3.-2.2.4. */
58 /* reference category 1                                                  */
59 /* lp2 = linear predictor for category 2 versus 1                        */
60 /* lp3 = linear predictor for category 3 versus 1                        */
61 /*************************************************************************/
62
63 title j=center font= 'Arial' c=blue bold "Estimation of calibration
       intercepts";
64 title2 j=center font= 'Arial' c=green italic "Calibration intercepts: alpha
       2 and alpha 3";
65
66 * we used proc nlmixed to fix some coefficients to 0 or 1 as appropriate;
67 proc nlmixed data=&indata;
68 parms   alpha2=0
69     alpha3=0;
70 eta1=0;
71 eta2= alpha2 + lp2;
72 eta3= alpha3 + lp3;
73 ARRAY exp_eta{3};
74 ARRAY estprobs{3};
75 exp_eta1=1;exp_eta2=exp(eta2);exp_eta3=exp(eta3);
76 bot=exp_eta1+exp_eta2+exp_eta3;
77 estprobs1=exp_eta{1}/bot;estprobs2=exp_eta{2}/bot;estprobs3=exp_eta{3}/bot;
78 p_outcome=exp_eta{outcome}/bot;
79 ll=log(p_outcome);
80 model outcome~GENERAL(ll);
81 id estprobs2 estprobs3;
82 ods output FitStatistics = ll_int Parameters = ll_orig;
83 run;
84
85
86 /*************************************************************************/
87 /* (3) estimation of calibration slope; cf sections 2.2.3.-2.2.4. */
88 /* reference category 1                                                  */
89 /* lp2 = linear predictor for category 2 versus 1                        */
90 /* lp3 = linear predictor for category 3 versus 1                        */
91 /*************************************************************************/
```

```
92
93  title j=center font= 'Arial' c=blue bold "Estimation of calibration slopes";
94  title2 j=center font= 'Arial' c=green italic "Calibration slopes: b_22 and b
        _33";
95
96  * we used proc nlmixed to fix some coefficients to 0 as appropriate;
97  proc nlmixed data=&indata;
98  parms   alpha2=0 b_22=1
99      alpha3=0 b_33=1;
100 eta1=0;
101 eta2= alpha2 + b_22*lp2;
102 eta3= alpha3 + b_33*lp3;
103 ARRAY exp_eta{3};
104 ARRAY estprobs{3};
105 exp_eta1=1;exp_eta2=exp(eta2);exp_eta3=exp(eta3);
106 bot=exp_eta1+exp_eta2+exp_eta3;
107 estprobs1=exp_eta{1}/bot;estprobs2=exp_eta{2}/bot;estprobs3=exp_eta{3}/bot;
108 p_outcome=exp_eta{outcome}/bot;
109 ll=log(p_outcome);
110 model outcome~GENERAL(ll);
111 id estprobs2 estprobs3;
112 ods output FitStatistics = ll_slopes;
113 run;
114
115
116 /*************************************************************/
117 /* (4) calibration testing; cf sections 2.2.3. and 2.2.4.  */
118 /*************************************************************/
119
120 /* The -2 log-likelihoods can be obtained in the output from proc nlmixed.
121    Overall calibration can be tested using a likelihood ratio chi-squared
122    test with 4 degrees of freedom to test the joint null hypothesis that
123    intercepts are zero and corresponding slopes 1. If the overall
124    calibration test is statistically significant, separate tests for the
125    calibration intercepts and calibration slopes may be considered, each
126    with 2 df.*/
127
128 /* overall calibration (4df): twice the difference in log-likelihood
129    between the model without adaptations (4) and the model with intercept
130    and slope adaptations (3) */
131
132 /* calibration intercept (2df): twice the difference in log-likelihood
133    between the model without adaptations (4) and the model with intercept
134    adaptation (2) */
135
136 /* calibration slopes (2df): twice the difference in log-likelihood between
137    the model with intercept adaptation (2) and the model with intercept and
138    slope adaptations (3) */
139
140 title j=center font= 'Arial' c=blue bold "Calibration test";
141 title2 j=center font= 'Arial' c=green italic "Test of overall calibration,
        and of calibration intercepts and slopes separately";
142
143 data _null_;set ll_orig;call symput('orig',2*negloglike);run;
144 data _null_;set ll_int;if descr =:'-2 Log' then call symput('int',value);run
        ;
145 data _null_;set ll_slopes;if descr =:'-2 Log' then call symput('sl',value);
        run;
146 data calibrationtest;length test $10;
147 test="Overall";chisq=&orig-&sl;df=4;pval=1-cdf('CHISQUARE',chisq,df);format
        pval pvalue6.4;output;
148 test="Intercepts";chisq=&orig-&int;df=2;pval=1-cdf('CHISQUARE',chisq,df);
        format pval pvalue6.4;output;
```

```
149 test="Slopes";chisq=&int-&sl;df=2;pval=1-cdf('CHISQUARE',chisq,df);format
        pval pvalue6.4;output;
150 run;
151 proc print data=calibrationtest;run;
152
153 %end;
154
155 /*******************************************************/
156 /* (5) Example of calibration plot; cf sections 2.2.2.  */
157 /*******************************************************/
158
159 title j=center font= 'Arial' c=blue bold "Parametric calibration plots";
160
161 data predictedobserved;
162 set mlrsep_logrecal (keep=pred estprobs2 estprobs3 lp2 lp3);
163 rename pred=op1 estprobs2=op2 estprobs3=op3;
164 p1 = 1/(1+exp(lp2)+exp(lp3));
165 p2 = exp(lp2)/(1+exp(lp2)+exp(lp3));
166 p3 = exp(lp3)/(1+exp(lp2)+exp(lp3));
167 run;
168
169 data predictedobserved2;
170 set predictedobserved(rename=op1=op rename=p1=p);cat=1;output;
171 set predictedobserved(rename=op2=op rename=p2=p);cat=2;output;
172 set predictedobserved(rename=op3=op rename=p3=p);cat=3;output;
173 run;
174
175 data predictedobserved;set predictedobserved end=eof;type='real';output;if
        eof then
176 do;type='ideal';p1=0;p2=0;p3=0;opi1=0;opi2=0;opi3=0;op1=.;op2=.;op3=.;output
        ;
177 type='ideal';p1=1;p2=1;p3=1;opi1=1;opi2=1;opi3=1;op1=.;op2=.;op3=.;output;
        output;end;
178 run;
179 data predictedobserved2;set predictedobserved2 end=eof;type='real';output;if
         eof then
180 do;type='ideal';p=0;op=0;cat=99;output;
181 type='ideal';p=1;op=1;cat=99;output;end;
182 run;
183
184 title2 j=center font= 'Arial' c=green italic "Calibration plot for category
        1";
185 axis1 label=(font=arial height=2 justify=center 'Predicted probability of
        cat 1') value=(font=arial height=1.5) length=10.0cm order=(0 to 1 by
        0.1);
186 axis2 label=(font=arial height=2 angle=90 rotate=0 'Observed proportion of
        cat 1') value=(font=arial height=1.5) length=10.0cm order=(0 to 1 by
        0.1);
187 legend1 position=(top left inside) mode=protect across=1 value=(h=1.5 font=
        arial "Calibration curve" "Ideal")  label=none ;
188 proc gplot data=predictedobserved;
189 symbol1 i=sm50s v=dot c=black cv=gray h=0.5 w=2 l=1;
190 symbol2 i=join v=none c=black h=0.5 w=2 l=2;
191 plot (op1 opi1)*p1/overlay haxis=axis1 vaxis=axis2 frame legend=legend1 ;run
        ;
192 title2 j=center font= 'Arial' c=green italic "Calibration plot for category
        2";
193 axis1 label=(font=arial height=2 justify=center 'Predicted probability of
        cat 2') value=(font=arial height=1.5) length=10.0cm order=(0 to 1 by
        0.1);
194 axis2 label=(font=arial height=2 angle=90 rotate=0 'Observed proportion of
        cat 2') value=(font=arial height=1.5) length=10.0cm order=(0 to 1 by
        0.1);
```

```
195 proc gplot data=predictedobserved;
196 symbol1 i=sm50s v=dot c=black cv=gray h=0.5 w=2 l=1;
197 symbol2 i=join v=none c=black h=0.5 w=2 l=2;
198 plot (op2 opi2)*p2/overlay haxis=axis1 vaxis=axis2 legend=legend1 ;run;
199 title2 j=center font= 'Arial' c=green italic "Calibration plot for category
        3";
200 axis1 label=(font=arial height=2 justify=center 'Predicted probability of
        cat 3') value=(font=arial height=1.5) length=10.0cm order=(0 to 1 by
        0.1);
201 axis2 label=(font=arial height=2 angle=90 rotate=0 'Observed proportion of
        cat 3') value=(font=arial height=1.5) length=10.0cm order=(0 to 1 by
        0.1);
202 proc gplot data=predictedobserved;
203 symbol1 i=sm50s v=dot c=black cv=gray h=0.5 w=2 l=1;
204 symbol2 i=join v=none c=black h=0.5 w=2 l=2;
205 plot (op3 opi3)*p3/overlay haxis=axis1 vaxis=axis2 legend=legend1 ;run;
206
207 proc sort data=predictedobserved2;by cat;run;
208 title2 j=center font= 'Arial' c=green italic "Overall calibration plot";
209 axis1 label=(font=arial height=2 justify=center 'Predicted probability')
        value=(font=arial height=1.5) length=10.0cm order=(0 to 1 by 0.1);
210 axis2 label=(font=arial height=2 angle=90 rotate=0 'Observed proportion')
        value=(font=arial height=1.5) length=10.0cm order=(0 to 1 by 0.1);
211 legend2 position=(top left inside) mode=protect across=1 value=(h=1.2 font=
        arial "Calibration curve, 1" "Calibration curve, 2" "Calibration curve,
        3" "Ideal")  label=none ;
212 proc gplot data=predictedobserved2;
213 symbol1 i=sm50s v=none c=blue h=0.5 w=2 l=1;
214 symbol2 i=sm50s v=none c=red h=0.5 w=2 l=1;
215 symbol3 i=sm50s v=none c=green h=0.5 w=2 l=1;
216 symbol4 i=join v=none c=black h=0.5 w=2 l=2;
217 plot (op)*p = cat/overlay haxis=axis1 vaxis=axis2 frame legend=legend2 ;run;
218
219 %mend;
```

Appendix C

Simulation study

This simulation study investigated the influence of the choice of reference category on the calibration slope. We considered a multinomial risk prediction model for a polytomous outcome with three categories. Two large data sets were simulated: development data (n=450000) to generate the prediction model using multinomial logistic regression, and validation data (n=450000) to evaluate the calibration slopes and the chi-squared test statistic (but not the p-value which is of little value with such large sample size). The calibration slopes were evaluated three times: every time another category was chosen as the reference category in the recalibration model. Finally, the three sets of results were compared.

Two normally distributed predictors x_1 and x_2 were considered for the development of the multinomial risk prediction model. The variances of both predictors were fixed at one throughout. Also, for patients with outcome category 1 the means of x_1 and x_2 were fixed at zero in the development as well as in the validation data. The means of x_1 and x_2 were varied for patients with outcome category 2 or 3. This allowed us to vary the amount of discrimination and the predictor distributions in the development and validation data, given the mathematical relation between AUCs and means under normality [141]. Further, we varied the event rates of the three categories in the development and validation data.

First, we varied whether the event rates of the three categories in the development data and validation data were

i identical or

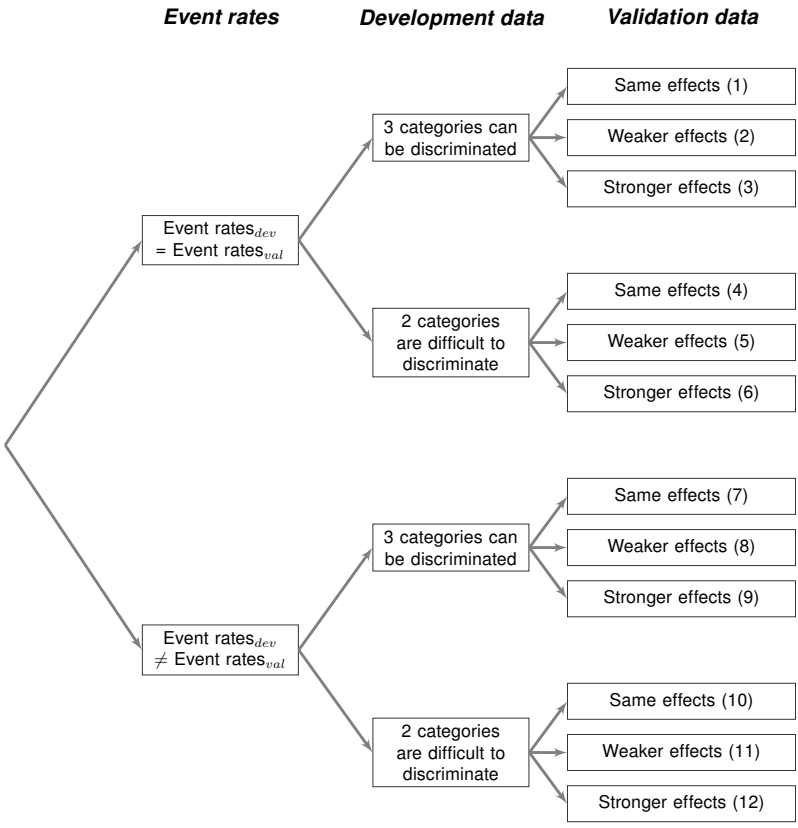


Figure C.1: Schematic overview of the 12 settings used in the simulation study.

ii different.

Second, discrimination between the three categories in the development data was varied by adapting the means of the predictor distributions for the patients in category 2 and in category 3:

i good discrimination between all three categories (c-statistics around 0.75 for each pair of categories) or

ii limited discrimination between two of the three categories (c-statistic of first category versus second and third category 0.73; c-statistic of second versus third category 0.58).

Finally, we varied the similarity of the predictor distributions in the validation data to those in the development data:

i distributions of predictors in the validation were identical to those in the development data,

ii predictor distributions in the validation data were changed to reflect weakened effects of the predictors (means for patients in categories 2 and 3 were moved towards zero),

iii predictor distributions were changed to reflect stronger effects of predictors in the validation data (means for patients in categories 2 and 3 were moved away from zero).

This leads to $2 \cdot 2 \cdot 3 = 12$ simulation settings.

Figure C.1 gives a schematic overview of the 12 settings, while table C.1 provides further details for each setting. Settings 1 and 4 can be labeled reference settings, as event rates and predictor distributions are identical in the development and validation data. In Table C.2 the conditional pairwise c-statistics for development and validation data for the different settings are given [199]. Finally, the results of the simulations can be found in Table C.3. Note that when category i is overfitted (respectively underfitted) with respect to category j, category j is also overfitted (respectively underfitted) with respect to category i (Figure 2.3). Thus we expect that the calibration slope for category 2 vs 1 is the same as the calibration slope for category 1 vs 2. We expect overfitting when effects in the validation data are weaker, and underfitting when effects in the validation data are stronger. This was confirmed by the results: calibration slopes were below 1 for settings 2, 5, 8, and 11, and were above 1 for settings 3, 6, 9, and 12. More importantly, the results suggest that the choice

Table C.1: Details for the 12 settings with respect to development and validation data. The settings with a star are reference settings with identical event rates and predictor distributions for the development and validation data.

Setting	Development data				Validation data				
	N	Event rates (%)	Mean of ($\mathbf{x}_1$,$\mathbf{x}_2$) in cat 2	Mean of ($\mathbf{x}_1$,$\mathbf{x}_2$) in cat 3	N	Event rates (%)	Predictor means in cat 2	Predictor means in cat 3	Change in effects vs development data
Event rates identical, good discrimination between all categories in development data									
1*	450000	50, 33, 17	(0.9,0.3)	(0.3,0.9)	15000	50, 33, 17	(0.9,0.3)	(0.3,0.9)	None
2	450000	50, 33, 17	(0.9,0.3)	(0.3,0.9)	15000	50, 33, 17	(0.5,0.1)	(0.1,0.5)	Weaker
3	450000	50, 33, 17	(0.9,0.3)	(0.3,0.9)	15000	50, 33, 17	(1.2,0.5)	(0.5,1.2)	Stronger
Event rates identical, limited discrimination between two of three categories in development data									
4*	450000	50, 33, 17	(0.7,0.5)	(0.5,0.7)	15000	50, 33, 17	(0.7,0.5)	(0.5,0.7)	None
5	450000	50, 33, 17	(0.7,0.5)	(0.5,0.7)	15000	50, 33, 17	(0.4,0.3)	(0.3,0.4)	Weaker
6	450000	50, 33, 17	(0.7,0.5)	(0.5,0.7)	15000	50, 33, 17	(1.0,0.7)	(0.7,1.0)	Stronger
Event rates different, good discrimination between all categories in development data									
7	450000	50, 33, 17	(0.9,0.3)	(0.3,0.9)	15000	25, 25, 50	(0.9,0.3)	(0.3,0.9)	None
8	450000	50, 33, 17	(0.9,0.3)	(0.3,0.9)	15000	25, 25, 50	(0.5,0.1)	(0.1,0.5)	Weaker
9	450000	50, 33, 17	(0.9,0.3)	(0.3,0.9)	15000	25, 25, 50	(1.2,0.5)	(0.5,1.2)	Stronger
Event rates different, limited discrimination between two of three categories in development data									
10	450000	50, 33, 17	(0.7,0.5)	(0.5,0.7)	15000	25, 25, 50	(0.7,0.5)	(0.5,0.7)	None
11	450000	50, 33, 17	(0.7,0.5)	(0.5,0.7)	15000	25, 25, 50	(0.4,0.3)	(0.3,0.4)	Weaker
12	450000	50, 33, 17	(0.7,0.5)	(0.5,0.7)	15000	25, 25, 50	(1.0,0.7)	(0.7,1.0)	Stronger

Table C.2: Pairwise c-statistics for development and validation data. The settings with a star are reference settings with identical event rates and predictor distributions for the development and validation data.

Setting	Category 1 vs category 2		Category 1 vs category 3		Category 2 vs category 3	
	Development	Validation	Development	Validation	Development	Validation
Event rates identical, good discrimination between all categories in development data						
1*	0.75	0.75	0.75	0.75	0.73	0.73
2	0.75	0.64	0.75	0.64	0.73	0.66
3	0.75	0.82	0.75	0.82	0.73	0.76
Event rates identical, limited discrimination between two of three categories in development data						
4*	0.73	0.73	0.73	0.73	0.58	0.58
5	0.73	0.64	0.73	0.64	0.58	0.54
6	0.73	0.81	0.73	0.81	0.58	0.62
Event rates different, good discrimination between all categories in development data						
7	0.75	0.75	0.75	0.75	0.73	0.73
8	0.75	0.64	0.75	0.64	0.73	0.66
9	0.75	0.82	0.75	0.82	0.73	0.76
Event rates different, limited discrimination between two of three categories in development data						
10	0.73	0.73	0.73	0.73	0.58	0.58
11	0.73	0.64	0.73	0.64	0.58	0.54
12	0.73	0.81	0.73	0.81	0.58	0.62

of reference category has limited influence on the estimation of the calibration slopes or on the test statistic.

Table C.3: Estimated calibration slopes and chi-squared test statistic (2 *df*) for different choices of reference category. The settings with a star are reference settings with identical event rates and predictor distributions for the development and validation data.

Setting	Measure	Reference category = 1		Reference category = 2		Reference category = 3	
		2 vs 1	3 vs 1	1 vs 2	3 vs 2	1 vs 3	2 vs 3
Event rates identical, good discrimination between all categories in development data							
1*	Calibration slope	1.00	1.00	1.00	1.00	1.00	1.00
	Test statistic		0.92		0.03		0.43
2	Calibration slope	0.56	0.55	0.53	0.63	0.52	0.62
	Test statistic		19718.5		200038.98		20196.56
3	Calibration slope	1.33	1.31	1.37	1.23	1.38	1.25
	Test statistic		8608.7742		9163.131		9257.922
Event rates identical, limited discrimination between two of three categories in development data							
4*	Calibration slope	1.00	0.88	0.99	1.01	1.00	0.99
	Test statistic		6.10		2.17		2.01
5	Calibration slope	0.59	0.58	0.59	0.49	0.59	0.52
	Test statistic		12242.44		12289.08		12279
6	Calibration slope	1.42	1.41	1.41	1.46	1.41	1.48
	Test statistic		9713.648		9709.512		9727.546
Event rates different, good discrimination between all categories in development data							
7	Calibration slope	0.99	0.99	0.99	0.99	0.99	0.99
	Test statistic		4.45		4.62		4.46
8	Calibration slope	0.59	0.57	0.52	0.63	0.52	0.65
	Test statistic		17814.81		18323.838		18346.5
9	Calibration slope	1.28	1.30	1.37	1.22	1.36	1.21
	Test statistic		7032.49		7788.99		7693.722

Continued on next page

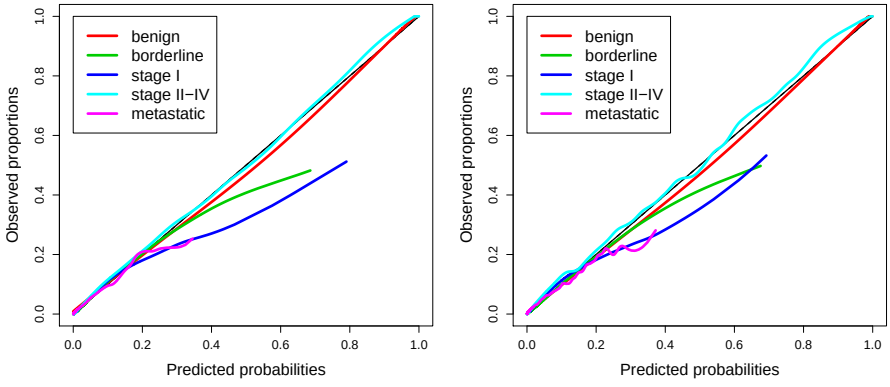
Table C.3 – *Continued from previous page*

Setting	Measure	Reference category = 1		Reference category = 2		Reference category = 3	
		2 vs 1	3 vs 1	1 vs 2	3 vs 2	1 vs 3	2 vs 3
Event rates different, limited discrimination between two of three categories in development data							
10	Calibration slope	1.01	1.00	1.00	1.01	1.00	1.01
	Test statistic		7.49		0.47		0.49
11	Calibration slope	0.58	0.57	0.58	0.51	0.58	0.52
	Test statistic		10737.36		10767.24		10750.5
12	Calibration slope	1.44	1.41	1.41	1.48	1.41	1.49
	Test statistic		8410.296		8406.225		8414.422

Appendix D

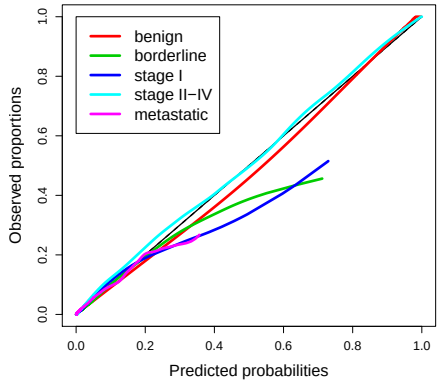
Multiclass calibration plots

D.1 Logistic regression models



(a) MLR (b) LR-BT

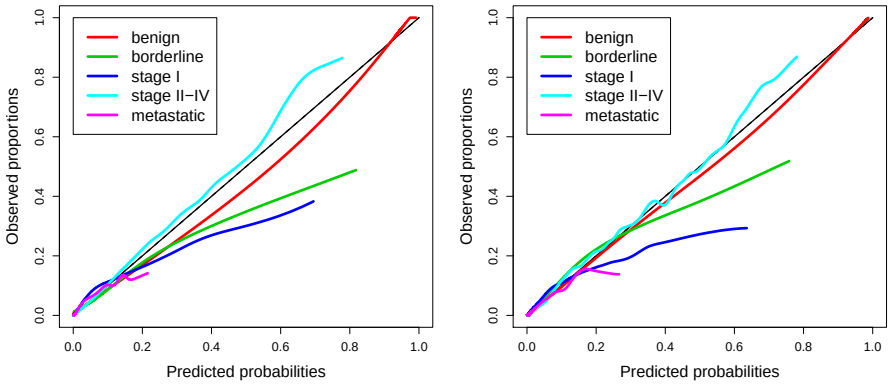
Figure D.1: Smoothed calibration plot of the validation data for (a) multinomial logistic regression, (b) nested dichotomous logistic regression models and (c) dichotomous logistic regression models combined via pairwise coupling.



(c) LR-PC

Continued Figure D.1: Smoothed calibration plot of the validation data for (a) multinomial logistic regression, (b) nested dichotomous logistic regression models and (c) dichotomous logistic regression models combined via pairwise coupling.

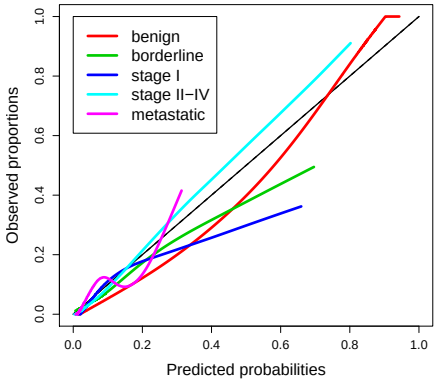
D.2 Machine learning models



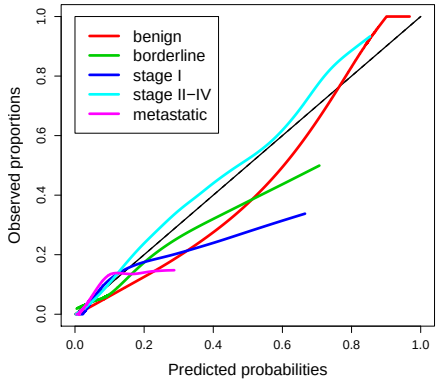
(a) SVM-BT accuracy

(b) SVM-BT logloss

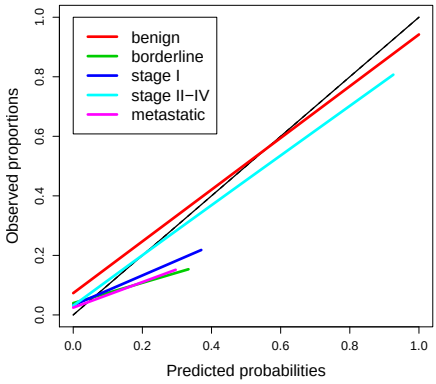
Figure D.2: Smoothed calibration plot of the validation data for (a-b) binary tree support vector machines, (c-d) support vector machines combined via pairwise coupling, (e-f) `k` nearest neighbors, (g-h) random forest, (i-j) naive bayes and (k-l) nearest shrunken centroid with accuracy and logloss as optimization method.



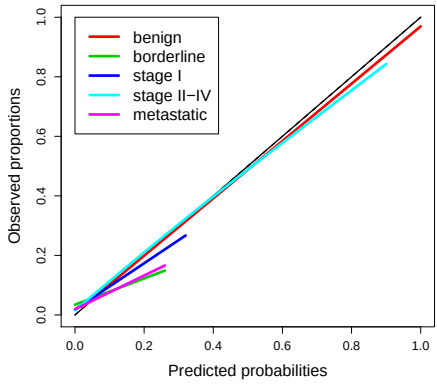
(c) SVM-PC accuracy



(d) SVM-PC logloss

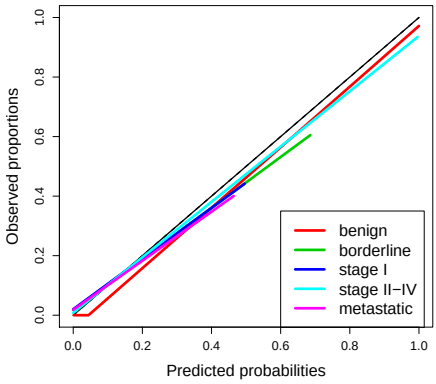


(e) kNN accuracy

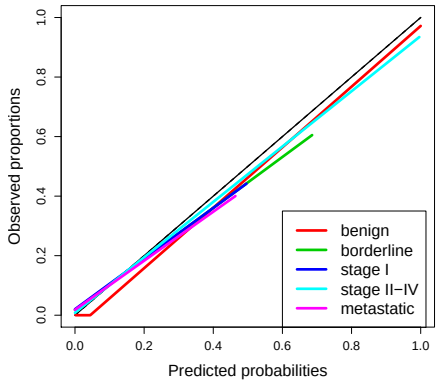


(f) kNN logloss

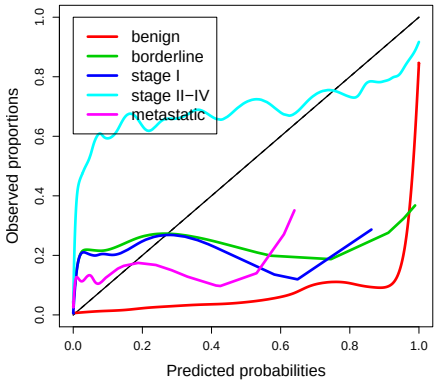
Continued Figure D.2: Smoothed calibration plot of the validation data for (a-b) binary tree support vector machines, (c-d) support vector machines combined via pairwise coupling, (e-f) k nearest neighbors, (g-h) random forest, (i-j) naive bayes and (k-l) nearest shrunken centroid with accuracy and logloss as optimization method.



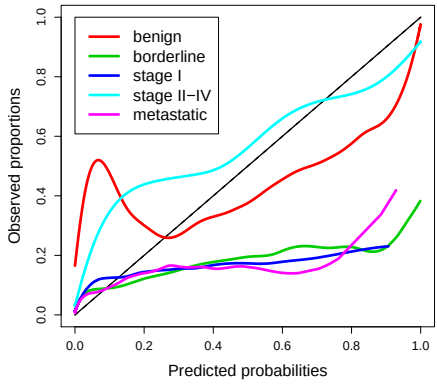
(g) RF accuracy



(h) RF logloss

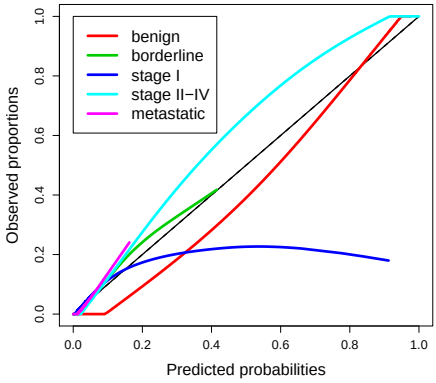


(i) NaB accuracy

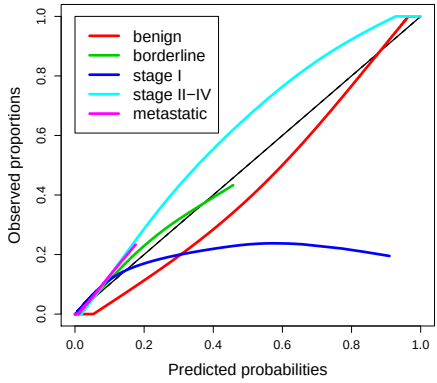


(j) NaB logloss

Continued Figure D.2: Smoothed calibration plot of the validation data for (a-b) binary tree support vector machines, (c-d) support vector machines combined via pairwise coupling, (e-f) `k` nearest neighbors, (g-h) random forest, (i-j) naive bayes and (k-l) nearest shrunken centroid with accuracy and logloss as optimization method.



(k) NSC accuracy



(l) NSC logloss

Continued Figure D.2: Smoothed calibration plot of the validation data for (a-b) binary tree support vector machines, (c-d) support vector machines combined via pairwise coupling, (e-f) `k` nearest neighbors, (g-h) random forest, (i-j) naive bayes and (k-l) nearest shrunken centroid with accuracy and logloss as optimization method.

Bibliography

[1] Abe, S. *Support vector machines for pattern classification*. Springer, 2010.

[2] Agresti, A. *An introduction to categorical data analysis*. Wiley series, United States of America, New York, 1996.

[3] Agresti, A. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2nd edition, 2010.

[4] Agresti, A. *Categorical Data Analysis*. Wiley series, United States of America, 3rd edition, 2013.

[5] Altman, D. and Royston, P. What do we mean by validating a prognostic model? *Statistics in Medicine*, 2000; 19: 453–473. doi:10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5.

[6] Altman, D., Vergouwe, Y., Royston, P., and Moons, K. Prognosis and prognostic research: validating a prognostic model. *BMJ*, 2009; 338: b605. doi:10.1136/bmj.b605.

[7] Austin, P. C. and Steyerberg, E. W. Graphical assessment of interal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, 2014; 33(3): 517–535. doi: 101002/sim.5941.

[8] Baker, S., Cook, N., Vickers, A., and Kramer, B. Using relative utilitiy curves to evaluate risk prediction. *Journal of the Royal Statistical Society Series A*, 2009; 172(4): 729–748. doi:10.1111/j.1467-985X.2009.00592.x.

[9] Baker, S. G. Putting risk prediction in perspective: relative utility curves. *Journal of the National Cancer Institute*, 2009; 101(22): 1538–1542. doi:10.1093/jnci/djp353.

[10] Banerjee, S., Aslam, N., Woelfer, B., Lawrence, A., Elson, J., and Jurkovic, D. Expectant management of early pregnancies of unknown location: a prospective evaluation of methods to predict spontaneous resolution of pregnancy. *BJOG: An International Journal of Obstetrics & Gynaecology*, 2001; 108(2): 158–163.

[11] Barnhart, K., Sammel, M., Appleby, D., Rausch, M., Molinaro, T., Van Calster, B., Kirk, E., Condous, G., Van Huffel, S., Timmerman, D., and Bourne, T. Does a prediction model for pregnancy of unknown location developed in the uk validate on a us population? *Human Reproduction*, 2010; 25(10): 2434–2440. doi: 10.1093/humrep/deq217.

[12] Barnhart, K., van Mello, N. M., Bourne, T., Kirk, E., Van Calster, B., Bottomley, C., Chung, K., Condous, G., Goldstein, S., Hajenius, P. J., Mol, B. W., Molinaro, T., O'Flynn O'Brien, K. L., Husicka, R., Sammel, M., and Timmerman, D. Pregnancy of unknown location: a consensus statement of nomenclature, definitions, and outcome. *Fertility and Sterility*, 2011; 95(3): 857–866. doi:10.1016/j.fertnstert.2010.09.006.

[13] Begg, C. B. and Gray, R. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 1984; 71(1): 11–18. doi:10.2307/2336391.

[14] Beygelzimer, A., Langford, J., and Zadrozny, B. Weighted one-against-all. In *In Proceedings of the 20th national conference on Artificial intelligence*, volume 2. 720–725.

[15] Biesheuvel, C., Vergouwe, Y., Steyerberg, E. W., Grobbee, D., and Moons, K. Polytomous logistic regression analysis could be applied more often in diagnostic research. *Journal of Clinical Epidemiology*, 2008; 61: 125–134. doi:10.1016/j.jclinepi.2007.03.002.

[16] Binder, H. What subject matter questions motivate the use of machine learning approaches compared to statistical models for probability prediction? *Biometrical Journal*, 2014; 56(4): 584–587. doi:10.1002/bimj.201300218.

[17] Bland, J., Altman, D., and Rohlf, F. In defence of logarithmic transformations. *Statistics in Medicine*, 2013; 32(21): 3766–3768. doi: 10.1002/sim.5772.

[18] Bland, J. M. and Altman, D. G. Statistics notes: Logarithms. *BMJ*, 1996; 312(7032): 700. doi:10.1136/bmj.312.7032.700.

[19] Bland, J. M. and Altman, D. G. Statistics notes: Transformations, means, and confidence intervals. *BMJ*, 1996; 312(7038): 1079. doi: 10.1136/bmj.312.7038.1079.

[20] Bouwmeester, W., Twisk, J., Kappen, T., van Klei, W., Moons, K., and Vergouwe, Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Medical Research Methodology*, 2013; 13: 19. doi:10.1186/1471-2288-13-19.

[21] Breiman, L. Random forests. *Machine learning*, 2001; 45(1): 5–32. doi:10.1023/A:1010933404324.

[22] Bull, S. B., Mak, C., and Greenwood, C. M. A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis*, 2002; 39: 57–74. doi:10.1023/A:1010933404324.

[23] Caputo, B., Sim., K., Furesjo, F., and Smola, A. Appearance-based object recognition using svms: which kernel should i use? In *Proceedings of NIPS workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision*. Whistler.

[24] Carley, M. E., Klingele, C. J., Gebhart, J. B., Webb, M. J., and Wilson, T. O. Laparoscopy versus laparotomy in the management of benign unilateral adnexal masses. *The Journal of the American Association of Gynecologic Laparoscopists*, 2002; 9(3): 321–326. doi: 10.1016/S1074-3804(05)60411-2.

[25] Charles, C., Gafni, A., and Whelan, T. Shared decision-making in the medical encounter: what does it mean?(or it takes at least two to tango). *Social Science & Medicine*, 1997; 44(5): 681–692. doi:10.1016/S0277-9536(96)00221-3.

[26] Chen, X. and Ishwaran, H. Random forests for genomic data analysis. *Genomics*, 2012; 99(6): 323–329. doi:10.1016/j.ygeno.2012.04.003.

[27] Condous, G., Okaro, E., and Bourne, T. Ultrasound diagnosis and management of pregnancy of unknown location and ectopic pregnancy. In D. Timmerman, J. Deprest, and T. Bourne, eds., *Ultrasound and Endoscopic Surgery in Obstetrics and Gynaecology*. Springer, 2003; 269–277.

[28] Condous, G., Timmerman, D., Goldstein, S., Valentin, L., Jurkovic, D., and Bourne, T. Pregnancies of unknown location: consensus statement. *Ultrasound in Obstetrics & Gynecology*, 2006; 28(2): 121–122. doi:10.1002/uog.2838.

[29] Condous, G., Van Calster, B., Kirk, E., Haider, Z., Timmerman, D., Van Huffel, S., and Bourne, T. Prediction of ectopic pregnancy in women with a pregnancy of unknown location. *Ultrasound in Obstetrics & Gynecology*, 2007; 29(6): 680–687. doi:10.1002/uog.4015.

[30] Cook, N. R. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation - Journal of the American Heart Association*, 2007; 115: 928–935. doi:10.1161/âĂŃCIRCULATIONAHA.106.672402.

[31] Cook, N. R. Statistical evaluation of prognostic versus diagnostic models: Beyond the roc curve. *Clinical Chemistry*, 2008; 54(1): 17–23. doi:10.1373/clinchem.2007.096529.

[32] Copas, J. B. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1983; 45(3): 311–354. doi:10.2307/2345402.

[33] Coppus, S., van der Veen, F., Opmeer, B., Mol, B., and Bossuyt, P. Evaluating prediction models in reproductive medicine. *Human Reproduction*, 2009; 24(8): 1774–1778. doi:10.1093/humrep/dep109.

[34] Coulter, A. Partnerships with patients: the pros and cons of shared clinical decision-making. *Journal of Health Services Research Policy*, 1997; 2(2): 112–121. doi:10.1177/135581969700200209.

[35] Courvoisier, D. S., Combescure, C., Agoritsas, T., Gayet-Ageron, A., and Perneger, T. V. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology*, 2011; 64(9): 993–1000. doi: 10.1016/j.jclinepi.2010.11.012.

[36] Cox, D. Two further applications of a model for binary regression. *Biometrika*, 1958; 45: 562–565. doi:10.2307/2333203.

[37] Cox, D. R. and Snell, E. J. *Analysis of binary data*, volume 32. CRC Press, 1989.

[38] Cruz, J. A. and Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2006; 2: 59.

[39] Dalton, J. Flexible recalibration of binary clinical prediction models. *Statistics in Medicine*, 2013; 32(2): 282–289. doi:10.1002/sim.5544.

[40] Daraï, E., Fauvet, R., Uzan, C., Gouy, S., Duvillard, P., and Morice, P. Fertility and borderline ovarian tumor: a systematic review of conservative management, risk of recurrence and alternative options.

Human Reproduction Update, 2013; 19: 151–166. doi:10.1093/humupd/dms047.

[41] De Wit, R., Robert, J., Wilkinson, P., de Mulder, P., Mead, G., Fossa, S., Cook, P., De Prijck, L., Stenning, S., and Collette, L. Final analysis demonstrating the equivalence of 3 bep vs 4 cycles and the 5 day schedule vs 3 days per cycle in good prognosis germ cell cancer. an eortc/mrc phase iii study. *Journal of Clinical Oncology*, 2001; 19: 1629–1640.

[42] Derksen, S. and Keselman, H. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 1992; 45(2): 265–282. doi:10.1111/j.2044-8317.1992.tb00992.x.

[43] Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. springer, 1996.

[44] Djavan, B., Remzi, M., Zlotta, A., Seitz, C., Snow, P., and Marberger, M. Novel artificial neural network for early detection of prostate cancer. *Journal of Clinical Oncology*, 2002; 20(4): 921–929.

[45] Dong, L., Frank, E., and Kramer, S. Ensembles of balanced nested dichotomies for multi-class problems. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Springer. Porto, Portugal, 84–95.

[46] Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*, volume 57. CRC press, 1994.

[47] Elwyn, G., Frosch, D., Thomson, R., Joseph-Williams, N., Lloyd, A., Kinnersley, P., Cording, E., Tomson, D., Dodd, C., Rollnick, S., et al. Shared decision making: a model for clinical practice. *Journal of General Internal Medicine*, 2012; 27(10): 1361–1367. doi:10.1007/s11606-012-2077-6.

[48] Fagerland, M., Hosmer, D., and Bofin, A. Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine.*, 2008; 27(21): 4238–4253. doi:10.1002/sim.3202.

[49] Ferri, C., Hernández-Orallo, J., and Salido, M. A. Volume under the roc surface for multi-class problems. In *Proceedings of the 14th European Conference on Machine Learning (ECML2003)*. Springer, Cavtat-Dubrovnick, Coratia, 2003; 108–120. doi:10.1007/978-3-540-39857-8_12.

[50] Fessler, J. A. Nonparametric fixed-interval smoothing with vector splines. *IEEE Transactions on Signal Processing*, 1991; 39(4): 852–859. doi:10.1109/78.80907.

[51] Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika*, 1993; 80(1): 27–38. doi:10.2307/2336755.

[52] Frank, E. and Kramer, S. Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the 21st International Conference on Machine Learning*. Banff, Canada, 39. doi:10.1145/1015330.1015363.

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

[54] Gerds, T. A., Cai, T., and Schumacher, M. The performance of risk prediction models. *Biometrical Journal*, 2008; 50(4): 457–479. doi:10.1002/bimj.200810443.

[55] Gerl, A., Clemm, C., Schmeller, N., Hartenstein, R., Lamerz, R., and Wilmanns, W. Advances in the management of metastatic non-seminomatous germ cell tumours during the cisplatin era: a single-institution experience. *British Journal of Cancer*, 1996; 74(8): 1280–1285.

[56] Goeman, J. and le Cessie, S. A goodness-of-fit test for multinomial logistic regression. *Biometrics*, 2006; 62(4): 980–985. doi:10.1111/j.1541-0420.2006.00581.x.

[57] Golub, G. H. G., Heath, M., and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 1979; 21(2): 215–223. doi:10.1080/00401706.1979.10489751.

[58] Hahlin, M., Thorburn, J., and Bryman, I. The expectant management of early pregnancies of uncertain site. *Human Reproduction*, 1995; 10(5): 1223–1227.

[59] Hand, D. J. Classifier technology and the illusion of progress. *Statistical Science*, 2006; 21(1): 1–14. doi:10.1214/088342306000000060.

[60] Hand, D. J. Naïve bayes. In X. Wu and V. Kumar, eds., *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC, 2009; 163–178.

[61] Hand, D. J. and Till, R. J. A simple generalization of the area under the roc curve for multiple class classification problems. *Machine Learning*, 2001; 45: 171–186. doi:10.1023/A:1010920819831.

[62] Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 1982; 143(1): 29–36. doi:10.1148/radiology.143.1.7063747. PMID: 7063747.

[63] Harrell, F. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, UK, 2001.

[64] Harrell, F. E. J., Lee, K. L., and Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing erros. *Statistics in Medicine*, 1996; 15: 361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[65] Harrell Jr, F. E., Lee, K. L., Matchar, D. B., and Reichert, T. A. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treatment Reports*, 1985; 69(10): 1071–1077.

[66] Harrison, D. A., Brady, A. R., Parry, G. J., Carpenter, J. R., and Kathy, R. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the united kingdom. *Critical Care Medicine*, 2006; 34(5): 1378–1388. doi:10.1097/01.CCM.0000216702.94014.75.

[67] Hastie, T. and Tibshirani, R. *Generalized additive models*. Chapman & Hall, London, 1990.

[68] Hastie, T. and Tibshirani, R. Classification by pairwise coupling. *Annals of Statistics*, 1998; 26(2): 451–471. doi:10.1214/aos/1028144844.

[69] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. *The elements of statistical learning*, volume 2. Springer, 2009.

[70] Hennessy, B. T., Coleman, R. L., and Markman, M. Ovarian cancer. *The Lancet*, 2009; 374(9698): 1371–1382. doi:10.1016/S0140-6736(09)61338-6.

[71] Hilbe, J. M. *Logistic regression models*. CRC Press Boca Raton, 2009.

[72] Hilden, J. Statistical diagnosis based on conditional independence does not require it. *Computers in Biology and Medicine*, 1984; 14(4): 429–435.

[73] Hilden, J. The area under the roc curve and its competitors. *Medical Decision Making*, 1991; 11(2): 95–101.

[74] Hilden, J. and Gerds, T. A. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine*, 2014; 33(19): 3405–3414. doi:10.1002/sim.5804.

[75] Hirsch, R. P. Validation samples. *Biometrics*, 1991; 47(3): 1193–1194.

[76] Hoffmann, T., Légaré, F., Simmons, M., McNamara, K., McCaffery, K., Trevena, L., Hudson, B., Glasziou, P., and Del Mar, C. Shared decision making: what do clinicians need to know and why should they bother? *The Medical Journal of Australia*, 2014; 201(1): 35–39. doi:10.5694/mja14.00002.

[77] Hornik, K., Meyer, D., and Karatzoglou, A. Support vector machines in r. *Journal of Statistical Software*, 2006; 15(9): 1–28.

[78] Hosmer, D., Hosmer, T., Le Cessie, S., and Lemeshow, S. A comparison of goodness-of-fit tests for the logistic regression models. *Statistics in Medicine*, 1997; 16: 965–980. doi:10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O.

[79] Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*. Wiley series, United States of America, 2000.

[80] Hossain, S., Ejaz Ahmed, S., and Howlader, H. A. Model selection and parameter estimation of a multinomial logistic regression model. *Journal of Statistical Computation and Simulation*, 2012; 84(7): 1412–1426. doi:10.1080/00949655.2012.746347.

[81] Ioannidis, J. P., Tarone, R., and McLaughlin, J. K. The false-positive to false-negative ratio in epidemiological studies. *Epidemiology*, 2011; 22(4): 450–456. doi:10.1097/EDE.0b013e31821b506e.

[82] Ivanov, J., Tu, J. V., and Naylor, C. D. Ready-made, recalibrated, or remodeled?: Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation - Journal of the American Heart Association*, 1999; 99: 2098–2104. doi:10.1161/âĂŃ01.CIR.99.16.2098.

[83] Janssen, K., Moons, K., Kalkman, C., Grobbee, D., and Vergouwe, Y. Updating methods improved the performance of a clinical prediction model in new patients. *Journal of Clinical Epidemiology*, 2008; 61: 76–86. doi:10.1016/j.jclinepi.2007.04.018.

[84] Jiang, X., Menon, A., Wang, S., Kim, J., and Ohno-Machado, L. Doubly optimized calibrated support vector machine (doc-svm): An

algorithm for joint optimization of discrimination and calibration. *PloS One*, 2012; 7(11): e48823. doi:10.1371/journal.pone.0048823.

[85] Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 2012; 19(2): 263–274. doi:10.1136/amiajnl-2011-000291.

[86] Job-Spira, N., Fernandez, H., Bouyer, J., Pouly, J.-L., Germain, E., and Coste, J. Ruptured tubal ectopic pregnancy: risk factors and reproductive outcome: results of a population-based study in france. *American Journal of Obstetrics and Gynecology*, 1999; 180(4): 938–944. doi:10.1016/S0002-9378(99)70665-4.

[87] Justice, A., Covinsky, K., and Berlin, J. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, 1999; 130: 515–524. doi:10.7326/0003-4819-130-6-199903160-00016.

[88] Kaijser, J., Bourne, T., Valentin, L., Sayasneh, A., Van Holsbeke, C., Vergote, I., Testa, A., Franchi, D., Van Calster, B., and Timmerman, D. Improving strategies for diagnosing ovarian cancer: a summary of the international ovarian tumor analysis (iota) studies. *Ultrasound in Obstetrics & Gynecology*, 2013; 41(1): 9–20. doi:10.1002/uog.12323.

[89] Kaijser, J., Sayasneh, A., Van Hoorde, K., Ghaem-Maghami, S., Bourne, T., Timmerman, D., and Van Calster, B. Preoperatief onderscheid maken tussen benigne en maligne adnexpathologie: nieuwe evidence-based adviezen voor de dagelijkse praktijk. *Nederlands Tijdschrift voor Obstetrie en Gynaecologie*, 2014; 127: 274–282.

[90] Kaijser, J., Sayasneh, A., Van Hoorde, K., Ghaem-Maghami, S., Bourne, T., Timmerman, D., and Van Calster, B. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. *Human Reproduction Update*, 2014; 20(3): 449–462. doi:10.1093/humupd/dmt059.

[91] Kaijser, J., Van Gorp, T., Van Hoorde, K., Van Holsbeke, C., Sayasneh, A., Vergote, I., Bourne, T., Timmerman, D., and Van Calster, B. A comparison between an ultrasound based prediction model (lr2) and the risk of ovarian malignancy algorithm (roma) to assess the risk of malignancy in women with an adnexal mass. *Gynecologic Oncology*, 2013; 129: 377–383. doi:10.1016/j.ygyno.2013.01.018.

[92] Kattan, M. W. Comparison of cox regression with other methods for determining prediction models and nomograms. *The Journal of Urology*, 2003; 170(6): S6–S10. doi:10.1097/01.ju.0000094764.56269.2d.

[93] Kirk, E., Bottomley, C., and Bourne, T. Diagnosing ectopic pregnancy and current concepts in the management of pregnancy of unknown location. *Human Reproduction Update*, 2014; 20(2): 250–261. doi:10.1093/humupd/dmt047.

[94] Kirk, E., Condous, G., Van Calster, B., Van Huffel, S., Timmerman, D., and Bourne, T. Rationalizing the follow-up of pregnancies of unknown location. *Human Reproduction*, 2007; 22(6): 1744–1750. doi:10.1093/humrep/dem073.

[95] Kleinbaum, D. G. and Klein, M. *Logistic regression: a self-learning text*. Springer, 2010.

[96] Klement, R. J., Allgäuer, M., Appold, S., Dieckmann, K., Ernst, I., Ganswindt, U., Holy, R., Nestle, U., Nevinny-Stickel, M., Semrau, S., Sterzing, F., Wittig, A., Andratschke, N., and Guckenberger, M. Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer. *International Journal of Radiation Oncology* Biology* Physics*, 2014; 88(3): 732–738. doi:10.1016/j.ijrobp.2013.11.216.

[97] König, I., Malley, J., Weimar, C., Diener, H.-C., and Ziegler, A. Practical experiences on the necessity of external validation. *Statistics in Medicine*, 2007; 26: 5499–5511. doi:10.1002/sim.3069.

[98] König, I., Malley, J. D., Pajevic, S., Weimar, C., Diener, H.-C., and Ziegler, A. Patient-centered yes/no prognosis using learning machines. *International Journal of Data Mining and Bioinformatics*, 2008; 2(4): 289–341.

[99] Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D., and Ziegler, A. Probability estimation with machine learning methods for dichotomous and multi-category outcome: Theory. *Biometrical Journal*, 2014; 56(4): 534–563. doi:10.1002/bimj.201300068.

[100] Kruppa, J., Liu, Y., Diener, H.-C., Holste, T., Weimar, C., König, I. R., and Ziegler, A. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical Journal*, 2014; 56(4): 564–583. doi:10.1002/bimj.201300077.

[101] Kruppa, J., Ziegler, A., and König, I. R. Risk estimation and risk prediction using machine-learning methods. *Human Genetics*, 2012; 131(10): 1639–1654. doi:10.1007/s00439-012-1194-y.

[102] Kuss, O. and McLerran, D. A note on the estimation of the multinomial logistic model with correlated responses in sas. *Computer Methods and Programs in Biomedicine*, 2007; 87: 262–269.

[103] Le Cessie, S. and Van Houwelingen, J. Ridge estimators in logistic regression. *Applied Statistics*, 1992; 41(1): 191–201.

[104] Leening, M. J., Steyerberg, E. W., Van Calster, B., D'Agostino, R. B., and Pencina, M. J. Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. *Statistics in Medicine*, 2014; 33(19): 3415–3418. doi:10.1002/sim.6133.

[105] Leening, M. J., Vedder, M. M., Witteman, J. C., Pencina, M. J., and Steyerberg, E. W. Net reclassification improvement: Computation, interpretation, and controversies: A literature review and clinician's guide. *Annals of Internal Medicine*, 2014; 160(2): 122–131. doi:10.7326/M13-1522.

[106] Li, J., Jian, B., and Fine, J. P. Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. *Biostatistics*, 2013; 14(2): 382–394. doi:10.1093/biostatistics/kxs047.

[107] Lisboa, P. J. and Taktak, A. F. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, 2006; 19(4): 408–415. doi:10.1016/j.neunet.2005.10.007.

[108] Macaskill, P. Empirical bayes estimates generated in a hierarchical summary roc analysis agreed closely with those of a full bayesian analysis. *Journal of Clinical Epidemiology*, 2004; 57(9): 925–932. doi:10.1016/j.jclinepi.2003.12.019.

[109] Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R., and Takwoingi, Y. Analysing and presenting results. In J. Deeks, P. Bossuyt, and C. Gatsonis, eds., *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, chapter Chapter 10: Analysing and Presenting Results. The Cochrane Collaboration, version 1.0 edition, 2010; Available from: http://srdta.cochrane.org/.

[110] Mallett, S., Halligan, S., Thompson, M., Collins, G. S., and Altman, D. G. Interpreting diagnostic accuracy studies for patient care. *BMJ*, 2012; 345: e3999. doi:10.1136/bmj.e3999.

[111] Malley, J., Kruppa, J., Dasgupta, A., Malley, K., and Ziegler, A. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, 2012; 51(1): 74. doi:10.3414/ME00-01-0052.

[112] Matheny, M. E., Resnic, F. S., Arora, N., and Ohno-Machado, L. Effects of svm parameter optimization on discrimination and calibration for post-procedural pci mortality. *Journal of Biomedical Informatics*, 2007; 40(6): 688–697. doi:10.1016/j.jbi.2007.05.008.

[113] McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1980; 42(2): 109–142.

[114] Menard, S. Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 2000; 54(1): 17–24. doi:10.2307/2685605.

[115] Menke, J. Bivariate random-effects meta-analysis of sensitivity and specificity with sas proc glimmix. *Methods of Information in Medicine*, 2010; 49(1): 54–62. doi:10.3414/ME09-01-0001.

[116] Miller, M. E. and Hui, W., Siu L.and Tierney. Validation techniques for logistic regression models. *Statistics in Medicine*, 1991; 10: 1213–1226. doi:10.1002/sim.4780100805.

[117] Miller, M. E., Langefeld, C. D., Tierney, W. M., Hui, S. L., and McDonald, C. J. Validation of probabilistic predictions. *Medical Decision Making*, 1993; 13: 49–58. doi:10.1177/0272989X9301300107.

[118] Mittlböck, M. and Schemper, M. Explained variation for logistic regression. *Statistics in Medicine*, 1996; 15(19): 1987–1997.

[119] Moons, K., Altman, D., Vergouwe, Y., and Royston, P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*, 2009; 338: b606. doi:10.1136/bmj.b606.

[120] Moons, K., Royston, P., Vergouwe, Y., Grobbee, D., and Altman, D. Prognosis and prognostic research: what, why, and how? *BMJ*, 2009; 338: b375. doi:10.1136/bmj.b375.

[121] Moons, K. G., de Groot, J. A., Linnet, K., Reitsma, J. B., and Bossuyt, P. M. Quantifying the added value of a diagnostic test or marker. *Clinical Chemistry*, 2012; 58(10): 1408–1417. doi:10.1373/clinchem.2012.182550.

[122] Moons, K. G., Stijnen, T., Michel, B. C., Büller, H. R., Van Es, G.-A., Grobbee, D. E., and Habbema, J. D. F. Application of treatment thresholds to diagnostic-test evaluation an alternative to the comparison of areas under receiver operating characteristic curves. *Medical Decision Making*, 1997; 17(4): 447–454. doi:10.1177/0272989X9701700410.

[123] Moore, R. G., McMeekin, D. S., Brown, A. K., DiSilvestro, P., Miller, M. C., Allard, W. J., Gajewski, W., Kurman, R., Bast Jr, R. C., and Skates, S. J. A novel multiple marker bioassay utilizing he4 and ca125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecologic oncology*, 2009; 112(1): 40–46. doi: 10.1016/j.ygyno.2008.08.031.

[124] Morise, A. P., Diamond, G. A., Detrano, R., Bobbio, M., and Gunel, E. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. *Medical Decision Making*, 1996; 16: 133–142. doi:10.1177/0272989X9601600205.

[125] Mossman, D. Three-way rocs. *Medical Decision Making*, 1999; 19: 78–89. doi:10.1177/0272989X9901900110.

[126] Mulley, A. G., Trimble, C., and Elwyn, G. Stop the silent misdiagnosis: patients' preferences matter. *BMJ*, 2012; 345(1): e6572. doi:10.1136/bmj.e6572.

[127] Nagelkerke, N. J. A note on a general definition of the coefficient of determination. *Biometrika*, 1991; 78(3): 691–692.

[128] Nattino, G., Finazzi, S., and Bertolini, G. Comments on 'graphical assessment of internal and external calibration of logistic regression models by using loess smoothers' by peter c. austin and ewout w. steyerberg. *Statistics in Medicine*, 2014; 33(15): 2696–2698. doi:10.1002/sim.6126.

[129] Nattino, G., Finazzi, S., and 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(14): 2390–2407. doi:10.1002/sim.6100.

[130] Obuchowski, N. A. Receiver operating characteristic curves and their use in radiology. *Radiology*, 2003; 229: 3–8. doi:10.1148/radiol.2291010898.

[131] Obuchowski, N. A., Goske, M. J., and Applegate, K. E. Assessing physicians' accuracy in diagnosing paediatric patients with acute abdominal pain: measuring accuracy for multiple diseases. *Statistics in Medicine*, 2001; 20: 3261–3278. doi:10.1002/sim.944.

[132] Obuchowski, N. A., Lieber, M. L., and Wians, F. H. Roc curves in clinical chemistry: uses, misuses, and possible solutions. *Clinical Chemistry*, 2004; 50(7): 1118–1125. doi:10.1373/clinchem.2004.031823.

[133] Pauker, S. G. and Kassirer, J. P. The threshold approach to clinical decision making. *The New England Journal of Medicine*, 1980; 302(20): 1109–1117. doi:10.1056/NEJM198005153022003.

[134] Peduzzi, P., Concato, J., Kemper, E., Holford, T., and Feinstein, A. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 1996; 49: 1373–1379. doi:10.1016/S0895-4356(96)00236-3.

[135] Peek, N. and Abu-Hanna, A. Clinical prognostic methods: trends and developments. *Journal of Biomedical Informatics*, 2014; 48: 1–4. doi:10.1016/j.jbi.2014.02.016.

[136] Peek, N., Arts, D., Bosman, R., van der Voort, P., and de Keizer, N. External validation of prognostic models for critically ill patients required substantial sample sizes. *Journal of Clinical Epidemiology*, 2007; 60: 491–501. doi:10.1016/j.jclinepi.2006.08.011.

[137] Peirce, C. S. The numerical measure of the success of predictions. *Science*, 1884; 4(93): 453–454.

[138] Pencina, M., D' Agostino, R., D' Agostino, R., and Vasan, R. Evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond. *Statistics in Medicine*, 2008; 27: 157–172. doi:10.1002/sim.944.

[139] Pencina, M., D'Agostino, R., and Steyerberg, E. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine.*, 2011; 30: 11–21. doi:10.1002/sim.4085.

[140] Pencina, M. J., D'Agostino, R. B., D'Agostino, R. B., and Vasan, R. S. Comments on 'integrated discrimination and net reclassification improvements: Practical advice'. *Statistics in Medicine*, 2008; 27(2): 207–212. doi:10.1002/sim.3106.

[141] Pencina, M. J., D'Agostino, R. B., and Massaro, J. M. Understanding increments in model performance metrics. *Lifetime Data Analysis*, 2013; 19(2): 202–218. doi:10.1007/s10985-012-9238-0.

[142] Pepe, M., Fang, J., Feng, Z., Gerds, T., and Hilden, J. The net reclassification index (nri): a misleading measure of prediction improvement with miscalibrated or overfit models. *UW Biostatistics Working Paper Series*, 2013; Working Paper 392.

[143] Pepe, M., Feng, Z., and Gu, J. Comments on 'evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond' by m.j. pencina et al, statistics in medicine. *Statistics in Medicine*, 2008; 27: 173–181. doi:10.1002/sim.2987.

[144] Pepe, M. and Janes, H. Methods for evaluating prediction performance of biomarkers and tests. In M.-L. T. Lee, M. Gail, R. Pfeiffer, G. Satten, T. Cai, and A. Gandy, eds., *Risk Assessment and Evaluation of Predictions*, Lecture Notes in Statistics. Springer New York. ISBN 978-1-4614-8980-1, 2013; 107–142. doi:10.1007/978-1-4614-8981-8_7.

[145] Pepe, M. S. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, 2003.

[146] Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., and Zheng, Y. Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiologyo*, 2008; 167(3): 362–368. doi:10.1093/aje/kwm305.

[147] Pepe, M. S., Janes, H., Kerr, K. F., and Psaty, B. M. Net reclassification index: a misleading measure of prediction improvement. *UW Biostatistics Working Paper Series*, 2013; Working Paper 394.

[148] Pigeon, J. and Heyse, J. An improved goodness of fit statistic for probability prediction models. *Biometrical Journal*, 1999; 41: 71–82. doi: 10.1002/(SICI)1521-4036(199903)41:1<71::AID-BIMJ71>3.0.CO;2-O.

[149] Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola, P. Bartlette, B. Schökopf, and D. Schuurmans, eds., *Advances in large margin classifiers*. Citeseer, MIT Press, Cambridge, 61–74.

[150] Provost, F. and Domingos, P. Tree induction for probability-based ranking. *Machine Learning*, 2003; 52(3): 199–215. doi:10.1023/A: 1024099825458.

[151] Reilly, B. M. and Evans, A. T. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of Internal Medicine*, 2006; 144(3): 201–209. doi:10.7326/ 0003-4819-144-3-200602070-00009.

[152] Rifkin, R. and Klautau, A. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 2004; 5: 101–141.

[153] Rish, I. An empirical study of the naive bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, volume 3. 41–46.

[154] Roukema, J., van Loenhout, R. B., Steyerberg, E. W., Moons, K. G., Bleeker, S. E., and Moll, H. A. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. *Journal of Clinical Epidemiology*, 2008; 61: 135–141. doi:10.1016/j.jclinepi.2007.07.005.

[155] Royston, P. and Sauerbrei, W. *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*, volume 777. John Wiley & Sons, 2008.

[156] Sackett, D. L. and Rosenberg, W. M. C. On the need for evidence-based medicine. *Journal of Public Health Medicine*, 1995; 17(3): 330–334.

[157] Sajda, P. Machine learning for detection and diagnosis of disease. *Annual Review of Biomedical Engineering*, 2006; 8: 537–565. doi:10.1146/annurev.bioeng.8.061505.095802.

[158] Sauerbrei, W., Meier-Hirmer, C., Berner, A., and Royston, P. Multivariable regression model building by using fractional polynomials: description of sas, stata and r programs. *Computational Statistics & Data Analysis*, 2006; 50(12): 3464–3485. doi:10.1016/j.csda.2005.07.015.

[159] Schenker, N. and Taylor, J. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 1996; 22: 425–446. doi:10.1016/0167-9473(95)00057-7.

[160] Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[161] Schwarz, D. F., König, I. R., and Ziegler, A. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, 2010; 26(14): 1752–1758. doi:10.1093/bioinformatics/btq257.

[162] Steinbach, M. and Tan, P.-N. knn: k-nearest neighbors. In X. Wu and V. Kumar, eds., *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC, 2009; 151–162.

[163] Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., Wood, A., and Carpenter, J. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 2009; 338: b2393. doi:10.1136/bmj.b2393.

[164] Steyerberg, E., Eijkemans, M., and Habbema, J. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, 1999; 52: 935–942. doi:10.1016/S0895-4356(99)00103-1.

[165] Steyerberg, E., Harrell, F. J., Borsboom, G., Eijkemans, M., Vergouwe, Y., and Habbema, J. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 2001; 54: 774–781. doi: 10.1016/S0895-4356(01)00341-9.

[166] Steyerberg, E., Keizer, H., Fosså, S., Sleijfer, D., Toner, G., Schraffordt Koops, H., Mulders, P., Messemer, J., Ney, K., and Donohue, J. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. *Journal of Clinical Oncology*, 1995; 13: 1177–1187.

[167] Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, United States of America, 2009.

[168] Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C., and Habbema, J. D. F. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine*, 2004; 23: 2567–2586. doi: 10.1002/sim.1844.

[169] Steyerberg, E. W., Calster, B. V., and Pencina, M. J. Performance measures for prediction models and markers: evaluation of predictions and classifications. *Revista Española de Cardiología (English Edition)*, 2011; 64(9): 788–794. doi:10.1016/j.rec.2011.05.004.

[170] Steyerberg, E. W., Eijkemans, M. J. C., and Habbema, J. D. F. Application of shrinkage techniques in logistic regression analysis: a case study. *Statistica Neerlandica*, 2001; 55(1): 76–88. doi:10.1111/1467-9574.00157.

[171] Steyerberg, E. W., Pencina, M. J., Lingsma, H. F., Kattan, M. W., Vickers, A. J., and Van Calster, B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *European Journal of Clinical Investigation*, 2012; 42(2): 216–228. doi:10.1111/j.1365-2362.2011.02562.x.

[172] Steyerberg, E. W., Schemper, M., and Harrell, F. E. Logistic regression modeling and the number of events per variable: selection bias dominates. *Journal of Clinical Epidemiology*, 2011; 64(12): 1464–1465. doi:10.1016/j.jclinepi.2011.06.016.

[173] Steyerberg, E. W., van der Ploeg, T., and Van Calster, B. Risk prediction with machine learning and regression methods. *Biometrical Journal*, 2014; 56(4): 601–606. doi:10.1002/bimj.201300297.

[174] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 2010; 21(1): 128–138. doi:10.1097/EDE.0b013e3181c30fb2.

[175] Strobl, C., Malley, J., and Tutz, G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 2009; 14(4): 323. doi:10.1037/a0016973.

[176] Taktak, A. F., Eleuteri, A., Lake, S. P., and Fisher, A. C. A web-based tool for the assessment of discrimination and calibration properties of prognostic models. *Computers in Biology and Medicine*, 2008; 38: 785–791. doi:10.1016/j.compbiomed.2008.04.005.

[177] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996; 58(1): 267–288.

[178] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 2002; 99(10): 6567–6572. doi:10.1073/pnas.082099299.

[179] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, 2003; 18(1): 104–117.

[180] Timmerman, D., Ameye, L., Fischerova, D., Epstein, E., Melis, G. B., Guerriero, S., Van Holsbeke, C., Savelli, L., Fruscio, R., Lissoni, A., Testa, A., Veldman, J., Vergote, I., Van Huffel, S., Bourne, T., and Valentin, L. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by iota group. *BMJ*, 2010; 341: c6839. doi:10.1136/bmj.c6839.

[181] Timmerman, D., Bourne, T., De Rijdt, S., Kaijser, J., and Van Calster, B. Characterizing ovarian pathology: refining the performance of ultrasonography. *International Journal of Gynecological Cancer*, 2012; 22: S9–S11. doi:10.1097/IGC.0b013e318251c9df.

[182] Timmerman, D., Bourne, T. H., Tailor, A., Collins, W. P., Verrelst, H., Vandenberghe, K., and Vergote, I. A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model. *American Journal of Obstetrics and Gynecology*, 1999; 181(1): 57–65. doi:10.1016/S0002-9378(99)70436-9.

[183] Timmerman, D., Testa, A., Bourne, T., Ferrazzi, E., Ameye, L., Konstantinovic, M., Van Calster, B., Collins, W., Vergote, I., Van Huffel, S., and Valentin, L. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group. *Journal of Clinical Oncology*, 2005; 23: 8794–8801. doi:10.1200/JCO.2005.01.7632.

[184] Timmerman, D., Valentin, L., Bourne, T., Collins, W., Verrelst, H., and Vergote, I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (iota) group. *Ultrasound in Obstetrics & Gynecology*, 2000; 16: 500–505. doi:10.1046/j.1469-0705.2000.00287.x.

[185] Timmerman, D., Van Calster, B., Testa, A., Guerriero, S., Fischerova, D., Lissoni, A., Van Holsbeke, C., Fruscio, R., Czekierdowski, A., Jurkovic, D., Savelli, L., Vergote, I., Bourne, T., Van Huffel, S., and Valentin, L. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the iota group. *Ultrasound in Obstetrics & Gynecology*, 2010; 36: 226–234. doi:10.1002/uog.7636.

[186] Toll, D., Janssen, K., Vergouwe, Y., and Moons, K. Validation, updating and impact of clinical prediction rules: A review. *Journal of Clinical Epidemiology*, 2008; 61: 1085–1094. doi:10.1016/j.jclinepi.2008.04.008.

[187] Vach, W. Calibration of clinical prediction rules does not just assess bias. *Journal of Clinical Epidemiology*, 2013; 66: 1296–1301. doi:10.1016/j.jclinepi.2013.06.003.

[188] Van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 2011; 45(3): 1–67.

[189] Van Calster, B., Abdallah, Y., Guha, S., Kirk, E., Van Hoorde, K., Condous, G., Preisler, J., Hoo, W., Stalder, C., Bottomley, C., Timmerman, D., and Bourne, T. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. *Human Reproduction*, 2013; 28(3): 609–616. doi:10.1093/humrep/des440.

[190] Van Calster, B., Condous, G., Kirk, E., Bourne, T., Timmerman, D., and Van Huffel, S. An application of methods for the probabilistic three-class classification of pregnancies of unknown location. *Artificical Intelligence in Medicine*, 2009; 46: 139–154. doi:10.1016/j.artmed.2008.12.003.

[191] Van Calster, B., Luts, J., Suykens, J. A., Condous, G., Bourne, T., Timmerman, D., and Van Huffel, S. Comparing methods for multi-class probabilities in medical decision making using ls-svms and kernel logistic regression. In J. Marques de Sá, L. Alexandre, W. Duch, and D. Mandic, eds., *Artificial Neural Networks - ICANN 2007 lecture notes in computer science*. 139–148. doi:10.1007/978-3-540-74695-9_15.

[192] Van Calster, B., Timmerman, D., Bourne, T., Testa, A., Van Hosbeke, C., Domali, E., Jurkovic, D., Neven, P., Van Huffel, S., and Valentin, L. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum ca-125. *Journal of the National Cancer Institute*, 2007; 5: 1706–1714. doi:10.1093/jnci/djm199.

[193] Van Calster, B., Valentin, L., Van Holsbeke, C., Testa, A. C., Bourne, T., Van Huffel, S., and Timmerman, D. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. *BMC Medical Research Methodology*, 2010; 10(96): 1–12. doi:10.1186/1471-2288-10-96.

[194] Van Calster, B., Van Belle, V., Condous, G., Bourne, T., Timmerman, D., and Van Huffel, S. Multi-class auc metrics and weighted alternatives. In D. Liu and R. Kozma, eds., *Proceedings of the 21st International Joint Conference on Neural Networks*. IEEE press, Los Alamitos, 1391–1397.

[195] Van Calster, B., Van Belle, V., Vergouwe, Y., and Steyerberg, E. Discrimination ability of prediction models for ordinal outcomes: Relationships between existing measures and a new measure. *Biometrical Journal*, 2012; 54(5): 674–685. doi:10.1002/bimj.201200026.

[196] Van Calster, B., Van Belle, V., Vergouwe, Y., Timmerman, D., Van Huffel, S., and Steyerberg, E. W. Extending the c statistic to nominal polytomous outcomes: The polytomous discrimination index. *Statistics in Medicine*, 2012; 31: 2610–2626. doi:10.1002/sim.5321.

[197] Van Calster, B., Van Hoorde, K., Valentin, L., Testa, A. C., Fischerova, D., Van Holsbeke, C., Savelli, L., Franchi, D., Epstein, E., Kaijser, J., Van Belle, V., Czekierdowski, A., Guerriero, S., Fruscio, R., Lanzani, C., Scala, F., Bourne, T., Timmerman, D., and International Ovarian Tumour Analysis (IOTA) group. Evaluating the risk of ovarian cancer prior to surgery using the adnex risk model: diagnostic study to differentiate between benign, borderline, stage i invasive, advanced stage invasive, and secondary metastatic tumours. *The BMJ*, 2014; 349: g5920. doi:10.1136/bmj.g5920.

[198] Van Calster, B. and Van Huffel, S. Integrated discrimination improvement and probability-sensitive auc variants. *Statistics in Medicine*, 2010; 29(2): 318–319. doi:10.1002/sim.3761.

[199] Van Calster, B., Vergouwe, Y., Van Belle, V., Looman, C., Timmerman, D., and Steyerberg, E. Assessing the discriminative ability of risk models for more than two outcome categories: a perspective. *European Journal of Epidemiology*, 2012; 27(10): 761–770. doi:10.1007/s10654-012-9733-3.

[200] Van Calster, B., Vickers, A. J., Pencina, M. J., Baker, S. G., Timmerman, D., and Steyerberg, E. W. Evaluation of markers and risk prediction models overview of relationships between nri and decision-analytic measures. *Medical Decision Making*, 2013; 33(4): 490–501. doi:10.1177/0272989X12470757.

[201] van der Ploeg, T., Smits, M., Dippel, D. W., Hunink, M., and Steyerberg, E. W. Prediction of intracranial findings on ct-scans by alternative modelling techniques. *BMC Medical Research Methodology*, 2011; 11(1): 143. doi:10.1186/1471-2288-11-143.

[202] Van Esbroeck, A., Rubinfeld, I., Hall, B., and Syed, Z. Quantifying surgical complexity with machine learning: Looking beyond patient factors to improve surgical models. *Surgery*, 2014; doi:10.1016/j.surg.2014.04.034.

[203] Van Holsbeke, C., Van Calster, B., Bourne, T., Ajossa, S., Testa, A., Guerriero, S., Fruscio, R., Lissoni, A., Czekierdowski, A., Savelli, L., Van Huffel, S., Valentin, L., and D, T. External validation of diagnostic models to estimate the risk

of malignancy in adnexal masses. *Clinical Cancer Research*, 2012; 18: 815–825. doi:10.1158/1078-0432.CCR-11-0879.

[204] Van Holsbeke, C., Van Calster, B., Testa, A., Domali, E., Lu, C., Van Huffel, S., Valentin, L., and Timmerman, D. Prospective interval validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. *Clinical Cancer Research*, 2009; 15: 648–691. doi:10.1158/1078-0432.CCR-08-0113.

[205] Van Hoorde, K., Vergouwe, Y., Timmerman, D., Van Huffel, S., Steyerberg, E., and Van Calster, B. Simple dichotomous updating methods improved the validity of polytomous models. *Journal of Clinical Epidemiology*, 2013; 66: 1158–1165. doi:10.1016/j.jclinepi.2013.04.014.

[206] Van Hoorde, K., Vergouwe, Y., Timmerman, D., Van Huffel, S., Steyerberg, E. W., and Van Calster, B. Assessing calibration of multinomial risk prediction models. *Statistics in Medicine*, 2014; 33(15): 2585–2596. doi:10.1002/sim.6114.

[207] van Houwelingen, H. and Thorogood, J. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine*, 1995; 14: 1999–2008. doi:10.1002/sim.4780141806.

[208] van Houwelingen, J. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica*, 2001; 55(1): 17–34. doi:10.1111/1467-9574.00154.

[209] van Houwelingen, J. and Le Cessie, S. Logistic regression, a review. *Statistica Neerlandica*, 1988; 42(4): 215–232. doi:10.1111/j.1467-9574.1988.tb01237.x.

[210] van Houwelingen, J. C. and Le Cessie, S. Predictive value of statistical models. *Statistics in Medicine*, 1990; 9: 1303–1325. doi:10.1002/sim.4780091109.

[211] Vandorpe, T., Smeets, A., Van Calster, B., Van Hoorde, K., Leunen, K., Amant, F., Moerman, P., Deraedt, K., Brouckaert, O., Van Huffel, S., Wildiers, H., Christiaens, M., and Neven, P. Lobular and non-lobular breast cancers differ regarding axillary lymph node metastasis: a cross-sectional study on 4,292 consecutive patients. *Breast Cancer Research and Treatment*, 2011; 128(2): 429–435. doi:10.1007/s10549-011-1565-4.

[212] Vapnik, V. *The nature of statistical learning theory*. Springer, 2000.

[213] Vapnik, V. N. and Vapnik, V. *Statistical learning theory*, volume 2. Wiley New York, 1998.

[214] Vergouwe, Y., Moons, K. G., and Steyerberg, E. W. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology*, 2010; 172: 971–980. doi:10.1093/aje/kwq223.

[215] Vergouwe, Y., Nieboer, D., Debray, T., Murray, G., Kattan, M., Oostenbrink, R., Koffijberg, H., Moons, K., and Steyerberg, E. A closed testing procedure to select an appropriate method to update prediction models. *Submitted to Statistics in Medicine*; .

[216] Vergouwe, Y., Royston, P., Moons, K. G., and Altman, D. G. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology*, 2010; 63(2): 205–214. doi:10.1016/j.jclinepi.2009.03.017.

[217] Vergouwe, Y., Steyerberg, E., Eijkemans, M., and Habbema, J. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology*, 2005; 58(5): 475–483. doi:10.1016/j.jclinepi.2004.06.017.

[218] Vergouwe, Y., Steyerberg, E. W., Foster, R. S., Habbema, J., and Donohue, J. P. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. *The Journal of Urology*, 2001; 165(1): 84–88. doi: 10.1097/00005392-200101000-00021.

[219] Vergouwe, Y., Steyerberg, E. W., Foster, R. S., Sleijfer, D. T., Fosså, S. D., Gerl, A., de Wit, R., Roberts, J. T., and Habbema, J. D. F. Predicting retroperitoneal histology in postchemotherapy testicular germ cell cancer: A model update and multicentre validation with more than 1000 patients. *European Urology*, 2007; 51: 424–432. doi:10.1016/j.eururo.2006.06.047.

[220] Viallon, V., Ragusa, S., Clavel-Chapelon, F., and Bénichou, J. How to evaluate the calibration of a disease risk prediction tool. *Statistics in Medicine*, 2009; 28: 901–916. doi:10.1002/sim.3517.

[221] Vickers, A., Cronin, A., and Begg, C. One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology*, 2011; 11: 13. doi:10.1186/1471-2288-11-13.

[222] Vickers, A. J. and Elkin, E. B. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 2006; 26: 565–574. doi:10.1177/0272989X06295361.

[223] Vittinghoff, E. and McCulloch, C. E. Relaxing the rule of 10 events per variable in logistic and cox regression. *American Journal of Epidemiology*, 2007; 165(6): 710–718. doi:10.1093/aje/kwk052.

[224] Wu, T., Lin, C., and Weng, R. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 2004; 5: 975–1005.

[225] Wynants, L., Bouwmeester, W., Moons, K., Moerbeek, M., Timmerman, D., Van Huffel, S., Van Calster, B., and Vergouwe, Y. Prediction performance of multilevel logistic regression models: a simulation study of the number of events per variable. *Submitted to Journal of Clinical Epidemiology*; .

[226] Wynants, L., Timmerman, D., Bourne, T., Van Huffel, S., and Van Calster, B. Screening for data clustering in multicenter studies: the residual intraclass correlation. *BMC Medical Research Methodology*, 2013; 13(1): 128. doi:10.1186/1471-2288-13-128.

[227] Xue, H., Yang, Q., and Chen, S. Svm: Support vector machines. In X. Wu and V. Kumar, eds., *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC, 2009; 37–59.

[228] Yates, J. F. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 1982; 30(1): 132–156. doi:10.1016/0030-5073(82)90237-9.

[229] Yee, T. The vgam package for categorical data analysis. *Journal of Statistical Software*, 2010; 32(10): 1–34.

[230] Yee, T. and Wild, C. Vector generalized additive models. *Journal of the Royal Statistical Society Series B (Methodological)*, 1996; 58(3): 481–493.

[231] Yee, T. W. *The Analysis of binary data in quantitative plant ecology*. Ph.D. thesis, Department of Mathematics and Statistics, University of Auckland, Auckland, 1993.

[232] Yee, T. W. Vector splines and other vector smoothers. In J. G. Betlehem and P. G. M. van der Heijden, eds., *COMPSTAT: Proceedings in Computational Statistics*. Springer, 529–534.

[233] Youden, W. Index for rating diagnostic tests. *Cancer*, 1950; 3(1): 32–35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[234] Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eight International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, 694–699.

[235] Ziegler, A. Rejoinder. *Biometrical Journal*, 2014; 56(4): 607–613. doi:10.1002/bimj.201400010.

Curriculum Vitae

Kirsten Van Hoorde was born in Oudenaarde on April 6, 1987. In 2004 she started studying Mathematics at the KU Leuven. After obtaining her Bachelor in Mathematics, she started a Master of Statistics, options Biometrics and Research, at the KU Leuven. In addition, she started studying Specific Teacher Training in Mathematics ('Specifieke Lerarenopleiding Wiskunde') in 2008 and obtained her degree, magna cum laude, in July 2010. She received the degree of Master of Statistics, magna cum laude, with a thesis on the development of a risk of unviable pregnancy index, under supervision of prof. dr. ir Sabine Van Huffel and prof. dr. ir. Johan Suykens in July 2010. In August 2010, she started her PhD at the Department of Electrical Engineering (ESAT) of the KU Leuven under the supervision of prof. dr. ir. Sabine Van Huffel and prof. dr. Ben Van Calster resulting in this thesis. She received a grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen) in 2010.

Publication list

Papers in international journals

Kaijser, J., Van Gorp, T., **Van Hoorde, K.**, Van Holsbeke, C., Sayasneh, A., Vergote, L., Bourne, T., Timmerman, D., Van Calster, B. A comparison between an ultrasound based prediction model (LR2) and the Risk of Ovarian Malignancy Algorithm (ROMA) to assess the risk of malignancy in women with an adnexal mass. *Gynecologic Oncology*, **129**(2), 377–383 (2013).

Kaijser, J.*, Sayasneh, A.*, **Van Hoorde, K.**, Ghaem–Maghami, S., Bourne, T., Timmerman, D., Van Calster, B. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. *Human Reproduction Update*, **20**(3), 449–462 (2014).

Kaijser, J.*, Sayasneh, A.*, **Van Hoorde, K.**, Ghaem–Maghami, S., Bourne, T., Timmerman, D., Van Calster, B. Preoperatief onderscheid maken tussen benigne en maligne adnexpathologie: nieuwe evidence-based adviezen voor de dagelijkse praktijk. *Nederlands Tijdschrift voor Obstetrie en Gynaecologie*, **127**, 274–282 (2014).

Timmerman, D., Van Calster, B., Vergote, I., **Van Hoorde, K.**, Van Gorp, T., Valentin, L., Bourne, T. Letter to the Editor of Obstetrics and Gynecology: Performance of the American College of Obstetricians and Gynecologists' ovarian tumor referral guidelines with a multivariate index assay. *Obstetrics & Gynecology* **118**(5), 1179–1181 (2011).

Van Calster, B., Abdallah, Y., Guha, S., Kirk, E., **Van Hoorde, K.**, Condous, G., Preisler, J., Hoo, W., Stalder, C., Bottomley, C., Timmerman, D., Bourne, T. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. *Human Reproduction* **28**(3), 609–616 (2013).

Van Calster, B., **Van Hoorde, K.**, Valentin, L., Testa, A.C., Fischerova, D.,

Van Holsbeke, C., Savelli, L., Franchi, D., Epstein, E., Kaijser, J., Van Belle, V., Czekierdowski, A., Guerriero, S., Fruscio, R., Lanzani,C., Scala, F., Bourne, T., Timmerman, D., International Ovarian Tumour Analysis (IOTA) group. Evaluating the risk of ovarian cancer prior to surgery using the ADNEX risk model: diagnostic study to differentiate between benign, borderline, stage I invasive, advanced stage invasive, and secondary metastatic tumours. *BMJ* **349**, g5920 (2014).

Van Hoorde, K., Vergouwe, Y., Timmerman, D., Van Huffel, S., Steyerberg, E.W., Van Calster, B. Simple dichotomous updating methods improved the validity of polytomous models. *Journal of Clinical Epidemiology* **66**(10), 1158–1165 (2013).

Van Hoorde, K., Vergouwe, Y., Timmerman, D., Van Huffel, S., Steyerberg, E., Van Calster, B. Assessing calibration of multinomial risk prediction models. *Statistics in Medicine* **33**(15), 2585–2596 (2014).

Vandorpe, T., Smeets, A., Van Calster, B., **Van Hoorde, K.**, Leunen, K., Amant, F., Moerman, P., Deraedt, K., Brouckaert, O., Van Huffel, S., Wildiers, H., Christiaens, M. R., Neven, P. Lobular and non–lobular breast cancers differ regarding axillary lymph node metastasis: a cross–sectional study on 4,292 consecutive patients *Breast Cancer Research and Treatment* **128**(2), 429–435 (2011).

Papers submitted to international journals

Abdallah, Y., **Van Hoorde, K.**, Preisler, J., Naji, O., Sayasneh, A., Saso, S., Afshan, I., Bradburn, E., Stalder, C., Barnhart, K., Timmerman, D., Van Calster, B., Bourne, T. How do we define a "pregnancy of unknown location": can we reliably diagnose an intrauterine or ectopic pregnancy when a yolk sac or embryo has not been visualised? Submitted to *Obstetrics & Gynecology*.

Kaijser, J., **Van Hoorde, K.**, Van Calster, B., Bourne, T., Vergote, I., Timmerman, D. Diagnosing adnexal tumours before surgery: a critical appraisal of recent evidence. Submitted to *The Obstetrician & Gynaecologist*.

Van Hoorde, K., Vergouwe, Y., Steyerberg, E.W., Bobdiwala, S., Condous, G., Kirk, E., Van Huffel, S., Bourne, T., Van Calster, B. Updating risk prediction models for multinomial outcomes: overview and illustration. Submitted to *Statistics in Medicine*.

Van Hoorde, K., Van Huffel, S., Timmerman, D., Van Calster, B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. Submitted to *Journal of Biomedical Informatics*.

Papers in preparation

Pexsters, A., **Van Hoorde, K.**, Bobdiwala, S., Van Huffel, S., Schildermans, S., D'Hooghe, T. , Van Calster, B., Bourne, T., Timmerman, D. Predicting viability in IVF pregnancies using a model based on serum biochemistry at 4 and 5 weeks gestation.

Van Calster, B., Bobdiwala, S., Guha, S., **Van Hoorde, K.**, Al-Memar, M., Kirk, E., Condous, G., Sur, S., Stalder, C., Timmerman, D., Bourne, T. Managing pregnancy of unknown location: a two-step approach using initial serum progesterone and a novel risk score based on serial serum hCG levels.

Abstracts in international journals

Abdallah, Y., Condous, G., **Van Hoorde, K.**, Kirk, E., Hoo, P., Stalder, C., Bottomley, C., Van Huffel, S., Timmerman, D., Van Calster, B., Bourne, T. Rationalizing the management of pregnancies of unknown location (PULs) by reliably selecting failed PUL and intrauterine pregnancies for minimal follow up. *Ultrasound in Obstetrics and Gynecology* **38**(Supplement 1), 62 (2011).

Abdallah, Y., Hoo, P., **Van Hoorde, K.**, Naji, O., Van Calster, B., Kirk, E., Pexsters, A., Bottomley, C., Stalder, C., Van Huffel, S., Timmerman, D., Bourne, T. Serial hCG assays for the management of pregnancies of unknown location (PUL): does the timing of hCG samples matter? *Ultrasound in Obstetrics and Gynecology* **38**(Supplement 1), 12-13 (2011).

Abdallah, Y., **Van Hoorde, K.**, Naji, O., Van Calster, B., Kirk, E., Pexsters, A., Husicka, R., Stalder, C., Van Huffel, S., Timmerman, D., Bourne, T. Comparing the subjective impression of the examiner compared to the hCG ratio and mathematical models for the management of pregnancies of unknown location (PUL). *Ultrasound in Obstetrics and Gynecology* **38**(Supplement 1), 14 (2011).

Kaijser, J., Van Gorp, T., **Van Hoorde, K.**, Van Holsbeke, C., Vergote, I., Timmerman, D., Van Calster, B. Serum CA-125 and HE-4 versus an ultrasound based predictive model to assess the risk of malignancy in women with an adnexal mass. *The International Journal of Gynecological Cancer* **22**(8, Supplement 3), 149–150 (2012).

Pexsters, A., **Van Hoorde, K.**, Van Calster, B., Van Huffel, S., Kirk, E., Abdallah, Y., Bottomley, C., Schildermans, S., D'Hooghe, T., Bourne, T., Timmerman, D. Using a model based on serum biochemistry at 4 and 5 weeks gestation to predict immediate viability and viability and 12 weeks in IVF

pregnancies. *Ultrasound in Obstetrics and Gynecology* **38**(Supplement 1), 25 (2011).

Pexsters, A., **Van Hoorde, K.**, Van Calster, B., Van Huffel, S., Kirk, E., Abdallah, Y., Naji, O., Schildermans, S., D'Hooghe, T., Bourne, T., Timmerman, D. Behaviour of serum hormone levels at 4 and 5 weeks gestation in an IVF treatment population. *Ultrasound in Obstetrics and Gynecology* **38**(Supplement 1), 61 (2011).

Abstracts presented at international conferences

Van Hoorde, K., Vergouwe, Y., Timmerman, D., Van Huffel, S., Steyerberg, E.W., Van Calster, B. Updating of polytomous risk prediction models based on sequential dichotomous modeling improves the performance. Abstract presented at the 33th Annual Conference of the International Society for Clinical Biostatistics, Bergen, Norway 2012.

Van Hoorde, K., Van Huffel, S., Timmerman, D., Van Calster, B. Assessing calibration of risk prediction models for polytomous outcomes. Abstract presented at the ENAR Spring meeting, Orlando, Florida 2013.

Van Hoorde, K., De Raedt, K., Timmerman, D., Van Huffel, S., Van Calster, B. Comparing different methods to develop prediction models for polytomous outcomes. Abstract presented at the 35th Annual Conference of the International Society for Clinical Biostatistics, Vienna, Austria 2014.

Invited talks

Van Hoorde, K., Wynants, L. Meet the expert 'round table' sessions: Biostatistics made easy at the 1st IOTA (International Ovarian Tumor Analysis) Congress, Leuven, Belgium 2013.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING (ESAT)
STADIUS: BIOMED
Kasteelpark Arenberg 10 box 2446
B-3001 Leuven
Kirsten.VanHoorde@esat.kuleuven.be
http://www.esat.kuleuven.be

