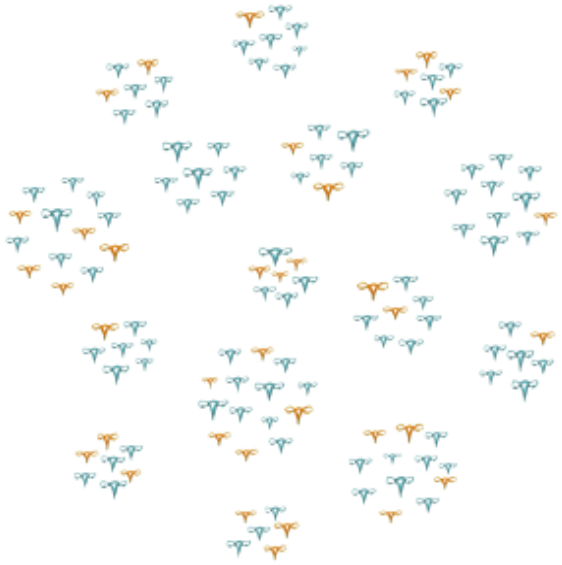
KU LEUVEN

ARENBERG DOCTORAL SCHOOL
FACULTY OF ENGINEERING SCIENCE

Clinical Risk Prediction Models Based On Multicenter Data

Methods for Model Development and Validation



Laure Wynants

Supervisors:
Prof. dr. ir. S. Van Huffel
Prof. dr. B. Van Calster
Prof. dr. D. Timmerman

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Engineering Science

December 2016

CLINICAL RISK PREDICTION MODELS BASED ON MULTICENTER DATA

METHODS FOR MODEL DEVELOPMENT AND VALIDATION

Laure WYNANTS

Supervisors:
Prof. dr. ir. S. Van Huffel
Prof. dr. B. Van Calster
Prof. dr. D. Timmerman

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Engineering Science (PhD)

Members of the Examination Committee:
Prof. dr. Y. Vergouwe
Prof. dr. A. De Meyer
Prof. dr. T. Debray
Prof. dr. T. Bourne
Prof. dr. ir. H. Hens

December 2016

© 2016 KU Leuven, Science, Engineering & Technology
Uitgegeven in eigen beheer, Laure Wynants, Kasteelpark Arenberg 10 box 2446, B-3001 Leuven

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 voorafgaandelijke 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.

Acknowledgments

I wish to thank my supervisors, Professor Sabine Van Huffel and Professor Ben Van Calster, for creating the opportunity to embark on this research project and for their generous advice throughout the past years. Sabine and Ben, thank you for your confidence in me. I am also grateful to my co-supervisor, Professor Dirk Timmerman, for allowing me to work with a wonderful dataset and for the clinical insights he gave me.

I am indebted to the members of my jury, Professor Anne-Marie De Meyer, Professor Yvonne Vergouwe, Professor Thomas Debray, Professor Tom Bourne, and Professor Hugo Hens. Thank you for your interest in my work. Yvonne, thank you for the nice collaboration during my research stay in your group at the Erasmus MC. Tom, thank you for the exciting research projects I got to work on.

I have been a PhD fellow of the Flemish institute for Innovation by Science and Technology (IWT) and received travel grants from the Research Foundation–Flanders (FWO). I am grateful for the financial support I received.

My life as a PhD researcher wouldn't have been half as much fun without all my BioMEd colleagues, past and present. Thanks for all the discussions, the laughs, the food, the sports and the parties. I will miss you all. I also thank Chiara, Wouter, Bavo and Jan, my collaborators at the university hospital. I look forward to our future working together.

I am obliged for the help I received from the administrative and technical staff at ESAT (and for their patience with me).

Last but certainly not least, I'd like to thank my friends, family and family-in-law, for their encouraging words and their interest in my research. To my lovely, crazy sisters, Amber and Myrthe, thank you for designing the artwork and the invitation for this PhD. To my parents, thank you for everything. Papa, saddle up, it is time to live up to your promise. To my cat, my most faithful home-office mate, thanks for keeping me company. To my partner, Stijn, thank you for your help with proofreading and editing this dissertation. But most of all, thank you for your love and support.

Nederlandse samenvatting

Risicopredictiemodellen worden ontwikkeld om artsen te helpen bij het stellen van een diagnose, het maken van een medische beslissing of het geven van een prognose. Om er voor te zorgen dat risicomodellen veralgemeenbaar zijn, verzamelen onderzoekers steeds vaker data van patiënten in verschillende ziekenhuizen door samen te werken aan zogenaamde multicentrische projecten. De resulterende datasets zijn geclusterd: patiënten van eenzelfde centrum kunnen meer met elkaar gemeen hebben dan patiënten van verschillende centra. Dat komt bijvoorbeeld door regionale verschillen in populaties en lokale doorverwijspatronen. Bijgevolg is de assumptie van onafhankelijke observaties, die gemaakt wordt door de meest gebruikte analysetechnieken (bv. logistische regressie), ongeldig. Dit wordt genegeerd in het meeste predictieonderzoek. Onderzoek dat foute assumpties maakt kan niettemin misleidende resultaten geven en leiden tot suboptimale verbeteringen in de patiëntenzorg.

Om dit probleem aan te kaarten heb ik onderzocht wat de gevolgen zijn van de aanname dat observaties onafhankelijk zijn en alternatieve technieken bestudeerd die clustering erkennen, voor het hele proces van het plannen van een studie, het bouwen van een model en het valideren van voorspellingen in nieuwe data. Ik gebruikte hiervoor *mixed* en *random effects* modellen, omdat hiermee verschillen tussen centra gemodelleerd kunnen worden. De voorgestelde oplossingen werden geëvalueerd met simulatiestudies en klinische data. Dit proefschrift behandelt de benodigde steekproefgrootte, de dataverzameling en de selectie van predictoren, het schatten van het model, en de validatie van risicovoorspellingen in nieuwe datasets. Hierbij ligt de focus voornamelijk op diagnostische modellen. De belangrijkste casestudy is het ontwikkelen en valideren van modellen voor de preoperatieve diagnose van

ovariumkanker, waarvoor de multicentrische dataset verzameld door de *International Ovarian Tumor Analysis*-groep (IOTA) gebruikt werd.

De resultaten suggereren dat mixed effects logistische regressiemodellen centrumspecifieke predicties geven die beter presteren in nieuwe patiënten dan de predicties van standaard logistische regressiemodellen. Hoewel simulaties aantoonden dat modellen toevallige patronen ongewenst oppikten in kleine datasets, hadden mixed effects modellen niet meer data nodig dan standaard logistische regressiemodellen. Uit een casestudy van predictoren voor ovariumkanker bleek dat metingen systematisch kunnen verschillen tussen centra in multicentrische datasets. Deze predictoren konden gedetecteerd worden met de residuele intraklasse correlatiecoëfficiënt en kunnen uitgesloten worden bij het bouwen van een risicomodel. Bovendien toonde een casestudy aan dat mixed effects modellen nodig zijn in elke stap van de selectieprocedure wanneer statistische variabelenselectiemethoden gebruikt worden, dit om te voorkomen dat de inferenties incorrect zijn. Tot slot demonstreerden casestudy's over ovariumkanker dat de voorspellende kracht van risicomodellen verschilde van centrum tot centrum. Dit kon aangetoond worden door het gebruik van modellen voor de meta-analyse van discriminatie, kalibratie en klinisch nut.

Ter conclusie, het in rekening brengen van verschillen tussen centra gedurende het plannen van predictieonderzoek, de ontwikkeling van een model en de validatie van de voorspelde risico's in nieuwe patiënten biedt inzichten in de heterogeniteit en betere predicties in lokale omstandigheden. Er resten nog veel methodologische uitdagingen, waaronder de inclusie van interacties tussen predictoren en centra, de optimale toepassing van mixed effects modellen in nieuwe centra en de verfijning van technieken om klinisch nut samen te vatten op basis van multicentrische data. De bevindingen in dit proefschrift wijzen er echter op dat toegepast predictieonderzoek baat zou hebben bij de implementatie van mixed en random effects technieken om ten volle gebruik te maken van alle informatie aanwezig in multicentrische studies.

Abstract

Risk prediction models are developed to assist doctors in diagnosing patients, decision-making, counseling patients or providing a prognosis. To enhance the generalizability of risk models, researchers increasingly collect patient data in different settings and join forces in multicenter collaborations. The resulting datasets are clustered: patients from one center may have more similarities than patients from different centers, for example, due to regional population differences or local referral patterns. Consequently, the assumption of independence of observations, underlying the most often used statistical techniques to analyze the data (e.g., logistic regression), does not hold. This is mostly ignored in much of the current clinical prediction research. Research that relies on faulty assumptions may yield misleading results and lead to suboptimal improvements in patient care.

To address this issue, I investigated the consequences of ignoring the assumption of independence and studied alternative techniques that acknowledge clustering throughout the process of planning a study, building a model and validating models in new data. I used mixed and random effects methods throughout the research as they allow to explicitly model differences between centers, and evaluated the proposed solutions with simulations and real clinical data. This dissertation covers sample size requirements, data collection and predictor selection, model fitting, and the validation of risk models in new data, focusing mainly on diagnostic models. The main case study is the development and validation of models for the pre-operative diagnosis of ovarian cancer, for which the multicenter dataset collected by the International Ovarian Tumor Analysis (IOTA) consortium is used.

The results suggested that mixed effects logistic regression models offer center-specific predictions that have a better predictive performance in new patients than the predictions from standard logistic regression models. Although simulations showed that models were severely overfitted with only five events per variable, mixed effects models did not require more demanding sample size guidelines than standard logistic regression models. A case study on predictors of ovarian malignancy demonstrated that in multicenter data, measurements may vary systematically from one center to another, indicating potential threats to generalizability. These predictors could be detected using the residual intraclass correlation coefficient and may be excluded from risk models. In addition, a case study showed that, if statistical variable selection is used, mixed effects models are required in every step of the selection procedure to prevent incorrect inferences. Finally, case studies on risk models for ovarian cancer demonstrated that the predictive performance of risk models varied considerably between centers. This could be detected using meta-analytic models to analyze discrimination, calibration and clinical utility.

In conclusion, taking into account differences between centers during the planning of prediction research, the development of a model and the validation of risk predictions in new patients offers insight in the heterogeneity and better predictions in local settings. Many methodological challenges remain, among which the inclusion of predictor-by-center interactions, the optimal application of mixed effects models in new centers, and the refinement of techniques to summarize clinical utility in multicenter data. Nonetheless, the findings in this dissertation imply that current clinical prediction research would benefit from adopting mixed and random effects techniques to fully employ the information that is available in multicenter data.

Nomenclature

Symbols

α	intercept
β	regression coefficient
γ	mean of the random effects distribution
Δ	difference
θ_j	true statistic in center j
σ_j	within-center standard error in center j
$\boldsymbol{\Sigma}$	within-center variance-covariance matrix
τ	standard deviation of the random effects distribution
ρ	correlation
$\boldsymbol{\Omega}$	variance-covariance matrix of the random effects
a_j	random intercept for center j
b_j	random slope for center j
C	c-statistic
D	dummy variable
e_{ij}	error term for individual i in center j
f	multiplication factor
H_{ij}	health status for individual i in center j
i	patient indicator
j	center indicator
J	number of centers
n_j	cluster size of center j
n_0	number of non-cases
n_1	number of cases
N	total sample size
K	number of patient-level predictors
L	number of center-level predictors

LP	linear predictor
M	number of dummy variables
p	event rate
p_{ij}	probability for individual i in center j
r_{11}	number of true positives
r_{01}	number of false positives
r_{00}	number of true negatives
r_{10}	number of false negatives
S_{ij}	measurement score for individual i in center j
Se	sensitivity
Sp	specificity
U	uniformly distributed variable
X_{ij}	predictor for individual i in center j
Y_{ij}	dependent variable for individual i in center j
Z_j	center-level predictor for center j

Basic operations

Σ	summation
$\int$	integral
log	natural logarithm

Distributions

bin(n, p)	binomial distribution with n trials and success probability p
$N(\theta, \tau^2)$	normal distribitution with mean θ and variance τ^2
unif(a, b)	uniform distribution on the interval [a,b]

Metrics

kg	kilogram
m^2	square meter
mm	millimeter

Abbreviations

AUC	area under the curve
BMI	body mass index
CI	confidence interval
CS	Cesarean section
DFBETAS	difference in beta values
EDs	Easy Descriptors
EPV	events per variable

FIGO	International Federation of Gynecology and Obstetrics
GEE	generalized estimating equations
ICC	intraclass correlation
IDI	integrated discrimination improvement
IPD	individual patient data
IQR	interquartile range
IOTA	International Ovarian Tumor Analysis group
MAR	missingness at random
MD	doctor of medicine
MFP	multivariate fractional polynomials
MNAR	missingness not at random
NB	net benefit
NRI	net reclassification improvement
RICC	residual intraclass correlation
RMI	risk of malignancy index
RMT	residual myometrial thickness
ROC	receiver operating characteristic
ROMA	Risk of Ovarian Malignancy Algorithm
SA	subjective assessment
SRs	Simple Rules
SRsMal	Simple Rules, assuming malignancy in case of inconclusive results
TRIPOD	transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
VBAC	vaginal birth after a Cesarean section

Table of contents

Acknowledgments iii
Nederlandse samenvatting v
Abstract vii
Nomenclature ix
Table of contents xiii
List of figures xix
List of tables xxv
1 General introduction 1
1.1 Outline of the thesis 3
1.2 Intended audience 4
2 Development and validation of risk models 7
2.1 Methods to develop and validate risk models 7
2.1.1 What is a risk model? 7
2.1.2 Formulating the research question 9
2.1.3 Study design and setup 9
2.1.4 Modeling strategy 12
2.1.5 Fitting the model 15
2.1.6 Validation of model performance 15
2.1.7 Reporting 21

2.1.8 Impact studies 21
2.1.9 Model implementation 22
2.1.10 Conclusion 22
2.2 Predicting successful vaginal birth after a Cesarean section 22
2.2.1 The clinical need for a new VBAC model 23
2.2.2 Subjects and methods 24
2.2.3 Results 26
2.2.4 Discussion 31
3 Statistical methods for multicenter data 35
3.1 Issues and opportunities of clustered data 35
3.2 Methods for clustered data 36
3.2.1 Ignoring the clustered data structure 37
3.2.2 Center-specific models 39
3.2.3 Correcting for clustering 45
3.2.4 Combining within-cluster results 46
3.2.5 Discussion 47
3.3 Multicenter data from the International Ovarian Tumor Analysis group 47
3.3.1 Background 48
3.3.2 The IOTA dataset 49
3.3.3 IOTA models and classification rules 54
3.3.4 The Simple Rules risk scoring system 59
3.4 Conclusion 63
4 Sample size for multicenter studies 65
4.1 Background 65
4.2 Design of the simulation study 66
4.2.1 The source populations 67
4.2.2 Sampling 67
4.2.3 Model building 69

4.2.4 Model evaluation 69
4.3 Results of the simulation study 71
4.3.1 Data clustering and the number of events per variable 71
4.3.2 Variable selection 73
4.3.3 Sample size 74
4.3.4 Random cluster effects correlated with predictors 77
4.4 Empirical example 77
4.5 Discussion 80
4.6 Conclusion 82
5 Predictor selection for multicenter studies 83
5.1 Screening for data clustering in multicenter studies: the residual intraclass correlation 83
5.1.1 Background 83
5.1.2 Methods 85
5.1.3 Results 91
5.1.4 Discussion 97
5.2 Statistical variable selection in clustered data 100
5.2.1 Background 100
5.2.2 Methods 101
5.2.3 Results 102
5.2.4 Discussion 104
5.3 Conclusion 106
6 Performance evaluation in multicenter studies 107
6.1 Heterogeneity in predictive performance 108
6.2 Performance measures for multicenter validation 109
6.2.1 Sensitivity and specificity 109
6.2.2 The c-statistic 111
6.2.3 Calibration 112
6.2.4 Net benefit 113

6.2.5 Explaining heterogeneity 114

6.2.6 Leave-one-center-out cross-validation..... 114

6.3 Some examples 116

6.3.1 The validation of IOTA strategies on phase III data: a meta-analysis of discrimination and calibration 116

6.3.2 The validation of the IOTA Simple Rules risks scoring system on phase III data: a meta-analysis of discrimination and a graphical assessment of calibration in specialized oncology centers and other centers 124

6.3.3 The validation of IOTA models and RMI in the hands of users with varied training on phase IVb data: a meta-regression of test accuracy 129

6.3.4 The validation of the clinical utility of models on phase III data: net benefit in specialized oncology centers and other centers... 132

6.4 A meta-analysis of net benefit..... 137

6.4.1 Various fixed and random effects weights 137

6.4.2 Random effects meta-analysis of the net benefit: an example 141

6.4.3 Future research: a Bayesian approach..... 145

6.5 Conclusion 148

7 Does ignoring clustering in multicenter data influence the predictive performance of risk models? A simulation study 149

7.1 Introduction..... 149

7.2 A framework of performance evaluation of risk models in clustered data 150

7.3 Calibration slopes for marginal and center-specific logistic regression models 152

7.4 Simulation study..... 153

7.4.1 Design 153

7.4.2 Results..... 155

7.5 Empirical example..... 163

7.6 Discussion..... 165

7.7 Conclusion 168
8 General discussion 169
8.1 Implications and recommendations 171
8.2 Future research 173
Appendices 177
A1 Multiple imputation in the IOTA dataset 178
A2 Technical overview of the design of the EPV simulation study 179
A3 Examples of R code for the EPV simulation study 181
A3.1 Generation of source populations 181
A3.2 Sampling from the source population and model building within samples 182
A4 Additional results from the EPV simulation study 187
A4.1 Bias in the estimated regression coefficients 187
A4.2 Standard (population-level or "overall") c-statistics and calibration slopes 188
A4.3 Bias in the estimated random intercept variance 189
A5 SAS macro to estimate the residual intraclass correlation 190
A6 Difference in net sensitivity and net amount of avoided false positives per 100 patients 202
A7 Center-specific case-mix and population models to generate a heterogeneous multicenter dataset 204
A8 Observed differences in center-specific decision curves in simulated data 207
A9 Results for meta-analyses of center-specific net benefit using various weights 208
A10 Additional results of the random effects meta-analysis of NB in IOTA data 209
A11 Formulas for the c-statistic and logistic calibration in a comprehensive framework 210
A12 Calibration with a biased estimate of the between-center variance 211

A13 Correspondence between predictions from the standard logistic regression model and marginalized predictions from the mixed effects logistic regression model 212

A14 R code for the simulation study of the impact of ignoring clustering on predictive performance 213

A15 Additional results for the simulation study of the impact of ignoring clustering on predictive performance 222

A15.1 Detailed calibration results of simulations with EPV=100 and ICC=20% 222

A15.2 Relation between the estimated random intercept variance and the calibration slope 223

A15.3 Calibration intercepts and c-statistics obtained with development samples with EPV 100 in a population with ICC=5% 224

A15.4 Calibration intercepts and c-statistics obtained with development samples with EPV 5 in a population with ICC=20% 226

References 229

Curriculum vitae 249

List of publications 251

Papers in international journals 251

Letters and replies in international journals 252

Conference abstracts in international journals 252

Unpublished conference contributions 253

Invited talks 255

List of figures

Figure 1. Thesis outline and overview of the relations between chapters. 4

Figure 2. From conception to implementation: an overview of steps to develop and validate clinical risk models. 8

Figure 3. ROC curve for the ADNEX model to distinguish between malignant and benign lesions. 16

Figure 4. Calibration plot of the ADNEX model at external validation. 18

Figure 5. Net benefit of the ADNEX model at external validation. 19

Figure 6. Flow chart summarizing the study sample, showing the mode of delivery according to Cesarean section scar visibility. 26

Figure 7. Calibration plot for the model to predict successful vaginal delivery after a Cesarean section (CS), based on patient age, previous history of VBAC, residual myometrial thickness and change in RMT between first and second trimester, in 121 women with visible CS scar and one previous CS. 30

Figure 8. Decision curves for three strategies in the management of delivery in women with visible Cesarean section (CS) scars and one previous CS: using the prediction model for success of vaginal birth after CS (VBAC), giving all women a trial of VBAC and giving no woman a trial of VBAC. 31

Figure 9. Objectives of the different IOTA phases. 50

Figure 10. The prevalence of malignant tumors per center, all phases combined. 54

Figure 11. IOTA Simple Rules. 56

Figure 12. IOTA Easy Descriptors. 57

Figure 13. The relative bias in estimated regression coefficients and the predictive performance in relation to the number of events per variable and the amount of clustering. 72

Figure 14. The relative bias in estimated regression coefficients and the predictive performance in relation to the number of events per variable. 75

Figure 15. The relative bias in estimated regression coefficients and the predictive performance in relation to the number of events per variable and sample characteristics (ICC=20%). 76

Figure 16. The relative bias in the estimated regression coefficients and the predictive performance in relation to the number of events per variable when predictors are dependent versus independent of the random intercept (ICC=20%). 78

Figure 17. The residual intraclass correlation versus the proportion of explained variance.. 94

Figure 18. Average measurement or prevalence per physician versus the random intercept. 96

Figure 19. The marginal effect of age estimated by the standard logistic regression model, the marginalized effect of the random intercept model and effect in the average center estimated by the random intercept model. 103

Figure 20. Variable selection frequencies in 100 bootstrap samples, using backward standard logistic regression and backward random intercept logistic regression. 104

Figure 21. The c-statistic for LR2 and the Risk of Malignancy Index (RMI) per contributing center and for all centers combined using a meta-analysis approach and pooled data. 120

Figure 22. The sensitivity and specificity for LR2, the Risk of Malignancy Index and a two-stage strategy using the Simple Rules as a first stage test followed by subjective assessment by an expert if the Simple Rules are inconclusive. 121

Figure 23. Center-specific logistic calibration curves for the Risk of Malignancy Index, LR2 and the Simple Rules. 122

Figure 24. The c-statistic for the Simple Rules risk scoring system per contributing center, for all oncology centers, all other centers and all centers combined using a meta-analysis. 126

Figure 25. Calibration curves for all data, data from oncology centers and data from other centers. 127

Figure 26. Decision curves of the Risk of Malignancy Index, the LR2, the IOTA Simple Rules Risk scoring system and the ADNEX model. 134

Figure 27. Decision curves of the Risk of Malignancy Index, the LR2, the IOTA Simple Rules Risk scoring system, and the ADNEX model in oncology centers and other centers. 135

Figure 28. Decision curves using various weights for a meta-analysis to summarize center-specific net benefit..... 139

Figure 29. Observed center-specific decision curves, the decision curve obtained by pooling all data and the decision curve obtained by random effects net benefit..... 143

Figure 30. Decision curves obtained by random effects net benefit of oncology centers and other centers. 144

Figure 31. A comprehensive framework of options for model validation, subject to the type of risk model that is being evaluated (standard or mixed effects logistic regression) and the available validation dataset (one center or multicenter). 151

Figure 32. Center-level and population-level calibration slopes of the standard logistic regression model, the conditional linear predictor of the random intercept model and the linear predictor of the random intercept model assuming an average random intercept, by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.822 (ICC=20%). 156

Figure 33. Center-level and population-level calibration intercepts of the standard logistic regression model, the conditional linear predictor of the random intercept model and the linear predictor of the random intercept model assuming an average random intercept, by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.822 (ICC=20%). 157

Figure 34. Center-level and population-level calibration slopes of the standard logistic regression model, the conditional linear predictor of the random intercept model and the linear predictor assuming an average random intercept, by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.157 (ICC=5%). 158

Figure 35. Center-level and population-level c-statistics of the standard logistic regression model, the conditional linear predictor of the random intercept model and the linear predictor assuming an average random intercept, by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.822 (ICC=20%). 160

Figure 36. Center-level and population-level calibration slopes of the standard logistic regression model, the conditional linear predictor of the random intercept model and the linear predictor of the random intercept model assuming an average random intercept, by estimated random intercept variance in samples with 5 events per variable and true random effects variance=0.822 (ICC=20%). 162

Figure A 1. The relative standard c-statistic and the relative standard calibration slope in relation to the number of events per variable, the amount of clustering, backward variable selection and sample size characteristics. 188

Figure A 2. The relative bias in the estimated random intercept variance (ICC=20%) in relation to the number of events per variable and sample characteristics. 189

Figure A 3. Observed differences in center-specific decision curves in a simulated heterogeneous and homogeneous population. 207

Figure A 4. Decision curves for LR2 and a refitted LR2 model obtained by random effects analysis of net benefit in oncology centers. 209

Figure A 5. Decision curves for LR2, a refitted LR2 model, a probability model of RMI, The Simple Rules classifying inconclusive cases as malignant, expert's subjective assessment, treat all and treat none, obtained by random effects analysis of net benefit. 209

Figure A 6. Marginal predictions from a standard logistic regression model versus marginalized predictions from a mixed effects logistic regression model. 212

Figure A 7. Center-level calibration slopes of the standard logistic regression model and the random intercept model assuming average random intercepts, by estimated random intercept variance. 223

Figure A 8. Center-level and population-level calibration intercepts of the standard logistic regression model, the conditional linear predictor of the random intercept model and the linear predictor assuming an average random intercept, by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.157 (ICC=5%). 224

Figure A 9. Center-level and population-level c-statistics of the standard logistic regression model, the conditional linear predictor of the random intercept model, and the linear predictor assuming an average random intercept, by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.157 (ICC=5%). 225

Figure A 10. Center-level and population-level calibration intercepts of the standard logistic regression model, the conditional linear predictor of the random intercept model from the random intercept model and the linear predictor assuming an average random intercept, by estimated random intercept variance in samples with 5 events per variable and true random effects variance=0.822 (ICC=20%). 226

Figure A 11. Population-level and center-level c-statistics of the standard logistic regression model, the conditional linear predictor of the random intercept model and the linear predictor assuming an average random intercept, by estimated random intercept variance in samples with 5 events per variable and true random effects variance=0.822 (ICC=20%). 227

List of tables

Table 1. Overview of potential pitfalls when developing or validating risk models. 10

Table 2. Indications for emergency Cesarean sections in 47 women with visible CS scars and one previous CS, who attempted vaginal birth after a Cesarean section 27

Table 3. Descriptive statistics of patient and Cesarean section (CS) scar characteristics according to mode of delivery in 121 women with visible CS scar and one previous CS 28

Table 4. Estimated effects of patient and scar characteristics in a fitted logistic regression model to the predicted success of a trial of vaginal birth after a Cesarean section. 29

Table 5. Sensitivity and specificity for success of a trial of vaginal birth after a Cesarean section, using the prediction model at specified cut-off values for predicted probability 29

Table 6. Methods to analyze multicenter datasets 37

Table 7. Patient, tumor and center characteristics for different IOTA phases 51

Table 8. Number of patients contributed by each center to the IOTA database in phases I, Ib, II, III and IVb. 52

Table 9. Estimated regression coefficients, random intercept variance, residual intraclass correlation and proportion of explained variance for regression models based on the Simple Rules ultrasound features fitted on phase Ib and II data (n=2445). 60

Table 10. Updated model coefficients using all data (n=4848)..... 61

Table 11. Simple Rules feature combinations and the associated risk of malignancy (in %) 62

Table 12. Simulation conditions..... 68

Table 13. The selection frequency of predictors 74

Table 14. Estimated coefficients of the fitted models for ovarian tumor diagnosis in 100 bootstrap samples, based on EPV=5 versus EPV=51, without (model 1 and model 2) and with (model 3 and model 4) backward variable selection (α=0.10)..... 79

Table 15. The performance of the fitted models for ovarian tumor diagnosis, based on EPV=5 versus EPV=51 in 100 bootstrap samples, without (model 1 and model 2) and with (model 3 and model 4) backward variable selection (α=0.10) 80

Table 16. Variance partitioning of ultrasound measurements and patient characteristics 93

Table 17. Effect of ultrasound machine quality on the residual intraclass correlation. 95

Table 18. Estimated regression coefficients for the full and reduced standard and random intercept logistic regression models..... 105

Table 19. Predictive performance of subjective assessment by an expert, the IOTA strategies and the Risk of Malignancy Index, using a meta-analysis of center-specific data. 118

Table 20. Sensitivity and specificity for the Simple Rules risk scoring system in oncology centers and other centers. 128

Table 21. Sensitivity and specificity of LR2, RMI, the Simple Rules and subjective assessment by users with varied levels of experience and training 130

Table 22. A comparison of the summary estimates and prediction intervals of a random effects meta-analysis based on net benefit and a log(1-net benefit) transformation. 140

Table 23. Summary sensitivity, specificity, net benefit and the probability that ear thermometry will be harmful in a new study, calculated using a Bayesian approach. 147

Table 24. Estimated regression coefficients from a mixed effects logistic regression model and a standard logistic regression model to predict the risk of ovarian mass malignancy. 163

Table 25. Population-level and center-level discrimination and calibration statistics for the marginal predicted risks from the standard logistic regression model and for the marginalized predicted risks from the mixed effects model, the predictions assuming an average random intercept, and the conditional predicted risks..... 164

Table 26. Schematic overview of the effect of the type of prediction on the conditional and standard performance measures, in the absence of overfitting and assuming a representative development dataset. 166

Table A 1. The relative bias (%) in the estimated regression coefficients. 187

Table A 2. Increase in net sensitivity (%) of the IOTA Simple Rules risk scoring system compared to the Risk of Malignancy Index, the LR2 model by IOTA, and the ADNEX model by IOTA for all patients and for oncology and other centers. 202

Table A 3. Net amount of additional false positives avoided per 100 patients by the IOTA Simple Rules risk scoring system compared to the Risk of Malignancy Index, the LR2 model by IOTA, and the ADNEX model by IOTA for all patients, oncology centers and other centers. 203

Table A 4. The center-specific distribution of the predictor X and the true association between the predictor X and the outcome Y, the number of patients sampled for the validation dataset from center j, the outcome prevalence in center j, the observed center-specific c-statistic, the observed calibration intercept and the observed calibration slope. 204

Table A 5. Results for meta-analyses of center-specific net benefit using various weights..... 208

Table A 6. Formulas for the c-statistic and logistic calibration in a comprehensive framework of the standard and within-center validation of marginal predictions, conditional predictions and predictions assuming an average random intercept. 210

Table A 7. Calibration slopes and intercepts at the center and population level for marginal predictions from the standard logistic regression model and predictions assuming an average random center intercept and conditional predictions from the random intercept logistic regression model. 222

1 General introduction

David Hand, the former president of the Royal Statistical Society, once said: "In general, when building statistical models, we must not forget that the aim is to understand something about the real world. Or predict, choose an action, make a decision, summarize evidence, and so on, but always about the real world, not an abstract mathematical world."[1] It is precisely this philosophy that statisticians adhere to when they engage in prediction modeling. When building a risk model with an intended use in clinical practice, the statistician desires to assist the clinician in the complex task of making a diagnosis, a prognosis or a medical decision. Complex indeed, as the model and the clinician alike need to integrate various pieces of information to enable an informed decision, for example regarding the treatment the patient requires. The nature of the considered information may be very diverse, from the patient's symptoms, medical history and demographics to blood work and test results or even genetic information and family history. Mathematical models use formulas to combine and weigh all information, and render explicit the underlying synthesis of evidence that experienced clinicians perform more or less subconsciously. They are not built to replace clinical judgement but should be regarded as an additional tool to provide evidence-based care, the ultimate goal being an improvement of patient outcomes.

Statistical models are developed from data. By collecting information on many patients and using mathematical techniques to distinguish between meaningful relations among variables and random patterns, researchers aim to build models that are useful in future patients. To ensure that the model may be used in diverse clinical settings, they increasingly collaborate to collect data from different hospitals in so-called multicenter studies. As much as a statistical model should be useful in the real world, it should also be chosen to

match the reality it models, since all models make assumptions regarding the world they describe. This is exactly what is ignored in much of the current prediction research. The standard mathematical techniques to analyze the patient data assume independence of the observations, and this cannot be taken for granted when data comes from multiple hospitals, as patients from the same hospital may be more similar than patients from different hospitals. Many factors such as regional differences in populations and local referral patterns contribute to noteworthy differences in patient case-mix across hospitals. When the assumptions underlying the mathematical techniques used to analyze data are wrong, we cannot simply expect the results to be right. All things considered, one must ask whether models relying on faulty presumptions yield the best predictions possible and, in the end, the best results for patients.

In my PhD research, I set out to investigate the consequences of wrongly assuming independence of observations and studied alternative models and techniques that acknowledge the multicenter nature of the data, focusing mainly on diagnostic models. I considered not only the model building itself but the entire process that precedes the introduction of a risk model to clinical practice, including the determination of the required sample size when planning a multicenter study, the selection of relevant predictors for inclusion in the risk model, and the evaluation of the predictive performance of the risk model in new patients. Throughout my research I used mixed and random effect models. In contrast to the most used statistical models, they do not assume that observations are independent. They explicitly model the dependence structure and acknowledge that the centers we study are part of a broader population of centers, to which we may want to generalize the results. To illustrate the methods that I studied, I made use of simulated data and real data. Imitating real-world data-generating mechanisms, simulation studies allow to study methods under various circumstances. Using real clinical data provides useful illustrations of how the methods may be implemented in practice. The case study that I used all through my research is the diagnosis of ovarian cancer. Making use of the multicenter dataset collected by the International Ovarian Tumor Analysis (IOTA) consortium, I developed and tested risk models to distinguish pre-operatively between benign and malignant tumors based on patient information and sonographic tumor characteristics.

1.1 Outline of the thesis

This dissertation summarizes my PhD research. Chapter 2 reviews the methodological literature concerning the development and validation (or testing) of risk models. It distills best practices that are common to high-quality prediction research and uncovers some common pitfalls. The recommended methodology is illustrated by a case study on the development and validation of a model to predict the probability that a woman can have a vaginal delivery after a previous Cesarean section. Chapter 3 provides a synopsis of methods to deal with multicenter data, among which the mixed effects model. Next, it discusses the multicenter IOTA dataset, introduces some of the previously developed models for ovarian tumor diagnosis, and shows how the multicenter nature of the data was accounted for during the development of the Simple Rules risk scoring system. Together, these two chapters provide the necessary basis for the following chapters (see Figure 1).

Chapters 4, 5 and 6 discuss methods to acknowledge the multicenter nature of data during different phases of prediction research. Chapter 4 presents a simulation study on the required sample size, and specifically, the required number of events per variable to build a generalizable risk model. Chapter 5 is concerned with variable selection. It proposes a novel metric (the residual intraclass correlation) to quantify to which extent measurements for predictors differ from one center to another. If large differences are present, this may be an argument to omit the predictor from the risk model. Next, the chapter compares an automated stepwise predictor selection procedure that accounts for the multicenter nature of data to a procedure that does not. Chapter 6 covers methods to evaluate the predictive performance of risk models in multicenter data. It describes how certain techniques developed for meta-analysis can be used to evaluate the performance of models in multiple centers and offers examples of the validation of models for ovarian tumor diagnosis. Besides that, the chapter also explores novel methods to evaluate the clinical utility of a risk model in multicenter data, which transcends common statistical methods to evaluate predictive performance in terms of the benefit of the model for clinical decision-making.

Chapter 7 synthesizes many elements introduced in chapters 4, 5 and 6 (see Figure 1). It presents a generalized framework of performance evaluation and shows the results of a simulation study to answer one important remaining question: does acknowledging the multicenter nature of the data improve the predictive performance of a risk model? Chapter 8 concludes: it summarizes the main results of my PhD research, discusses their implications and proposes some directions for future research.

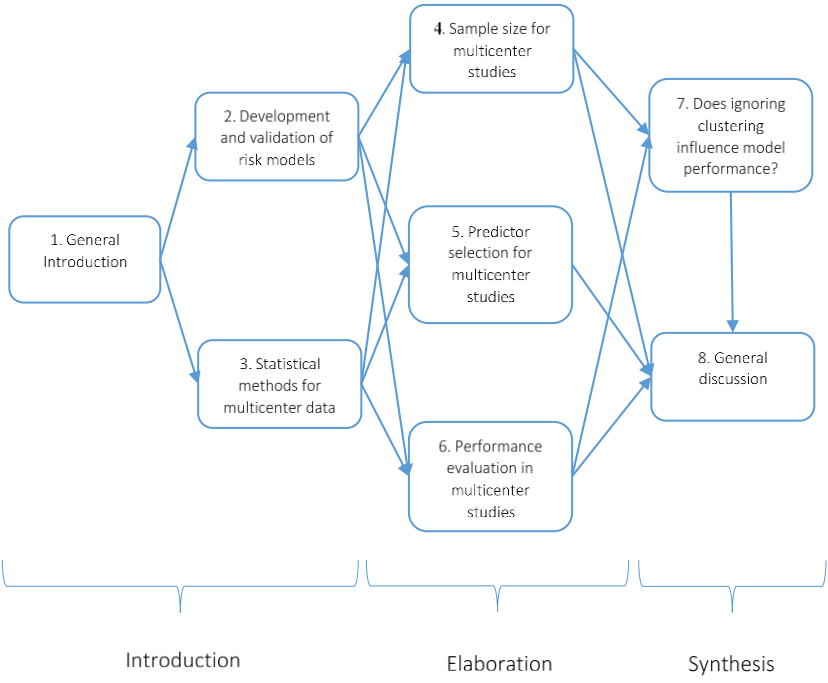


Figure 1. Thesis outline and overview of the relations between chapters.

1.2 Intended audience

This dissertation may be of interest to methodologists and clinicians alike.

Some of the sections in this dissertation describe existing methods that are relevant in the context of the development or validation of risk models in multicenter data (sections 2.1, 3.2 and 6.2). The review of methods for prediction research presented in section 2.1 was previously published in *BJOG*. The other sections were written to inform the reader on the state-of-the-art methodology. Other parts of this dissertation present novel methodological research (Chapter 4, Chapter 5, section 6.4 and Chapter 7). Chapter 4, on the role of the number of events per variable, is published in the *Journal of Clinical Epidemiology*, and section 5.1, on the residual intraclass correlation, is published in *BMC Medical Research Methodology*. Chapter 7, on the effect of acknowledging the multicenter nature of data on the predictive performance of models, will be published in *Statistical Methods in Medical*

Research. The research presented in sections 5.2, on statistical predictor selection, and 6.4, on the meta-analysis of clinical utility, has previously been presented at international conferences (of the International Society of Clinical Biostatistics and the Society for Medical Decision Making). The research on the meta-analysis of clinical utility was awarded the 2014 Lee Lusted Student Prize in Quantitative Methods and Theoretical Developments at the annual meeting of the Society for Medical Decision Making.

The most important applications of the discussed methodology on clinical data are described in sections 2.2, 3.3, 5.1.3, 6.3 and 6.4.2. Besides illustrations on real-world data, these sections offer meaningful clinical insights as well. Section 2.2, on the development of a prediction model for vaginal birth after a Cesarean section, was published in *Ultrasound in Obstetrics and Gynecology*. I was the main statistician involved in this research. Part of the material covered in sections 3.3.4 and 6.3.2, on the development and validation of the Simple Rules risk scoring system, was published in the *American Journal of Obstetrics & Gynecology*. I participated in this research as a statistician. The results of the validation of diagnostic models for ovarian cancer on IOTA phase III data presented in section 6.3.1 were published in a paper in the *British Journal of Cancer*, which I co-authored as one of the statisticians involved. The results presented in section 6.3.3 on the validation of IOTA models in the hands of users with varied training were also published in a paper in the *British Journal of Cancer*, for which I was the main statistician. The research presented in section 6.3.4, on the clinical utility of models for ovarian tumor diagnosis, was previously presented at the conference of the Society for Medical Decision Making and at the symposium on Methods for Evaluating Medical Tests and Biomarkers (MEMTAB), and is the subject of a paper in preparation of which I am the main author. In addition, sections 4.4, 5.2, 6.4.3 and 7.5 alo make use of clinical datasets in empirical examples but the main purpose of these sections is to illustrate the methods, and their conclusions may be less relevant for clinical practice.

2 Development and validation of risk models

This chapter is an introduction to risk models for clinical predictions. Section 2.1 offers an overview of the methodology for the development and validation of risk models. Section 2.2 presents an example on the development of a model to predict the probability of vaginal delivery after a Cesarean section for a previous pregnancy.

This chapter is based on the following research papers:

Wynants L, Collins GS and Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG*. 2016: e-pub ahead of print.

Naji O, Wynants L, Smith A, et al. Predicting successful vaginal birth after Cesarean section using a model based on Cesarean scar features examined by transvaginal sonography. *Ultrasound Obstet Gynecol*. 2013; 41: 672-8.

2.1 Methods to develop and validate risk models

In recent years we have seen an increasing number of papers about risk models.[2] To maximize their potential, it is imperative that risk models are developed carefully and validated rigorously. We address the steps between the formulation of a research question and the use of the model in clinical practice one by one (Figure 2) and highlight common pitfalls (Table 1).

2.1.1 What is a risk model?

Risk models predict the risk that a condition is present (diagnostic) or will develop in the future (prognostic).[3] Risk models aim to aid clinicians in making treatment decisions based on patient-specific measurements and to discuss risks with patients. In that sense risk models have the potential to

facilitate personalized medicine and enhance shared decision-making.[4] Risk prediction is usually performed using multivariable models that aim to provide reliable predictions in new patients.[5, 6] This is an important distinction with etiological models, which are used to study causality, and explanatory models, which are used to provide a useful explanation of a phenomenon with the available, albeit incomplete, knowledge on the subject.

Examples of diagnostic models include the models of the IOTA (international ovarian tumor analysis) group. The LR2 model is a logistic regression model that calculates the risk that a tumor is malignant, based on ultrasound characteristics of the tumor and clinical characteristics of the patient.[7] The ADNEX (Assessment of Different NEoplasias in the adnexa) model is similar, but discriminates between five different subtypes of adnexal tumors and gives risk estimates for each.[8] The results of these models can be used to triage patients with suspicious adnexal masses, and refer them to a specialized gynecological oncologist if the risk of malignancy is high. An example of a prognostic model is the well-known Gail model. This model uses a woman's medical and reproductive history and the history of breast cancer among first-degree relatives to estimate her risk of developing invasive breast cancer in a specific period of time, for example in the next five years.[9]

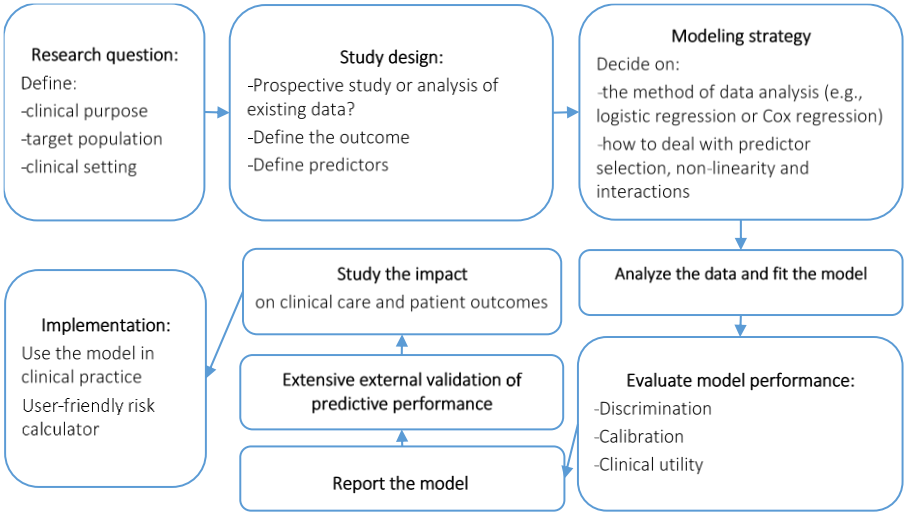


Figure 2. From conception to implementation: an overview of steps to develop and validate clinical risk models.

2.1.2 Formulating the research question

Before starting the development of a risk model, it is pivotal to define the clinical purpose it should serve,[10] as well as the target population and the clinical setting. A mismatch with clinical needs is not uncommon and makes the model useless (Table 1, pitfall 1). Models for the same purpose often already exist. The characteristics of relevant models (e.g., the clinical setting and included predictors), and the extent to which these models have been validated, should be checked. Search strategies for finding relevant existing risk models are available.[3, 11] Systematic reviews indicate that models are often developed in vain because there was no clear clinical need, models for the same purpose already existed, or existing models had not yet been the subject of validation studies. [2, 12-15]

2.1.3 Study design and setup

To enhance the transparency and reproducibility of prediction research, the study setup, along with the details of the planned data analysis, should be discussed beforehand and written down in a study protocol.[16] Recently, researchers have called for the publication and registration of protocols for prediction research.[17]

The preferred study design is a cross-sectional cohort study for diagnostic risk models and a longitudinal cohort study for prognostic risk models. In the former design, the sample consists of patients suspected of having a particular disease, for example patients with certain symptoms. In the latter design, a cohort is defined by the presence of particular characteristics, for example, having a certain disease, and followed over time to register the occurrence of a certain event. If the outcome to be predicted is rare, it may be useful to identify sufficient cases and recruit a random sample of non-cases, using a case-cohort or nested case-control design.[18] The risk model then needs to be adjusted for the sampling frequency of non-cases to ensure reliable predictions.[19]

In general, the patient sample should be representative of the intended target population for the model. In order to prevent bias, a consecutive sample of patients is preferred. To enhance model generalizability and transportability, multicenter data collection is recommended. The ideal study setup is one where data is collected with the primary aim to develop the risk model. This is often referred to as a prospective study,[3, 20] as opposed to a retrospective study where data already existed for another purpose. However, one should be careful using these terms as they have been defined in conflicting ways.[21]

Table 1. Overview of potential pitfalls when developing or validating risk models.

Pitfall	Description	Solution(s)
1. The model does not match the clinical needs	The model is not clinically relevant or does not fit in the clinical workflow, the selected variables lack credibility, or the wrong population is targeted.	* Do not build the model * Discuss with clinicians
2. Bias due to a complete case analysis	Patients with missing data are omitted. This may lead to an unrepresentative sample of complete cases that results in incorrect risk predictions.	* Use imputation methods * Omit variables with excessive missingness (e.g., >50%)
3. The number of events is small relative to the number of examined parameters	There are not enough events to develop a model that is likely to be robust and validate. This leads to overfitting.	* Increase the sample size (e.g., collaborate with other centers) * Reduce the complexity of the model * Use subject matter knowledge to select predictors * Do not build the model
4. A too strong focus on statistical selection and significance	Predictors included for the final model are merely selected using statistical procedures (e.g., stepwise selection) with a strong focus on significance levels to include or exclude a variable.	* Use subject matter knowledge, even if the number of events per variable (EPV) is acceptable * Rely on clinical relevance and the sign of effects to aid selection
5. Too much complexity	Too many complex terms are considered, for example with respect to interaction terms or nonlinear terms. This leads to an increased risk of overfitting and to an unattractive and complicated model.	* Limit complex terms to what is important, reasonable and plausible * Define a strategy to deal with model complexity upfront
6. An inefficient internal validation strategy	Data are randomly divided into a small development dataset and a small internal validation dataset. This jeopardizes EPV and the reliability of results.	* Use bootstrapping or cross-validation instead, certainly when EPV is small
7. Inappropriate reporting of the model	The scientific report is incomplete (e.g., an incomplete account of the population, setting, data collection, reference standard, or modeling strategy, not reporting the model coefficients).	* Use the TRIPOD guidelines when preparing the scientific report [3]
8. External validation is ignored	New models are built, instead of validating existing models.	* Set up external validation studies * Directly compare different models for the same purpose

Examples of existing data sources that may be used for prediction research include registry studies, randomized clinical trials, and hospitals' medical health records. A limitation of using existing data is that they were collected for a different purpose and that important risk factors may not have been collected properly or may be lacking altogether, thereby reducing the performance and credibility of the resulting model. Data from randomized trials may lack generalizability due to strong inclusion or exclusion criteria, or volunteer bias. Furthermore, in the presence of an effective treatment, thought needs to be given to how treatment should be handled in the analysis.[22]

It is important to carefully identify the predictor variables for inclusion in the study, such as established risk factors, variables that are easy to collect,[23] and promising variables of which the predictive value has yet to be established. The potential predictors should be available when the risk model will be used during patient care. For example, when predicting the risk of a birth defect at the nuchal scan, it is not useful to consider birth weight as a predictor.

For diagnostic models, the outcome should be measured by the best available and commonly accepted method, the "reference standard", for example the histological examination of tissue to determine whether a tumor is benign or malignant. The outcome can typically be determined after the candidate predictors have been assessed. The time interval between data collection on predictors and the outcome should be specified and kept short. Preferably, patients receive no treatment during this interval. For prognostic models the outcome should be a clinically relevant event, occurring in a specified period in the future. In both diagnostic and prognostic settings, blinded outcome assessment (i.e., without knowledge of the predictors) avoids unwanted review bias.

The definitions of variables and outcomes should be standardized in advance. Measurement error should be avoided. Unfortunately, measurement error is common for many predictors and influences model performance.[24] The estimated regression coefficients are often diluted (biased towards 0), although the opposite may also occur.[25, 26] It may be useful to set up an interobserver reliability study, or at least discuss the likely reliability of potential predictors, particularly when predictors contain some level of subjectivity.[27] In multicenter datasets, differences in measurement equipment, local procedures, or center populations can give rise to systematic differences in measurements.

However carefully the study is designed, data are rarely complete. A common mistake is to exclude patients with missing data, and perform the analysis only

on complete cases (Table 1, pitfall 2). This approach is inefficient as it reduces the sample size, and potentially introduces a bias, because there is usually a reason for missing values.[28] Hence, the resulting model may generalize poorly to new patients. A superior approach is to replace ("impute") the missing data with sensible values. Preferably several plausible values are imputed ("multiple imputation") to acknowledge that imputed values are themselves estimates and thus uncertain.[29, 30] A popular method is (multiple) imputation by "chained equations", also called fully conditional specification.[31] Using this approach, missing information can be estimated with related variables for which information is available, including the outcome and variables that are related to the missingness.[28-30, 32, 33] Most imputation approaches assume that values are "missing at random" (MAR).[34] This means that missing values do not occur in a completely random fashion, but are random conditional on the available data. For example, if values are more often missing in older patients, then MAR holds if patient age is available. Sometimes, missingness may be related to the value of the missing variable itself, or to characteristics that are not available in the dataset. This is called missingness not at random (MNAR). An example is depression in a questionnaire: the most severely depressed respondents may not have been able to complete the questionnaire due to their condition.[35] MNAR cannot be detected by any tests, but its potential impact can be assessed by sensitivity analyses.

2.1.4 Modeling strategy

Regression methods such as logistic regression (for dichotomous outcomes) and Cox regression (for time-to-event outcomes) are the most frequently used algorithms to fit risk models. The machine learning literature offers an alternative class of models that focus on computational flexibility.[36] Examples include neural networks, support vector machines, and random forests.[37] However, these methods struggle with reliable probabilistic estimation[38, 39] and interpretability. Moreover, at least for clinical applications, machine learning methods generally do not yield better performance than regression methods.[39-43]

An important consideration for a modeling strategy is the complexity of the model, relative to the available sample size, in order to avoid overfitting.[5] Capturing random peculiarities in a sample (overfitting) should be avoided,[44] and we should focus on identifying the "true" underlying associations with the outcome. Overly complex models are often overfitted: they perform well on the model development data, but poorly on new data (Table 1, pitfall 3).[44] It is recommended to limit the ratio of the number of events to the number of

variables examined, referred to as events per variable (EPV). Although EPV is a common term, it is more appropriate to consider the total number of estimated parameters (i.e., all regression coefficients, including all those considered prior to statistical variable selection) rather than just the number of variables in the final model.[5] For binary outcomes, the number of events is the number of individuals in the smallest outcome category.

For small samples, shrinkage methods may be considered, whereby model coefficients are penalized (i.e., shrunken towards zero) to avoid overestimation of effects.[45] A common penalization method is ridge regression.[46] Even though standard maximum likelihood yields unbiased parameter estimates provided the sample size is sufficiently large,[47] overfitting may occur in relatively large datasets. Each parameter is estimated with uncertainty, and multiplying multiple parameter estimates with their associated predictor values may yield predicted risks that are too extreme at the extremes of the linear predictor (i.e., high risks are overestimated and low risks are underestimated). Using shrinkage techniques introduces a bias in the estimated regression coefficients, but yields a gain in precision of the predictions that offsets the introduced bias. This is an example of Stein's paradox.[48, 49]

The required sample size thus depends on the complexity of the model, although shrinkage techniques may be used to reduce overfitting. A minimum of 10 EPV is frequently recommended,[50, 51] but the following is more realistic: at least 10 EPV when the model is specified a priori, although additional shrinkage of model coefficients, using for example ridge regression, may be required; at least 20 EPV to alleviate the need for shrinkage; and at least 50 EPV when statistical variable selection is used.[6, 45, 52, 53] Research on EPV guidelines is ongoing.[54]

It is common practice to use statistical variable selection to reduce the number of variables. Backwards elimination is often preferred because this approach starts with a full model (i.e., includes all variables) and eliminates non-significant variables one by one.[5] Although convenient, such methods have important limitations, especially in small datasets.[5, 6, 44, 45, 52] Due to repeated statistical testing, such selection methods lead to overestimated coefficients (resulting in overfitting) and overoptimistic p-values (Table 1, pitfall 4).[55] The same is true when selecting variables based on their univariate association with the outcome, which should be avoided.[5] Statistical significance is often not the best criterion for inclusion or exclusion of predictors. Instead of using purely statistical selection, a preferable strategy is to rely on subject matter knowledge to select predictors before any multivariable analysis.[5, 6] Based on

experience and scientific literature, domain experts can often judge the likely importance of predictors. In addition, predictors can be judged in terms of subjectivity,[27, 56] financial cost, and invasiveness. For example, the variable age is easy and cheap, whereas measurements based on magnetic resonance imaging can be difficult and expensive. Nevertheless, statistical selection remains of interest when subject matter knowledge falls short.[5, 57] Subject matter knowledge may also prove useful when scrutinizing a fitted model. For example, an effect in opposite direction to what is expected is suspicious. The exclusion of such predictors can be beneficial for the robustness and transportability of the model.[45]

When dealing with categorical variables, categories with little data may be combined to reduce the number of parameters. Categorizing continuous variables should be avoided, because this entails a substantial loss of information.[58]

To obtain an optimal model fit, it is advisable to assess whether the variable has a linear effect or might need a transformation. For example, biomarker effects are often logarithmic: for low biomarker values, small differences have a strong influence on the risk of disease, whereas for high values small differences do not matter much. Popular methods to model nonlinearity are spline functions (e.g., restricted cubic splines) and multivariable fractional polynomials.[5, 57] The latter approach combines the assessment of nonlinearity with backward variable selection. Investigations of nonlinearity should be kept within reasonable limits relative to the number of available events to avoid overly complex models (Table 1, pitfall 5). To avoid overfitting when the number of events is small, one can either assume linear effects or investigate nonlinearity for the most important predictors only.

Including interactions may also improve the fit of the model. An interaction occurs when the coefficient of a predictor depends on the value of another predictor. To avoid overly complex models (especially in small samples), it is recommended to specify in advance which interaction terms are known or potentially relevant, instead of testing all possible interaction terms.[5, 6] The number of possible interaction terms grows exponentially with the number of predictors, and the power to detect interactions is low.[59] As a result, many statistically significant interactions are hard to replicate. Moreover, it is unlikely that interaction terms will dramatically improve predictions. An interesting exception is the interaction between fetal heart activity and gestational sac size when predicting the chance of pregnancy viability beyond the first trimester: larger gestational sac sizes increase the chance of viability

when fetal heart activity was observed but decrease the chance of viability when no such activity was seen.[60]

It is recommended to define the modeling strategy before performing the analysis and to describe it in the study protocol, including how to deal with interactions and nonlinearities.[5, 6]

2.1.5 Fitting the model

Once the modeling strategy is defined and the data is collected, the model can be fitted to the data using statistical software. However, prior to fitting the model to the data, it is important to scrutinize the data with descriptive statistics and correct implausible values or fill in missing values, if possible. Strong associations between predictor variables (multicollinearity) can be detected with simple correlations or a more thorough assessment with variance inflation factors when fitting the model. When strong collinearity is present, coefficient estimates may become unstable. This can be avoided by excluding, transforming, or combining variables (e.g., by calculating summary scores).[6] Influence statistics can indicate outlying observations that change the estimated coefficients when removed from the sample.[5] It is useful to examine influential cases in more detail, and advisable to be careful when deciding whether or not to remove these cases.

2.1.6 Validation of model performance

The proof of the pudding is in the eating, certainly for risk models. The three most important types of predictive performance measures are discrimination measures, calibration measures and measures of clinical utility. Models can be validated using the data used for model development, or new data.

Performance measures

Discrimination assesses whether estimated risks are different for patients with and without the disease. The best-known measure is the c-statistic, which equals the area under the receiver operating characteristic (ROC) curve (AUC) for binary outcomes (Figure 3). It is the probability that a randomly selected patient experiencing an event has a higher predicted probability than a randomly selected patient not experiencing the event. The use of a risk threshold turns the risk model into a classification (or decision) rule. For example, a patient with a predicted probability of having a particular disease above 30% may be referred to a specialized clinician to undergo further tests.

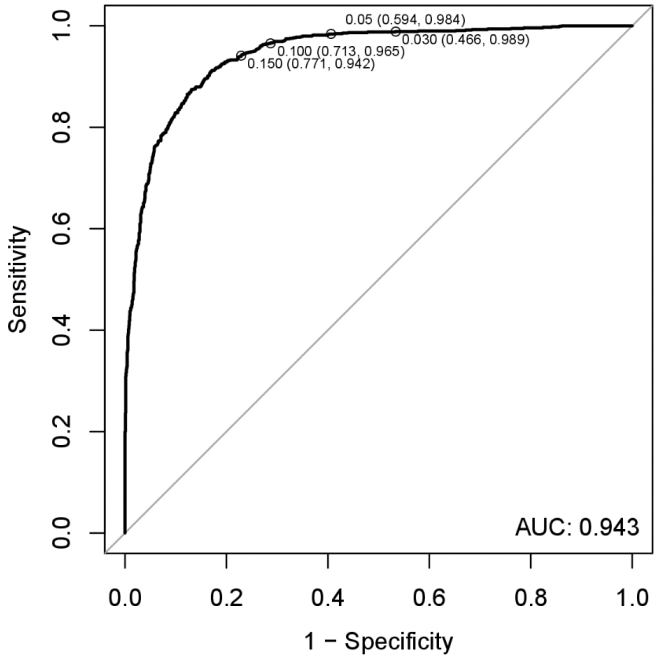


Figure 3. ROC curve for the ADNEX model to distinguish between malignant and benign lesions. The black line represents the ROC curve for the ADNEX model; the grey diagonal line presents the ROC curve for a model without discriminative ability; the dots represent the specificity and sensitivity for risk thresholds 0.03, 0.05, 0.10 and 0.15 (risk threshold (specificity, sensitivity)).

Thresholds can be selected in several ways.[61, 62] At the level of the individual patient, the relative consequences of a false positive and false negative decision can be taken into account. For example, a threshold below 50% indicates that a false negative is considered more harmful than a false positive. If the disease is life-threatening or has a serious impact on the patient's quality of life, a low risk threshold may be chosen, while a high risk threshold may be preferable if the treatment has important side-effects. Of course, patient characteristics could be important to motivate the choice of the risk threshold. For example, one may want to avoid loss of fertility as a side effect of an operation for young patients with a desire to have children, while it is not an important consideration for a patient who considers his family as complete. At the population level, a balance in terms of sensitivity and specificity can be sought. This framework is more useful for cost-effectiveness studies and for national guidelines that formulate recommendations for clinical practice, keeping in mind the national health care system. When a threshold is defined,

the patient classification and the actual disease status can be cross-tabulated and summarized with measures such as sensitivity, specificity, positive predictive value, and negative predictive value. The positive and negative predictive values directly depend on the event rate (i.e., the prevalence for diagnostic outcomes), but in practice sensitivity and specificity may also vary with event rate.[63]

Calibration assesses the agreement between the predicted risks and observed event rates. Different levels of calibration can be distinguished.[64] The first one is mean calibration and requires that the average predicted risk in the population corresponds to the observed event rate. This is also referred to as calibration-in-the large. The second level is weak calibration, and this requires that there should be no systematic over- or underestimation of risks and no systematic over- or underfitting of risks. Overfitting yields predicted risks for high-risk patients that are too high and predicted risks for low-risk patients that are too low, while underfitting yields predicted risks that are too moderate. This can be checked in the logistic calibration framework, in which a logistic regression model is used to model the outcome as a function of the linear predictor of the risk model. [6, 65] The third level, moderate calibration, requires that patients with a predicted risk of r% have an event rate of r%. This is easy to assess in a calibration plot (Figure 4).[6, 66,64] In this plot, the relationship between predicted risks and observed outcomes is visualized, using for example, a loess smoother or splines. A model that is well calibrated will yield a curve lying on the diagonal, suggesting that predicted risks are correct. For example, among patients with a predicted risk of 30%, 3 out of 10 have the event. The curve in Figure 4 suggests some underestimation of predicted risks. An alternative assessment of moderate calibration is to categorize predicted risk and assess whether the event rate corresponds to the average predicted risk in each risk group. The Hosmer-Lemeshow test is frequently used for this purpose, but neither assesses direction or magnitude of (mis)calibration, and suffers from serious drawbacks such as low power. Its use is therefore widely discouraged.[67, 68] The highest level of calibration, strong calibration, requires that the event rate equals the predicted risk for every covariate pattern. However, attempting to reach strong calibration during model development may be unrealistic and yield overly complex models.

Discrimination and calibration do not address the clinical consequences of using a risk model.[69-71] One measure to assess the *clinical utility* for decision-making is the net benefit, which is plotted in a decision curve (Figure 5).[72, 73] Net benefit (NB) relies on the relationship between the risk threshold and the relative importance of true versus false positives:[74] when a risk threshold of

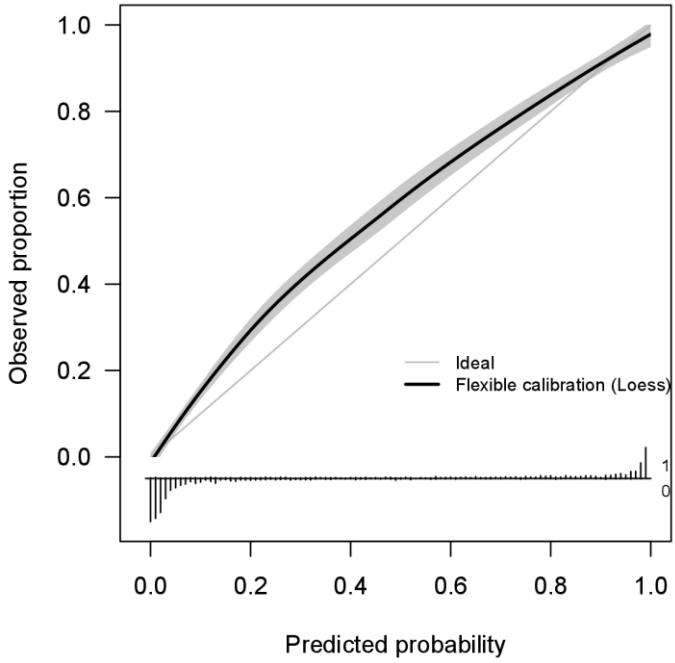


Figure 4. Calibration plot of the ADNEX model at external validation. Thick line: Loess smoother indicating the relation between the predicted probability of malignancy by the ADNEX model and observed probability. Grey band: 95% confidence interval for the loess smoother. The thin diagonal line indicates perfect calibration. The histogram on the x-axis shows the distribution of predicted probabilities of malignancy, with the frequencies of predicted probabilities for events plotted above and for non-events below the x-axis.

10% (odds 1:9) is adopted to select patients for treatment, this is equivalent to accepting up to 9 false positives per true positive. NB equals the net proportion of true positives: $(r_{11} – (1/9) \times r_{01}) / N$ for a 10% risk threshold, with N the total number of observations, r_{11} the number of true positives and r_{01} the number of false positives. The decision curve plots NB for different risk thresholds. The NB of the model can be compared with the NB of treating all patients or treating no patients (see Figure 5): if the model has a higher NB at a given threshold, this means that treatment decisions at that threshold have utility. If the model has a lower NB than any of the default strategies, (i.e., treat all or treat none), it is said to be harmful. It was proven that moderately

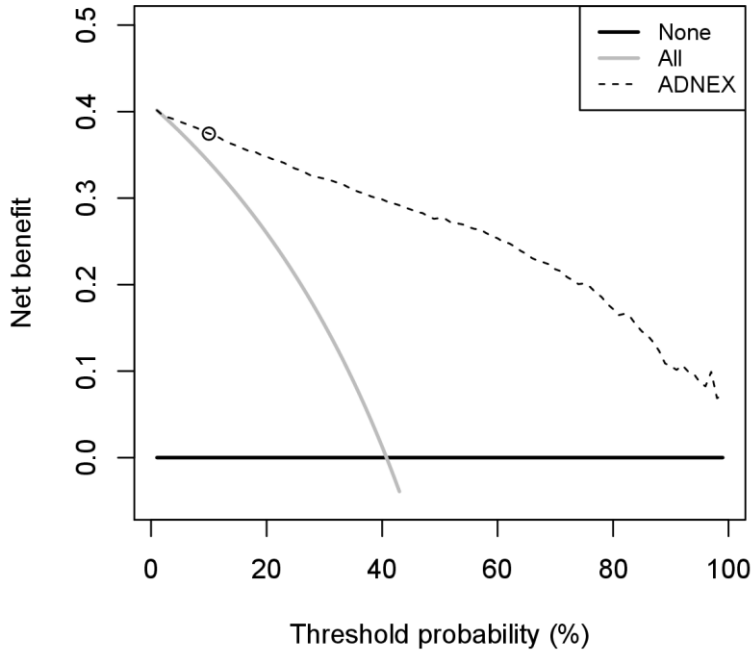


Figure 5. Net benefit of the ADNEX model at external validation. Dashed line: net benefit (NB) of the ADNEX model to distinguish between benign and malignant lesions. Grey line: NB of classifying all tumors as malignant. Black line at zero: NB of classifying all tumors as benign (none as malignant). Dot: NB of ADNEX at threshold probability 10%.

calibrated risk models are never harmful.[64] Poor discrimination and miscalibration reduce the NB, [75] but NB goes beyond these purely statistical measures to capture clinical utility. Despite being introduced recently, measures for clinical utility have already been recommended by several leading journals.[73, 76-80] Besides NB, also other measures of clinical utility exist, such as relative utility, which can be shown to be a function of the NB of the model and the NB of the best default strategy.[79, 81] However, measures for clinical utility are not always of interest. For example, a risk model to merely inform pregnant couples of the chance of a successful pregnancy beyond the first trimester[60] would not require an analysis of clinical utility, as no decision needs to be made. Calibration is the key measure in such contexts.

Models can be compared by calculating the difference in the c-statistic, by comparing calibration and by comparing clinical utility.[82] The significance of

an added predictor can be based on a test for the predictor's regression coefficient rather than on tests for performance improvement (e.g., change in AUC).[83] Reclassification statistics such as net reclassification improvement (NRI) and integrated discrimination improvement (IDI) have fueled intense debate.[84-86] NRI and IDI depend on calibration in the sense that models may appear advantageous simply because of poor calibration.[87-89]

Apparent, internal and external validation

Models can be validated internally, using the same dataset that was used to develop the model, or externally, using a new dataset. The model is usually evaluated first with internal validation.[90] The most popular approach is to randomly split the dataset into a development set and an internal validation set. However, this approach is inefficient and should be avoided, especially in small sample studies (Table 1, pitfall 6). Alternative approaches such as cross-validation or bootstrapping are recommended.[91] Internal validation should be distinguished from apparent validation,[6] which is the validation of the model on exactly the same dataset that was used for model development, without the use of split sample, cross-validation or bootstrapping techniques. Apparent validation typically yields overoptimistic performance measures, and calibration intercepts and slopes are by definition perfect. Hence, apparent performance results are not useful.

The real test of a model involves external validation on new data, either collected at a later point in time from the same center(s) (temporal validation) or collected in different centers (geographical validation).[90, 92] When several models for the same purpose exist, it is recommended to directly compare these models in a single external validation study.[82] However, it is disappointing that there are many more publications describing the development of new risk models than there are publications externally validating existing models (Table 1, pitfall 8).[23, 93, 94] It would be better to develop fewer models that hold more promise to be robust and useful.

A reliable external validation study requires a sufficiently large sample size. At least 100 but preferably 200 events are recommended.[64, 68, 95, 96] External validation often results in poorer performance.[93] There are many factors that can explain results at external validation, and hence these results have to be interpreted carefully, often in the context of differences in case-mix between the development and validation datasets.[97, 98] For example, discrimination is typically lower in more homogeneous populations. Even temporal validation may show performance degradation, for example because the center changed

from a secondary to a tertiary care center, or because new clinical guidelines have changed the population.[99]

If the performance of a model is disappointing, updating the model is a sensible solution that is more efficient than developing a completely new model, because the original model still contains useful information. Different approaches exist, such as intercept adjustment, rescaling of model coefficients, model refitting, or model extension.[100, 101, 102]

Publications describing the development of new risk models should always include an internal validation,[3, 19] but sometimes models are immediately validated externally (i.e., evaluated on a different dataset that is collected at a different time point and/or at a different location).[90] Assessment of discrimination and detecting lack of model fit using flexible calibration curves[64] are important considerations at internal validation, while calibration[64] and the assessment of clinical utility are important at external validation.

2.1.7 Reporting

Complete and transparent communication of model development and validation ensures that others can fully understand what was done. Reviews of reporting quality have repeatedly revealed severe shortcomings (Table 1, pitfall 7).[94, 103-105] This led to comprehensive guidelines and checklists for manuscript preparation such as the recent TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement.[3, 19]

The exact formula of the model (including the intercept or baseline survival at a given time point) should always be reported in the publication to allow independent users to validate or use it.[3, 106] To aid the uptake of risk models, they are often simplified or presented in alternative formats, such as score charts, nomograms, and color bars.[5, 6, 60, 78, 107] Clear eligibility criteria should be presented along with the model, including ranges of continuous predictors, so that users are made aware if they are extrapolating beyond the ranges observed in the development data.

2.1.8 Impact studies

The ultimate aim of many risk models is to improve patient care. If a model proves to maintain a good predictive performance at extensive external validation, the final step would be an impact study, perhaps together with a

cost-effectiveness analysis.[23, 108, 109] Ideally, impact studies are (cluster) randomized studies for which primary endpoint(s) are clinical care parameters such as length of hospitalization, number of unnecessary operations, days off work, time to diagnosis, morbidity, or quality of life.[108, 110, 111] Unfortunately, few risk models reach this stage of investigation,[23, 108] although predictive performance is no guarantee for beneficial impact on patient outcomes.[112, 113]

2.1.9 Model implementation

User-friendly implementations increase the uptake of a model.[114] The model can, for example, be implemented in a spreadsheet for use with office software, or made accessible on websites (e.g., www.qrisk.org). With the rise of smartphones and tablets, implementation is currently shifting towards mobile applications. In any case, disseminating insufficiently validated risk models has to be avoided.

2.1.10 Conclusion

The development and validation of risk models should be appropriately planned, conducted, presented, and reported. Unfortunately, reviews have demonstrated that methodological and reporting standards are often not met.[2, 94, 103, 105] Risk models should only be developed when there is a clinical need, and external validation studies should be given the attention and prominence they deserve. The development of a robust model and an informative validation is not straightforward, with several pitfalls looming along the way. Perfection is impossible, but adhering to established methodological standards is important to arrive at a good model that has the potential to be useful in clinical practice.

2.2 Predicting successful vaginal birth after a Cesarean section

The prediction of successful vaginal birth after a Cesarean section serves as an illustration of the development and internal validation of a risk model in a small sample. A woman who has had a Cesarean section (CS), may be scheduled for a CS again in a subsequent pregnancy, by choice or because of compelling medical reasons. This is called an elective repeated Cesarean delivery. However, women may also plan to deliver vaginally (VBAC, vaginal birth after Cesarean section). If the attempted VBAC is unsuccessful, an

emergency repeated Cesarean delivery is needed. In this section, we describe the development and validation of a model to predict the probability of a successful VBAC.

2.2.1 The clinical need for a new VBAC model

CS rates have increased globally over recent years. Accurately predicting the likely success of a trial of labor is clinically useful, as a failed trial of VBAC is associated with increased maternal and fetal morbidity.[115] Several authors have attempted to predict which patients are likely to have a successful VBAC by using clinical history and physical examination at the time of admission for delivery. Flamm and Geiger developed a scoring system to predict the success of a trial of labor using maternal age, history of previous vaginal delivery, the indication for a previous cesarean delivery, cervical effacement and dilation.[116] After external validation,[117] it was concluded that an unfavorable score might be helpful when counseling patients who consider a trial of labor. However, this system does not give a calibrated spectrum of risks and leaves any decision about VBAC until a woman is admitted to the delivery room in labor. A better system is needed, one that predicts the success of an attempted VBAC at an earlier stage.

There is a growing body of evidence that suggests that complete healing of the previous cesarean scar and myometrial thickness at the lower uterine segment are related to the chance of achieving a vaginal delivery in a subsequent pregnancy.[118] Over the last ten years there have been multiple attempts to study the prevalence and the clinical significance of apparently "defective" Cesarean scars visualized with ultrasonography.[119-126] In these studies, the prevalence of visible CS scars ranged from 19 to 88%.

Despite this interest in scar morphology there have been no longitudinal studies to demonstrate a link between scar appearances on ultrasound and the prediction of the outcome of a trial of VBAC. Some studies focused on single residual myometrial thickness (RMT) measurements at 36 weeks' gestation.[127-129] Eden carried out a systematic review to evaluate VBAC prediction models and concluded that none consistently identified women at risk of a failed trial of labor.[130] This study aims to use single measurements of CS scar dimensions as well as the change in these dimensions over time to predict the probability of successful vaginal delivery in women selected for a trial of VBAC.

2.2.2 Subjects and methods

Study design

The data used in this study was collected at the ultrasound department of a London university teaching hospital between June 2010 and July 2011, to develop reference charts of changes in CS scar size over the course of a pregnancy.[131] The local research and ethics committee approved the protocol and written informed consent was obtained from all participating women. In total, a cohort of 320 women with singleton pregnancies was followed up throughout their pregnancies. Women with a history of one previous CS were considered suitable for a trial of VBAC and were referred to a specialized VBAC clinic. Per hospital policy, women with two or more previous CS were booked for an elective repeated CS.

Women were scanned by transvaginal ultrasound by one of two clinicians, using a Voluson E8 Expert (GE healthcare, Wisconsin, USA) ultrasound system equipped with a 5-9 MHz vaginal probe. This occurred at 11-13, 19-21 and 32-34 weeks' gestational age, coinciding with their routine scans. At the first scan visit, demographic data were recorded and an obstetric history was obtained. CS scar measurements were repeated at the second and third visit. Operators performing the scans were blinded to this clinical information. A CS scar was defined as visible on ultrasound when it showed both a hypoechoic indentation representing myometrial discontinuity at the anterior wall of the lower uterine segment, with or without underlying residual myometrium. The delineation and measurements of the CS scar using ultrasound were carried out according to methods reported previously.[126, 131, 132] The RMT was defined as the thickness of the part of the scar contained in any residual myometrium, and was measured in the sagittal plane. Patients were followed up until the final pregnancy outcome was known. Clinicians responsible for the care of the patients in labor were blinded to the ultrasound features of the CS scar. All data were transmitted weekly to a designated research database and were checked for missing, out of range, and inconsistent values. Women with non-visible CS scars, women with a planned repeated CS and women with an emergency repeated CS before 38 weeks' gestational age were excluded from the analysis.

Modeling strategy

The factors considered for use in the predictive model were selected relying on experts' opinions. These included two demographic variables (patients'

age and pre-pregnancy body mass index (BMI)), one variable related to previous obstetrics history (prior successful VBAC) and two CS scar variables (the RMT measured at the second trimester and the change in RMT between the first and second trimester). Considering the purpose of this study and the limited sample size, the associations of pregestational diabetes mellitus, steroid use and smoking with VBAC were not considered. Logistic regression was used to assess the significance of predictors in a prediction model combining patient and scar characteristics. Because the sample size was small relative to the number of predictors, penalized maximum likelihood based on the corrected Akaike's information criterion was used to obtain reliable regression coefficients.[5] Linearity of predictor effects was checked using plots of partial residuals against the predictors. Interactions were not considered. Variance inflation factors were used to detect strong relationships among the predictors. Finally, the distributions of the DFBETAS (difference in beta values) were checked but no outliers were identified.

Model performance

The c-statistic and the sensitivity and specificity (at a range of cut-off values) were computed to assess and compare the discriminative performance of both models. Calibration plots were made to investigate the agreement between observed and predicted probabilities. Finally, the clinical utility of the model (see section 2.1.6) was assessed with decision curves, plotting the net benefit (NB) for threshold probabilities between 0.3 and 0.9.[72] A threshold probability of 0.7 to classify patients as suited for attempting vaginal delivery has often been used in earlier studies.[130] The odds of the threshold represent the relative misclassification costs, with a value above one indicating that an emergency repeated CS (i.e., a false positive) is worse than an unnecessary elective repeated CS (i.e., a false negative). For example, using a threshold probability of 0.7 (odds 2.3) implies that the benefit of avoiding one emergency CS is worth the harm of 2.3 unnecessary elective CSs. The NB of the prediction model is compared to the NB of two default strategies for which no model is needed: attempting a VBAC in all patients or attempting a VBAC in none of the patients.

To validate the prediction models internally we did not split the data into training and test sets due to the limited sample size. Rather, bootstrap resampling was used to estimate performance on new patients. The model was fitted on each of 1000 bootstrap samples. The optimism in model performance was estimated as the average decrease in performance when the model was tested on the original sample instead of the bootstrap sample. Finally, the

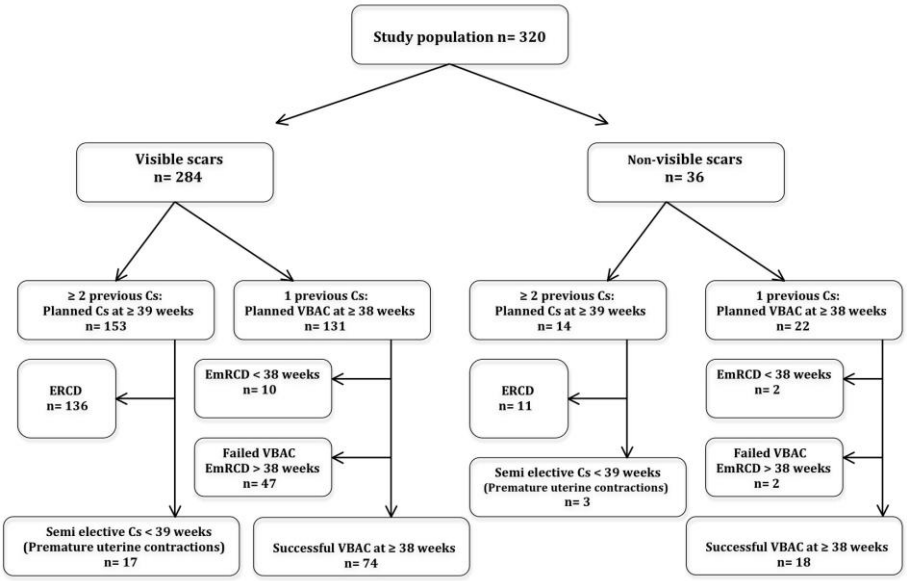


Figure 6. Flow chart summarizing the study sample, showing the mode of delivery according to Cesarean section (CS) scar visibility. EmRCD: emergency repeated Cesarean delivery; ERCD: elective repeated Cesarean delivery; VBAC: vaginal birth after Cesarean section.

performance estimate at internal validation was obtained by subtracting the estimated optimism from the original performance estimate.[6, 91]

To assess the added value of scar characteristics in a prediction model, the model was compared to a reduced regression model, including only patient characteristics. This model was built and evaluated using the same techniques as described above. Analyses were performed using R (http://www.r-project.org/).[133] The rms package was used to fit the logistic regression model, to calculate the penalized coefficients and the c-statistic, and to produce the calibration plot.[5]

2.2.3 Results

The scar was visible in 284 out of 320 cases (89%) (Figure 6). 153 cases had a history of more than one previous CS and therefore were assigned to have an elective CS at ≥ 39 weeks' gestation. 131 women with a history of 1 previous CS and visible scar were counseled and offered a trial of VBAC. 10

Table 2. Indications for emergency Cesarean sections in 47 women with visible CS scars and one previous CS, who attempted vaginal birth after a Cesarean section.

Indication	**N(%)**
Confirmed scar rupture	2 (4.3)
Antepartum hemorrhage	2 (4.3)
Placenta previa	2 (4.3)
Breech presentation	3 (6.4)
Fetal distress	14 (29.8)
Fetal distress secondary to placental abruption	1 (2.1)
Failure to progress	20 (42.6)
PET/GDM/PROM	2 (4.3)
Placental abruption	1 (2.1)

GDM: gestational diabetes mellitus; PET: pre-eclamptic toxemia, PROM: premature rupture of membranes.

women were subsequently excluded because they underwent CS due to various causes. Of the remaining 121 cases, successful VBAC was achieved in 74 (61%), while 47 cases had failed trials of labor, resulting in an intrapartum emergency Cesarean delivery for various intrapartum reasons (Table 2).

Women with successful and failed VBACs were similar in terms of age and pre-pregnancy BMI (Table 3). Women with a successful VBAC often had a previous successful VBAC (33.8%, versus 6.4% among women with a failed VBAC). The median RMT was higher for patients who subsequently underwent a successful VBAC than for patients with a failed VBAC (e.g., RMT in the second trimester was 4.2 versus 2.8 mm) and the median difference in RMT between the first and second semester was lower for women with a successful VBAC than for women with failed trials (0.8 vs. 2.8 mm).

In a regression model including age, BMI, RMT, ΔRMT, and previous VBAC, BMI had a small positive effect on the chance of successful vaginal delivery and did not discriminate between successful and failed trials (OR=1.08 per unit increase, 95% CI 0.95 to 1.23). This effect is contrary to findings in the literature[130] and was excluded from the prediction model to enhance generalizability.

Table 3. Descriptive statistics of patient and Cesarean section (CS) scar characteristics according to mode of delivery in 121 women with visible CS scar and one previous CS.

	Failed VBAC (n=47)	**Successful VBAC (n=74)**
Age (years)	33	32
	(20, 28, 35, 43)	(21, 29, 34, 39)
Body mass index (kg/m^2)	25	27
	(19, 23, 29, 40)	(18, 24, 30, 41)
Previous VBAC	3 (6.4)	25 (33.8)
RMT (mm)		
First trimester	5.5	5.9
	(3.0, 4.8, 6.4, 9.1)	(2.8, 3.5, 5.9, 7.6)
Second trimester	2.8	4.2
	(0.5, 2.6, 3.1, 4.2)	(2.6, 3.0, 4.6, 6.3)
Third trimester	2.5	3.6
	(0.5, 2.4, 2.6, 3.8)	(3.2, 2.6, 3.9, 5.9)
ΔRMT: second trimester– first trimester (mm)	2.8	0.8
	(0.1, 2.1, 3.4, 6.2)	(0.1, 0.4, 1.2, 4.7)

Results presented as median (minimum, 1st quartile, 3rd quartile, maximum) or n (%). GA: gestational age; RMT: residual myometrial thickness; VBAC: vaginal birth after a Cesarean section.

The fitted model suggests that the chance of a successful VBAC increases with RMT (OR=6.26 per mm increase, 95% CI 2.12 to 18.52). In addition, the larger the decrease in RMT from the first to the second trimester, the lower the probability of a successful VBAC (OR=0.25 per mm difference, 95% CI 0.13 to 0.48). The effect of maternal age tended to be negative while the effect of previous VBAC tended to be positive (OR=0.70, 95% CI 0.35 to 1.37, OR=3.28, 95% CI 0.50 to 21.47, respectively). The penalized effect estimates are given in Table 4 alongside the non-penalized effect estimates and are preferred for making predictions in new patients (an example is given in the discussion). Performance measurements are shown for the predictions using the penalized coefficients, but differences are minimal.

At a probability cut-off of 0.7, the prediction model had a optimism-corrected sensitivity and specificity of 90% (Table 5). The c-statistic on the development data was 0.96 (95% CI 0.90 to 0.98), while the optimism-corrected c-statistic was 0.94 (95% CI 0.89 to 0.96). In comparison, the reduced model without scar features had a c-statistic of 0.66 (CI 95% 0.55 to 0.77) on the development data and an optimism-corrected c-statistic of 0.62 (CI 95% 0.51 to 0.73).

Table 4. Estimated effects of patient and scar characteristics in a fitted logistic regression model to the predicted success of a trial of vaginal birth after a Cesarean section.

Effect	Penalized coefficient	Non-penalized coefficient	OR non-penalized (95% CI)
Intercept	–0.77	–1.34	—
Age (per 5 year increase)	–0.29	–0.36	0.70 (0.35–1.37)
RMT (per mm increase)	1.44	1.83	6.26 (2.12–18.52)
ΔRMT (per mm decrease)	–1.22	–1.39	0.25 (0.13–0.48)
Previous VBAC (yes/no)	1.09	1.19	3.28 (0.50–21.47)

OR: odds ratio; RMT: residual myometrial thickness; VBAC: vaginal birth after Cesarean section.

Table 5. Sensitivity and specificity for success of a trial of vaginal birth after a Cesarean section, using the prediction model at specified cut-off values for predicted probability.

Cut-off for probability	Sensitivity	Specificity
0.3	97% (97%)	74% (74%)
0.5	95% (94%)	83% (81%)
0.7	90% (90%)	91% (90%)
0.9	59% (59%)	98% (96%)

Apparent sensitivity/specificity (optimism-corrected sensitivity/specificity).

The estimated probabilities of a successful VBAC given by the full model were accurate (Figure 7). There is some underestimation of the probability (by 0.1 or less) for predicted probabilities between 0.2 and 0.7, but the vast majority of women have predicted probabilities outside this range. The reduced model yielded badly calibrated probabilities: the probability of a successful VBAC was severely overestimated (results not shown).

The model clearly has a greater NB than an elective Cesarean in all women or an attempted VBAC in all women (Figure 8). At the probability threshold of 0.7, the optimism-corrected NB of the full model was 0.45 higher than that of giving none of the women a trial at VBAC. This means that the benefit of using the model with scar characteristics in addition to patient characteristics

is equivalent to identifying 45 additional successful VBACs per 100 patients at the same rate of emergency Cesareans. The model also has clinical utility compared to the default strategy of giving all women a trial of VBAC. This was the case even at a threshold probability of 0.3, at which the optimism-corrected NB of the model was 0.11 higher. This is equivalent to 26 (= 0.11×100/(0.30/0.70)) fewer emergency CS deliveries per 100 patients with no increase in the number of unnecessary elective CS deliveries.

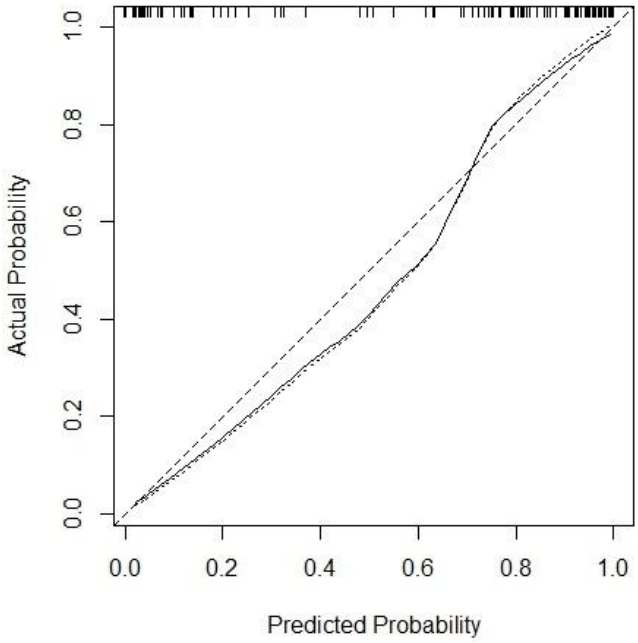


Figure 7. Calibration plot for the model to predict successful vaginal delivery after a Cesarean section (CS), based on patient age, previous history of VBAC, residual myometrial thickness and change in RMT between first and second trimester, in 121 women with visible CS scar and one previous CS: apparent (…), optimism-corrected (—) and ideal (---) calibration. The vertical lines at the top represent individual data points.

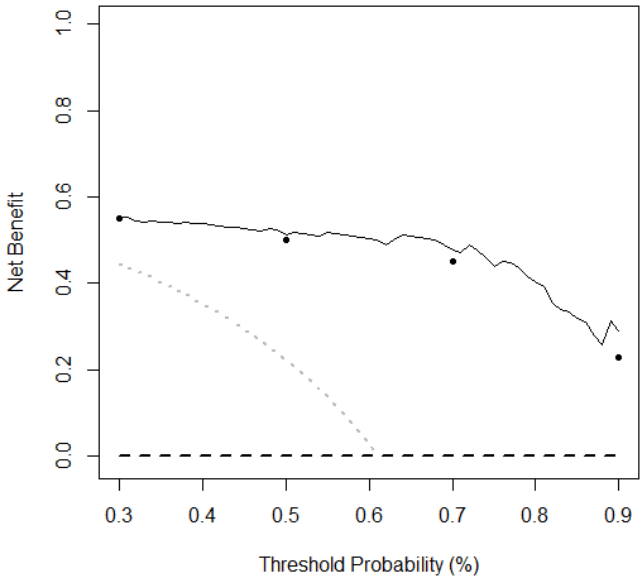


Figure 8. Decision curves for three strategies in the management of delivery in women with visible Cesarean section (CS) scars and one previous CS: using the prediction model for success of vaginal birth after CS (VBAC) (—), giving all women a trial of VBAC (...) and giving no woman a trial of VBAC (---). Optimism-corrected net benefit is indicated with dots.

2.2.4 Discussion

We have shown that when ultrasound measurements of CS scars during pregnancy are incorporated into a prediction model, the success of a trial of labor can be predicted with a high level of accuracy. Previously published models have focused on single RMT measurements made in the late third trimester. In contrast, our study suggests that if an assessment of the scar is made relatively early in the pregnancy (first and second trimester), the decisions regarding the mode of delivery can be made well in advance of any potential delivery date.

We found that women who underwent successful trials of labor had greater RMT measurements in the second trimester compared to women who had a failed VBAC. Furthermore, women with a successful VBAC showed minimal

changes in RMT between the first and second trimester. Our findings suggest that the observation of a large hypoechoic part of the scar may not indicate that scar integrity is compromised, provided that the RMT is sufficiently thick and remains so during the pregnancy.[131] The outcome when the scar could not be visualized may support this: when a trial of VBAC was attempted in these women, 82% of cases resulted in a successful vaginal delivery. An inability to visualize a scar (providing the assumed location of the CS scar can be located anatomically) probably reflects good scar healing, with no hypoechoic segment on ultrasound.

There have been several reports describing prediction models to help clinicians counsel women regarding the mode of delivery after a previous CS. However, the use of a combination of CS scar measurements with clinical variables to predict the success or failure of a trial of VBAC has not been described previously. As the uterus changes in size so rapidly during the course of a pregnancy, we have introduced the change in scar size over time as a predictive variable.

A strength of our model is that it provides both patient and clinician with a probability for the likely success of a trial of VBAC. We can illustrate this by giving two relatively extreme examples. The formula for the probability of success (based on the coefficients provided in Table 4) is $P=1/(1 + \exp(-z))$, where exp refers to the exponential function and $z= -0.77 - 0.29 \times \text{age}/5 + 1.44 \times \text{RMT} - 1.22 \times \Delta\text{RMT} + 1.09 \times$ (1 if previous VBAC). If we consider a woman aged 32, without a previous VBAC, with an RMT on her second-trimester scan of 2.7 mm and a decrease in RMT of 1.5 mm between the first and second trimesters, then, by entering the specified values, the estimated chance of successful vaginal delivery for this patient is 36%. In contrast, if the same woman had an RMT of 4.1 mm at the second-trimester ultrasound scan and a decrease in RMT of 0.9 mm between the first and second trimesters, she would have an estimated 90% chance of a successful vaginal delivery. There were two cases of confirmed uterine rupture in the group selected for a trial of VBAC in our study population. If we apply the model in these cases, the estimated chance of their achieving a successful VBAC would be 0.00005% and 0.00009%, respectively. Had the model been applied prospectively, both women would have been strongly advised against a trial of VBAC and so would have avoided the morbidity associated with uterine rupture.

The c-statistic for our model is very high. This may reflect the relatively small number of women in the study. However, even if interpreted with caution, the findings suggest that the scar is a major factor in the function of the uterus during a trial of labor. A key issue for any test is reproducibility. The

interobserver variability for measurement of CS scars is good in the first trimester and moderate in the second and third trimesters.[126] It is therefore reasonable to suggest that these measurements will be of value if used more widely. It is important to highlight that whilst this model and approach show promise, external validation in different units and populations is needed before introduction into clinical practice.

3 Statistical methods for multicenter data

This chapter explains how to handle a clustered dataset, encountered in multicenter studies. Section 3.1 explains how clustered data influence statistical analyses. Section 3.2 gives a birds-eye overview of methods that can be used to model clustered data. Section 3.3 introduces the multicenter International Ovarian Tumor Analysis (IOTA) dataset. Section 3.4 concludes. The focus in this chapter is on binary outcomes.

Section 3.3.4 is based on the following research paper:

Timmerman D, Van Calster B, Testa A, Savelli L, Fischerova D, Froyman W, Wynants L, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis (IOTA) group. *Am J Obstet Gynecol.* 2016; 214: 424-37.

3.1 Issues and opportunities of clustered data

Clustered data is characterized by a hierarchically nested structure, caused by a natural grouping in the population.[134] Examples include patients in hospitals, pupils in schools, and batches of a product. Each level of the nested structure represents a population and is a source of variability in the data: there is variability between patients but also between hospitals. Note that one can consider a population of hospitals but not a population of, e.g., genders. Although patients belong to a gender as much as they belong to a hospital, there is only a limited amount of genders, but a large number of hospitals. Gender is a patient-level characteristic, while hospital is an additional level in the nested structure. This chapter focusses on two-level structures but extensions to higher-level structures are straightforward, for instance patients within centers within countries.

A clustered data structure implies that observations within a cluster have more in common than observations belonging to different clusters. This violates the assumption of independence of observations, which is key in statistics.[134] As a consequence, estimated standard errors tend to be incorrect, affecting p-values and confidence intervals. Another consequence is that the effective sample size may be smaller than the total number of observations. This is because the variance of an estimator in a clustered data structure is typically larger than the variance of an estimator if observations are independent.[134]

Despite these complications, there are two important reasons to work with multicenter data: generalizability of the results and efficiency of the data collection.[135] Natural groupings are suited for a two-stage sampling procedure, that is, first select hospitals and subsequently select patients from the selected hospitals. Selecting a representative sample of hospitals enhances the generalizability of the results to patients in other hospitals. Moreover, collecting data at multiple sites at once reduces the recruitment time required to collect data on a sufficiently large sample of patients, which may be especially useful when studying a rare disease.

In a prediction framework, multicenter data offers one important additional opportunity: to make use of the information that characterizes a hospital yet is not captured by patient-level information. This could improve predictions, tailoring them to the center a patient belongs to. In an etiological or explanatory framework, multicenter data allows studying effects at the patient level and at the center level, and even cross-level interaction effects and variance at the patient and center level that is dependent on the value of predictors.

3.2 Methods for clustered data

There are many ways to deal with clustered data during the analysis. They can be grouped as follows: techniques that ignore the clustered data structure, techniques that can offer center-specific predictions and techniques that correct for clustering when making inferences but do not offer center-specific predictions (see Table 6). In addition, a meta-analysis can combine the results of multiple single-center studies.

Table 6. Methods to analyze multicenter datasets.

Method	Cluster-specific predictions	Correction during inference	Regression coefficients
Standard logistic regression	No	No	Marginal
Fixed effects logistic regression	Yes	Yes	Conditional
Mixed effects logistic regression	Yes	Yes	Conditional
Generalized estimating equations	No	Yes	Marginal
Logistic regression with cluster-robust standard errors	No	Yes	Marginal
Logistic regression with survey weights	No	Yes	Marginal
Logistic regression with bootstrap confidence intervals	No	Yes	Marginal
Conditional logististic regression	No	Yes	Conditional

3.2.1 Ignoring the clustered data structure

The first option is to analyze the patient data as if they were independent. Logistic regression is a common technique for estimating risk models for diagnosis.[3, 6] Let Y_{ij} be the 1 if individual i ($i=1,\ldots,n_j$) from center j ($j=1,\ldots J$) experiences the event of interest and 0 otherwise. X_{kij} is the kth predictor ($k=1,\ldots,K$) and $p_{ij}=P(Y_{ij}=1)$ the probability that an individual experiences the event of interest. The logistic regression model expresses p_{ij} as a linear combination of predictors X_{kij}, using the logit as a link function:

(1)

$$Y_{ij} \sim bin(1, p_{ij})$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha+\sum_{k=1}^{K}\beta_k X_{kij}\,.$$

The intercept α and regression coefficients β_k are estimated using maximum likelihood.

The standard logistic regression model is a population-averaged or marginal model.[136] Its regression coefficients have a marginal interpretation: they estimate the average effects in the population. Its predictions are marginal as well, meaning that the predicted probability for an individual patient reflects the average probability of patients with the same observed values of predictors, but ignores the centers the patients came from. The predicted probability of an event is computed by taking the inverse logit of the linear predictor (LP) of the estimated model:

(2)

$$LP_{LR\ ij}=\hat{\alpha}+\sum_{k=1}^{K}\hat{\beta}_k X_{kij}$$

(3)

$$\hat{p}_{LR\ ij}=\frac{1}{1+\exp(-LP_{LR\ ij})}\,.$$

Ignoring the clustered data structure provides correct results for statistical tests if only patient-level predictors are included and the effective sample size is used for the computation of standard errors.[134, 136, 137] Otherwise, the estimates of standard errors are wrong and confidence intervals are often too narrow.

Model (1) can be extended by including center characteristics Z_j:[134]

(4)

$$Y_{ij}\sim bin(1,\ p_{ij})$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha+\sum_{k=1}^{K}\beta_k X_{kij}+\sum_{l=1}^{L}\beta_l Z_{lj}\,.$$

The effect estimates and predictions are now adjusted for the center characteristics. However, this model does not take into account that the Z_{lj} are measured on only J centers, instead of N patients. Hence, there is a serious

risk that confidence intervals around the effect of the center-level variables Z_{lj} are too narrow.[134]

Analyzing the patient data as described in equations (1) and (4) is also known as disaggregation. The opposite method is aggregation, where all patient-level information is aggregated to the hospital level, for example, by averaging, and regressing the logit of the average event probability in center j ($\overline{p}_{.j}$) on the average predictors in center j ($\overline{X_{k.j}}$):[134]

(5)

$$\log\left(\frac{\overline{p}_{.j}}{1-\overline{p}_{.j}}\right)=\alpha+\sum_{k=1}^{K}\beta_k\overline{X_{k.j}}+\sum_{l=1}^{L}\beta_l Z_{lj}\,.$$

With this technique one may be tempted to make an ecological fallacy, that is, to make assertions about associations at the patient level based on associations between aggregated variables. Aggregation does not allow predictions for individual patients and hence, this technique is of little interest in a prediction framework.

3.2.2 Center-specific models

Center-specific models are able to offer center-specific predictions. In contrast to marginal models, they include center-specific parameters.[136] The effect estimates have a conditional interpretation: the regression coefficients obtained reflect the predictor effect within a center (i.e., the average center or the reference center).[138] Conditional regression coefficients are typically further from zero than marginal regression coefficients in the logistic regression framework (see chapter 5.2).[139] Importantly, center-specific models provide correct inferences. There are two frequently used ways to construct a center-specific model: the fixed effects model and the mixed effects model. Model updating[6] can be regarded as an additional way to obtain a center-specific risk model. This technique is suited when a risk model is already available and patient data from a specific center can be used to tailor it to the local setting.

A fixed effects model

Model (1) can be extended by treating center as a categorical variable and adding m=J-1 dummy variables D_{mj} to model the center effects:[136]

(6)

$$Y_{ij} \sim \text{bin}(1, p_{ij})$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + \sum_{m=1}^{J-1} a_m D_{mj} + \sum_{k=1}^{K} \beta_k X_{kij} .$$

Each center now has its own intercept: α for the reference center and $\alpha+a_m$ for the other centers. Center-level characteristics can no longer be included in this model, as all variation between centers is explained by the estimation of the center effects a_m.[134] The model can be extended further by including interactions between the center dummies and the individual-level X_{kij}.

As the number of parameters to be estimated in this model is strongly dependent on the number of centers, this method is only suited if there is a small number of centers, say <10.[134, 140] Otherwise predictor effects may be biased and standard errors may be underestimated.[140, 141] Moreover, large numbers of patients per center (>100) and a sufficient number of events per center are required. Otherwise, center effects may be overfitted and the standard errors of predictor effects may be inflated.[134, 140, 141]

The conditional linear predictor for the ith patient from the jth center includes the estimated center effect:

(7)

$$LP_{FLR\ ij} = \hat{\alpha} + \hat{a}_m + \sum_{k=1}^{K} \hat{\beta}_k X_{kij} .$$

Conditional predicted probabilities $\hat{p}_{FLR\ c\ ij}$ are obtained by taking the inverse logit of the conditional linear predictor.

The fixed effects model treats the centers in the dataset as if they are the only centers of interest, rather than as a sample of a population of centers. To apply a risk model developed with fixed effects logistic regression in a new center, an estimate of the center effect must somehow be obtained. Debray and colleagues proposed to estimate an average center intercept using meta-analytic techniques, to select the center effect of a similar center from the study population, or to estimate the center intercept from the outcome prevalence (but this requires that predictors in the model are all mean-centered).[142, 143]

A mixed effects model

If the centers have been chosen from a population of possible centers with the intention to answer the question "do centers differ?", then they are said to have random effects.[144] This is in contrast with the approach outlined above, where the centers in the model development dataset are the only centers of interest, and they are said to have fixed effects.[144] A model that includes both random (e.g., center) and fixed (e.g., predictor) effects, is a mixed effects model.

The simplest form of a mixed effects logistic regression model is a random intercept model. This can be used to model heterogeneity of the event rate across centers by allowing intercepts to vary between centers. In this case, one extra parameter needs to be estimated, alongside the beta coefficients of fixed effect predictors and the overall intercept: the random intercept variance τ^2. The random center intercepts a_j are assumed to be normally distributed with mean zero.

(8)

$$Y_{ij} \sim \text{bin}(1, p_{ij})$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + a_j + \sum_{k=1}^{K} \beta_k X_{kij}$$

$$a_j \sim N(0, \tau^2)$$

The proportion of unexplained variance that can be attributed to the center level, rather than the patient level, can be calculated from the random effect variance τ^2:

(9)

$$\frac{\tau^2}{\sigma^2_{error} + \tau^2},$$

where σ^2_{error} is the variance of the patient-level error terms. These have a logistic distribution with a variance of $\pi^2/3 \approx 3.29$ in a logistic regression model. The proportion given by equation (9) is often referred to as the intraclass correlation (ICC). In what follows, we will distinguish between the ICC, which is estimated from an empty random intercepts model without predictors, and the residual intraclass correlation (RICC), which is estimated from a model including predictors (such as model (8)). The former is the proportion of total variance that can be attributed to the center level, while the

latter is the proportion of unexplained variance that can be attributed to the center level. These quantities will be studied more in depth in section 5.1.

The mixed effects model can be extended in various ways and is a very versatile tool to investigate clustered data.[134] Center-level variables can be included in the model and may partly explain the differences between centers and reduce the variance of the random intercepts. Heterogeneity in the effects of the X_{kij} can be investigated by adding random slopes. Cross-level interactions between patient and center characteristics can be included to reduce the variance of the random slopes. The variance structure in the data can be investigated in depth. For example, one may allow covariance between random terms (e.g., between random slopes and random intercepts). Moreover, one may explicitly model heteroscedasticity at the patient and the center level by allowing the variance terms to depend, linearly or quadratically, on explanatory variables. In addition, the two-level model can also be extended to a higher-level model.

The presence of center-specific parameters and the non-linear link function complicate parameter estimation in the mixed effects model and requires alternative estimation methods to approximate the maximum likelihood estimates. The best and most frequently used are the Laplace approximation and adaptive numerical quadrature.[134] Alternatively, a Bayesian approach can be adopted.[145] Parameter estimates may differ, depending on the chosen estimation technique.

To obtain reliable parameter estimates, the sample size needs to be sufficient, and in a random effects model, the number of clusters, that is, centers, appears to be most important. It is comforting that the fixed effects parameters in the model are reliable even with small sample sizes and only a few clusters (five or more), and the standard errors of patient-level fixed effects are accurate.[141, 146] However, to obtain correct standard errors for center-level fixed effects and correct estimates of the random effect variance, a moderately large number of clusters is required.[146, 147] A rough guideline is to collect data of at least thirty observations per center, from at least thirty centers. Random effect variances tend to be underestimated if there are too few clusters,[146] although severe overestimation may occur when the cluster sizes are very small.[147] To obtain accurate estimates of the standard errors of random effect variances, very large sample sizes may be needed.[146, 147] Note that mixed effects models can handle unbalanced data, with the number of patients per hospitals varying widely.[53]

Mixed effects models assume that random effects are normally distributed, but they are found to be quite robust against misspecifications of the random

effects distribution.[148-150] Alternative techniques that deal with non-normal random effects distributions exist, for example latent class models and Bayesian models.[134]

The conditional linear predictor for the ith individual from the jth center includes the random intercept for the jth center:[151]

(10)

$$LP_{MLR\ c\ ij}=\hat{\alpha}+\hat{a}_j+\sum_{k=1}^{K}\hat{\beta}_k X_{kij}\ .$$

The random intercepts $\hat{a}_j$ can be estimated using Empirical Bayes estimation. Through the combination of population-level and center-level information, posterior intercepts are obtained, which are shrunken towards zero. The degree of shrinkage is higher if less center-level information is available (i.e., stronger shrinkage for small centers) and if the between-center variance τ^2 is lower (i.e., the same degree of shrinkage for all centers).[134] Conditional predicted probabilities $\hat{p}_{MLR\ c\ ij}$ are obtained by taking the inverse logit of the conditional linear predictor.

To obtain a prediction for a patient of a center not included in the development set, the random intercept can be replaced by the average random intercept, ($\hat{a}_j = 0$).[151, 152]

(11)

$$LP_{MLR\ a\ ij}=\hat{\alpha}+0+\sum_{k=1}^{K}\hat{\beta}_k X_{kij}\ .$$

Predicted probabilities assuming an average center effect (0), $\hat{p}_{MLR\ a\ ij}$, are obtained by taking the inverse logit of this linear predictor. This will yield the conditional prediction for an individual from a center with an average intercept. Due to the nonlinearity of the logit transformation, this does not correspond to the average but to the median probability of patients with the same observed values of predictors across centers.

Although the mixed effects logistic regression model is a center-specific model, marginal predictions can be obtained by integrating over the distribution of the random effects:[151]

(12)

$$\hat{p}_{MLR\ m\ ij} = \int_{-\infty}^{\infty} \frac{1}{1+\exp(-LP_{MLR\ cond\ ij})} f(\hat{a}_j) d\hat{a}_j$$

$$\hat{p}_{MLR\ m\ ij} = \int_{-\infty}^{\infty} \frac{1}{1+\exp[-(\hat{\alpha}+\hat{a}_j+\sum_{k=1}^{K}\hat{\beta}_k X_{kij})]} f(\hat{a}_j)\, d\hat{a}_j,$$

where $f(\hat{a}_j)$ is the density function of a normal distribution with mean zero and variance $\hat{\tau}^2$. In practice, the integral can often not be solved analytically and must be evaluated by numerical averaging by sampling a large number of random effects from their fitted distribution. The marginalized linear predictor of the mixed effects model $LP_{MLR\ m\ ij}$ can be obtained by performing a logit transformation on the marginal predicted probabilities.[153] These predictions are very similar to the marginal predictions obtained by the standard logistic regression model.

Model updating

A center-specific model can also be obtained by updating an existing risk model, using new data from the center of interest. Several updating techniques exist, [98,99,100] of which recalibration is the simplest. In this case, a logistic regression model is fitted in the new center, with the linear predictor (LP) of the existing model as a predictor:

(13)

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_{new} + \beta_{overall} LP_i.$$

To adjust only the model intercept (i.e., calibration-in-the-large), the LP can be included as an offset variable.

Model revision techniques allow for a different regression slope in the new center. One either tests in a stepwise forward manner if predictors have a significantly different effect in the new center, after recalibration of the model, or one completely re-estimates the regression coefficients in the new sample. These updating techniques can be further elaborated by considering the inclusion of additional predictors. The extent of the updating should depend on the available sample size: model revision should only be attempted in large samples, possibly in combination with shrinkage.[100]

Updating techniques for mixed effects models have been proposed but further research is required.[154]

3.2.3 Correcting for clustering

Several statistical techniques correct for clustering and provide better inferences than simply ignoring the clustered data structure. These methods are less useful for prediction purposes, as they do not yield center-specific predictions. Conditional logistic regression, like the center-specific methods discussed above, yields regression coefficients with conditional interpretations. Other methods, like generalized estimating equations, the use of cluster-robust standard errors, survey methods and bootstrap methods, yield parameters with marginal interpretations.

Conditional logistic regression is conditioned on sufficient statistics for the center-specific effects.[136] Hence, it can only estimate patient-level predictors; predictors at the center level are removed as "nuisance". For this model to work well, there should be a moderately large number of patients per center.[137]

Generalized estimating equations (GEE) can be used to estimate a marginal logistic regression model, while taking into account the clustered data structure for the estimation of standard errors.[139] Hence, correct inferences can be made but predictions are not center-specific. GEE allows the user to specify the correlation structure between observations. It is natural to assume that outcomes are correlated for patients from the same center and independent for patients from different centers. This assumption of an "exchangeable" correlation structure also underlies random effects models, which can give biased results if the assumption is violated. Luckily, in practice, this assumption is reasonable in multicenter studies.

Robust sandwich estimators for the standard error can be used if model assumptions are incorrect and are suited for any type of regression model. For example, robust standard errors can be used if the correlation structure is misspecified with GEE, when analyzing clustered data with a standard logistic regression model, or when analyzing data with a random effects model when the random effects distribution is non-normal. However, they require a large sample, which translates to a very large number of clusters in the case of clustered data. [140, 141] GEE with robust standard errors can lead to standard errors that are too low if the number of centers is not large, whereas GEE with non-robust standard errors performs well for multicenter studies, even with a small number of centers (e.g., five).[140, 141] Moreover, the quality of cluster-robust standard errors suffers when the clusters are very different in terms of

size and distributions of predictors.[134] However, with a large number of clusters and a fairly homogeneous population of clusters, a standard logistic regression with cluster-robust standard errors is a good alternative to GEE or random effects logistic regression.[26]

Survey methods are commonly used to correct for bias in a two-stage sampling design when computing averages and proportions in large surveys, but they can also be used to estimate parameters from regression models.[137] Typically, survey weights are constructed in relation to the probability of sampling for each hospital and each patient within the hospital, to take into account that not every patient has the same sampling probability. Hence, they can be used in multicenter studies that use two-stage random sampling from a finite sample, but they do not offer center-specific predictions.

Bootstrap methods can be used to construct confidence intervals for clustered data.[137, 155] By randomly sampling centers from the real dataset, the correlation structure is preserved in each bootstrap sample. Parameters of interest can be calculated in each bootstrap sample and bootstrap confidence intervals can be used to account for the effect of clustering.

3.2.4 Combining within-cluster results

When the relationship between a predictor and an outcome has been studied in various centers, a meta-analysis can be performed to combine the published results into a pooled estimate.[156-158] If it can be assumed that all studies estimate the same underlying true parameter (i.e., the studies investigate the same homogeneous population, use the same predictor and outcome definition, etc.), a fixed effects meta-analysis is suitable. If this assumption does not hold, for example because of differences between study populations, a random effects meta-analysis is preferable.

Methods have been proposed to combine pooled univariable or multivariable regression coefficients from the literature with multivariable regression coefficients obtained from individual patient data from a new study population.[159-161] These can then be combined in a risk model that uses the external information to enhance generalizability. Additionally, techniques have been proposed that combine existing risk models and simultaneously update the resulting model for a new setting for which IPD data is available.[162]

If individual patient data (IPD) is available from multiple studies, an IPD meta-analysis can be performed with the techniques proposed above. The data can be analyzed by ignoring the clustered data structure (“stacking”), using a

fixed effects model ("stratified estimation of the intercept"), or using a random effects model.[142]

3.2.5 Discussion

Several strategies have been proposed to deal with clustered data. Simply ignoring the clustered data structure yields incorrect inferences for the parameters of the regression model. Several alternative techniques do provide correct inferences. Among these, a distinction should be made between methods that provide marginal or population-averaged effects estimates, like the standard logistic regression model, and models that provide conditional effect estimates. GEE is an example of the first class, while conditional logistic regression is an example of the second class.

However, all these methods miss out on the opportunity to employ center-specific information to provide risk estimates for patients that are tailored to the centers they belong to. Center-specific methods provide correct inferences and center-specific predictions. Two center-specific methods suited to analyze multicenter datasets are fixed effects regression and mixed effects regression. Fixed effects models should only be used when the number of centers is low and the number of patients per center is large. Mixed effects models provide reliable estimates for fixed effects even for low sample sizes, and the shrinkage in Empirical Bayes estimates of cluster-specific intercepts and slopes prevents overfitting for small centers. Moreover, the philosophy underlying the mixed effects model puts a strong emphasis on the generalizability of the results to other centers, which is also key in prediction research.

Updating techniques also provide center-specific predictions, but can only be used to adjust an existing risk model to the specific characteristics of a new center. Finally, a meta-analysis can be performed to combine the results of various single-center studies.

3.3 Multicenter data from the International Ovarian Tumor Analysis group

The International Ovarian Tumor Analysis (IOTA) group is an example of a multicenter collaboration collecting clustered data. These data will be used throughout the thesis as a case study. This section discusses the heterogeneity

in ovarian cancer prevalence and how this was accounted for in the development of the Simple Rules Risk scoring system.[163]

3.3.1 Background

A tumor is a mass characterized by an abnormal growth of cells. Epithelial cells, which form the surface of the ovary, are the origin of the majority of ovarian tumors, while germ cell, sex cord and stromal tumors are much rarer.[164] Tumors can be benign or malignant, that is, cancerous. The malignant ovarian tumors are grouped into primary invasive tumors, which originate in the ovaries, borderline tumors, which are a subclass of primary invasive tumors with a more favorable outcome, and metastatic tumors, which have their origin elsewhere (e.g., breast) but have spread to the ovary. The International Federation of Obstetrics and Gynecology (FIGO) classifies ovarian cancer into four stages, without distinguishing between primary invasive and borderline tumors. In stage I the cancer is confined to the ovaries; in stage II the disease extends to other organs in the pelvis or the pelvic peritoneum; in stage III the disease has spread to the abdominal peritoneum and/or the lymph nodes; in stage IV there are distant metastases of the ovarian tumor.[165]

Approximately 1 in 52 women will be diagnosed with ovarian cancer during their lifetime.[166, 167] Worldwide, nearly 239,000 women have been diagnosed with ovarian cancer in 2012, with incidence rates varying across the world.[166] In Belgium, 766 women were diagnosed with ovarian cancer in 2013, making it the eight most common cancer among women.[167] The incidence increases with age, the average age at diagnosis being 66.[167]

When diagnosed at stage I, 90% of women with ovarian cancer will survive their disease for five years or more, compared to less than 5% of women when diagnosed at the stage IV.[168] Because ovarian cancer is mostly diagnosed at a late stage,[167] ovarian cancer is the most common cause of gynecological cancer death, responsible for approximately 150,000 deaths worldwide each year.[166] Unfortunately, there is no conclusive evidence that population-wide screening programs reduce mortality.[169] Guidelines suggest that women with an increased level of serum CA125 or symptoms suggestive of ovarian cancer (e.g., pelvic or abdominal pain, urinary frequency or urgency, increased abdominal size or bloating, difficulty eating or feeling full quickly) should be referred for an ultrasound scan.[170]

Ultrasound is the best technique to refer patients to the appropriate care pathway based on a presurgical diagnosis.[171] Patients with benign-looking

masses may be managed conservatively or treated with minimally invasive surgery to avoid patient morbidity, costs and loss of fertility. In contrast, when a mass looks malignant, specialized care should be taken to avoid tumor rupture during debulking surgery,[172] and supplementary investigations should be done to determine whether further treatment (e.g., chemotherapy) is required. Patient outcomes are better when cancer is treated by a gynecological oncologist than when it is treated by a general surgeon or a general gynecologist,[173] and when treated in a high-volume or specialized hospital.[174] However, the majority of patients do not receive treatment from a specialist.[174]

Subjective assessment of ultrasound images is the best method to distinguish between benign and malignant tumors before surgery, but the quality of the assessment seems to depend on the experience of the clinician.[175] In addition, the required expertise is not available in every center. This has inspired researchers to develop risk models and classification rules, in order to allow a more objective differentiation between masses. The early studies did not adhere to the methodological strategies discussed in chapter 2 and suffered from serious drawbacks such as low sample sizes, single center data and non-standardized measurements.[176]

The Risk of Malignancy Index (RMI)[177] remains the most used and recommended way of classifying ovarian tumors,[170] although evidence suggests that better models are available nowadays.[15] RMI is computed as U×M×CA125. U is an ultrasound score, considering the following features: the presence of a multilocular cyst, solid areas, evidence of metastases, ascites and bilateral lesions. The ultrasound score is 0 if none of the features is present, 1 if one is present and 3 if more than one is present. M is 1 for premenopausal patients and 3 for postmenopausal patients. Serum level CA125 is expressed in U/ml. A cut-off of 200 is often used with RMI to indicate malignancy.

3.3.2 The IOTA dataset

The International Ovarian Tumor Analysis (IOTA) consortium was established in 1999 to organize the collection of a large multicenter dataset in a standardized way and overcome the issues earlier studies suffered from. To this end, IOTA published a consensus opinion on definitions and measurements of sonographic characteristics of adnexal tumors.[178] The IOTA dataset includes clinical and ultrasound information on non-pregnant with at least one persistent adnexal mass. All women underwent surgery within 120

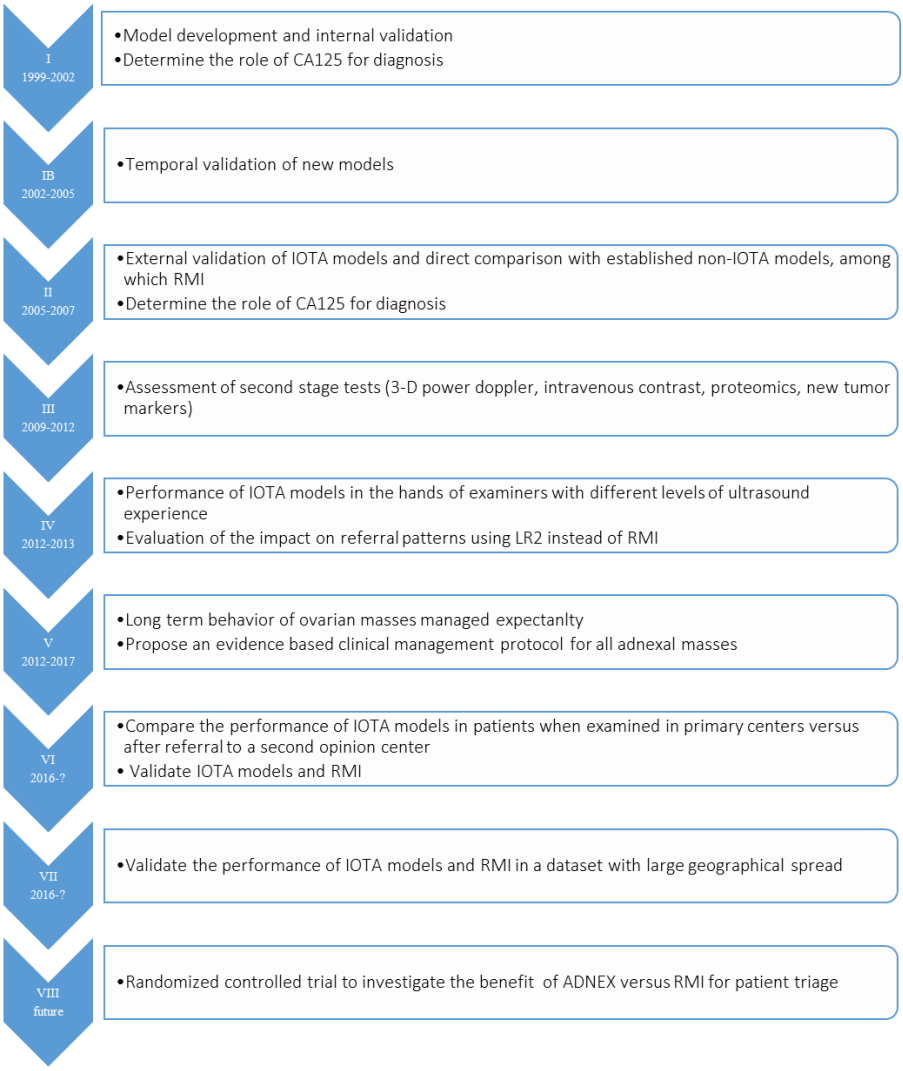


Figure 9. Objectives of the different IOTA phases.

Table 7. Patient, tumor and center characteristics for different IOTA phases.

Phase	Number of patients	Number of centers (oncology centers)	Number of countries	Patient age, mean (sd)	Benign	Borderline	Stage I	Stage II-IV	Metastatic
I	1066	9 (4)	5	47 (16)	800 (75%)	55 (5%)	42 (4%)	127 (12%)	42 (4%)
Ib	507	3 (2)	3	47 (17)	364 (72%)	20 (4%)	24 (5%)	79 (16%)	20 (4%)
II	1938	19 (11)	9	46 (16)	1396 (72%)	111 (6%)	70 (4%)	303 (16%)	58 (3%)
III	2403	18 (11)	6	50 (16)	1423 (59%)	153 (6%)	128 (5%)	573 (24%)	126 (5%)
IVb	255	3 (2)	1	46 (17)	181 (71%)	18 (7%)	23 (9%)	26 (10%)	7 (3%)

Table 8. Number of patients contributed by each center to the IOTA database in phases I, Ib, II, III and IVb.

Code	City	Country	Type	I	Ib	II	III	IVb	Total
LBE	Leuven	Belgium	Oncology referral center	263	287	252	129		931
RIT	Roma	Italy	Oncology referral center	126	96	122	443		787
MSW	Malmo	Sweden	General hospital	315	124	137	201		777
GBE	Genk	Belgium	General hospital			200	228		428
OIT	Monza	Italy	Oncology referral center	46		251	105		402
PCR	Prague	Czech Republic	Oncology referral center			90	264		354
CIT	Milano	Italy	Oncology referral center			94	218		312
LPO	Lublin	Poland	Oncology referral center			154	131		285
SIT	Sardinia	Italy	General hospital			154	107		261
MIT	Milano	Italy	Referral center, non-oncology	87		50	86		223
BIT ph 3	Bologna	Italy	Referral center, non-oncology				213		213
QCCH	London	UK	Oncology referral center					193	193
BIT ph 2	Bologna	Italy	Oncology referral center			135			135
SSW	Stockholm	Sweden	Oncology referral center				120		120
KUK	London	UK	Oncology referral center	54		65			119
NIT	Napoli	Italy	Referral center, non-oncology	31		64	8		103
BFR	Paris	France	General hospital	80					80
LSW	Lund	Sweden	Oncology referral center			38	39		77
BCH	Beijing	China	Oncology referral center			73			73
MFR	Maurepas	France	General hospital	64					64
UDI	Udinese	Italy	Oncology referral center			17	47		64
BSP	Barcelona	Spain	Referral center, non-oncology				37		37
SHH	Southampton	UK	Oncology referral center					33	33
WMUH	London	UK	General hospital					29	29
FIT	Florence	Italy	General hospital				21		21
VIT	Milano	Italy	General hospital			21			21
GIT	Napoli	Italy	Oncology referral center			9	6		15
OCA	Ontario	Canada	General hospital			12			12

days after a transvaginal ultrasound scan had been performed. Histopathologic examination of the excised tissue was performed in the local centers without knowledge of the ultrasound results. If more than one mass was present, only the mass that appeared most complex on ultrasound was included. Data collection occurred with dedicated software, which assured that all the questions were completed. Only serum level CA125 was an optional measurement, and multiple imputation was used to handle missing values (see Appendix 1).[28, 179] The quality of the data was ensured by built-in checks and manual reviews by clinicians and statisticians.

IOTA data collection has occurred in several phases since 1999 and is still ongoing (see Figure 9). Each phase was designed for a specific primary objective, although some classification and risk models were developed in secondary analyses of the data. Each phase was also used for the validation of established non-IOTA models or previously developed IOTA models. The study protocols for were approved by the Ethics Committee of the University Hospitals Leuven ('Commissie Medische Ethiek') and by the local Ethics Committee at each recruitment center. Throughout this thesis, data from IOTA phases I to III are used. In chapter 6, a subset of data from phase IV, referred to as IVb, is used. Together, these phases contain information of 6169 women with 2005 (33%) malignant tumors (Table 7).

In total, 28 centers recruited data to phases I, Ib, II, III and IVb. The number of patients recruited per center varied between 12 and 931 (median 120, IQR 51 to 266). Centers did not recruit patients in all phases (Table 8). The prevalence of malignant tumors in IOTA phases I, Ib, II, III and IVb combined is 32%. The prevalence of malignancy is highly heterogeneous across centers, the lowest observed being 0% and the highest 66% (median 24%, interquartile range (IQR) 13% to 35%). Part of the heterogeneity in prevalence is explained by the nature of the centers (see Figure 10). 15 centers are oncology centers, that is, tertiary referral centers with a specific gynecologic oncology unit, while the other 13 centers are general hospitals or hospitals with a gynecologic ultrasound unit not linked to an oncology unit. For example, BIT was a referral center with a gynecologic ultrasound unit in phase II, but became specialized in gynecologic oncology in phase III. The prevalence in this center increased from 8% to 31%, because the patient population changed as the center changed its orientation.

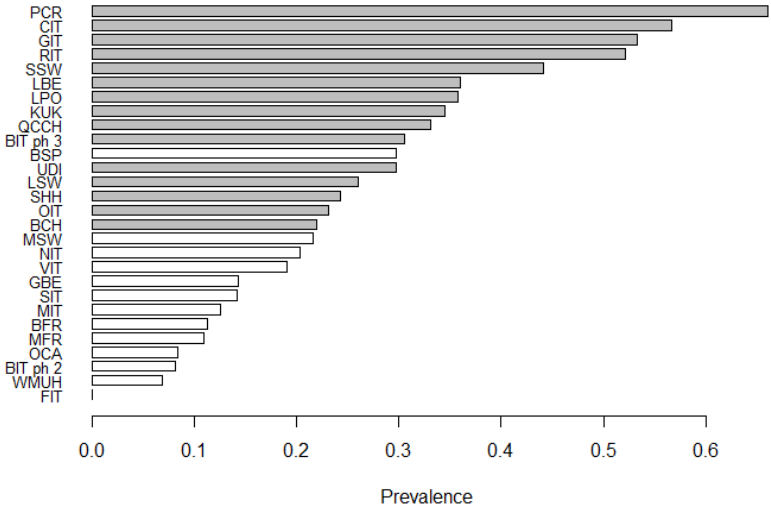


Figure 10. The prevalence of malignant tumors per center, all phases combined. Grey bars indicate oncology centers; white bars indicate non-oncology centers.

3.3.3 IOTA models and classification rules

Several IOTA models have been developed over time. By extensively validating them and comparing them to other models, IOTA investigators have attempted to identify the best strategy to manage patients with adnexal masses.[176] This section discusses established IOTA models and classification rules that will be referred to in later chapters.

The LR2 model is a logistic regression model developed on a training dataset of 754 patients from IOTA phase I, using age, the presence of ascites, the presence of blood flow within a solid papillary projection (papflow), the maximal diameter of the largest solid component (soldmax, in millimeter and truncated at 50 mm), the presence of irregular cyst walls (wallreg) and the presence of acoustic shadows (shadows) as predictors for the probability of tumor malignancy (p_{ij}):[7]

(14)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = -5.3718 + 0.0354 \times age_{ij} + 1.6159 \times ascites_{ij} + 1.1768 \times papflow_{ij} + 0.0697 \times soldmax_{ij} + 0.9586 \times wallreg_{ij} - 2.9486 \times shadows_{ij}.$$

It was internally validated on the remaining 312 patients from phase I,[7] temporally validated in phases Ib,[180] II,[41, 181, 182] III,[183] validated in new centers in phases II,[41, 181, 182] III[183] and IVb,[184] and validated externally by other researchers.[185, 186] LR2 has a good discriminatory performance, maintaining c-statistics >.9 at validation. Using a probability threshold of 10% to classify tumors as malignant yields a sensitivity of 93% and a specificity of 84%, according to a recent meta-analysis.[187] More complex mathematical models developed on the IOTA I dataset, such as artificial neural networks and support vector machines, showed no benefit over the simpler logistic regression model.[41, 180] Studies have shown that LR2 outperforms RMI[15, 187, 188] (see section 3.3.1) and ROMA,[189] which is a logistic regression model combining two biomarkers, with -12.0 + 2.38 × log(HE4) + 0.0626 × log(CA125) the linear predictor for premenopausal patients and -8.09 + 1.04 × log(HE4) + 0.732 × log(CA125) the linear predictor for postmenopausal patients.[190]

The LR1 model was developed and validated simultaneously with the LR2 model. It is comparable to LR2 but includes additional predictors:[7] the personal history of ovarian cancer (pershistovca), the current use of hormonal therapy (hormtherapy), the maximum diameter of the lesion (lesdmax), the experience of pain during the examination (pain), the presence of entirely solid tumors (solid) and the color score of intratumoral blood flow (colscore, ordinal score from 1, indicating no blood flow, to 4, indicating a very strong blood flow):

(15)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = -6.7468 + 1.5985 \times pershistovca_{ij} - 0.9983 \times hormtherapy_{ij} + 0.0326 \times age_{ij} + 0.00841 \times lesdmax_{ij} - 0.8577 \times pain_{ij} + 1.5513 \times ascites_{ij} + 1.1737 \times papflow_{ij} + 0.9281 \times solid_{ij} + 0.0496 \times soldmax_{ij} + 1.1421 \times wallreg_{ij} - 2.3550 \times shadows_{ij} + 0.4916 \times colscore_{ij}.$$

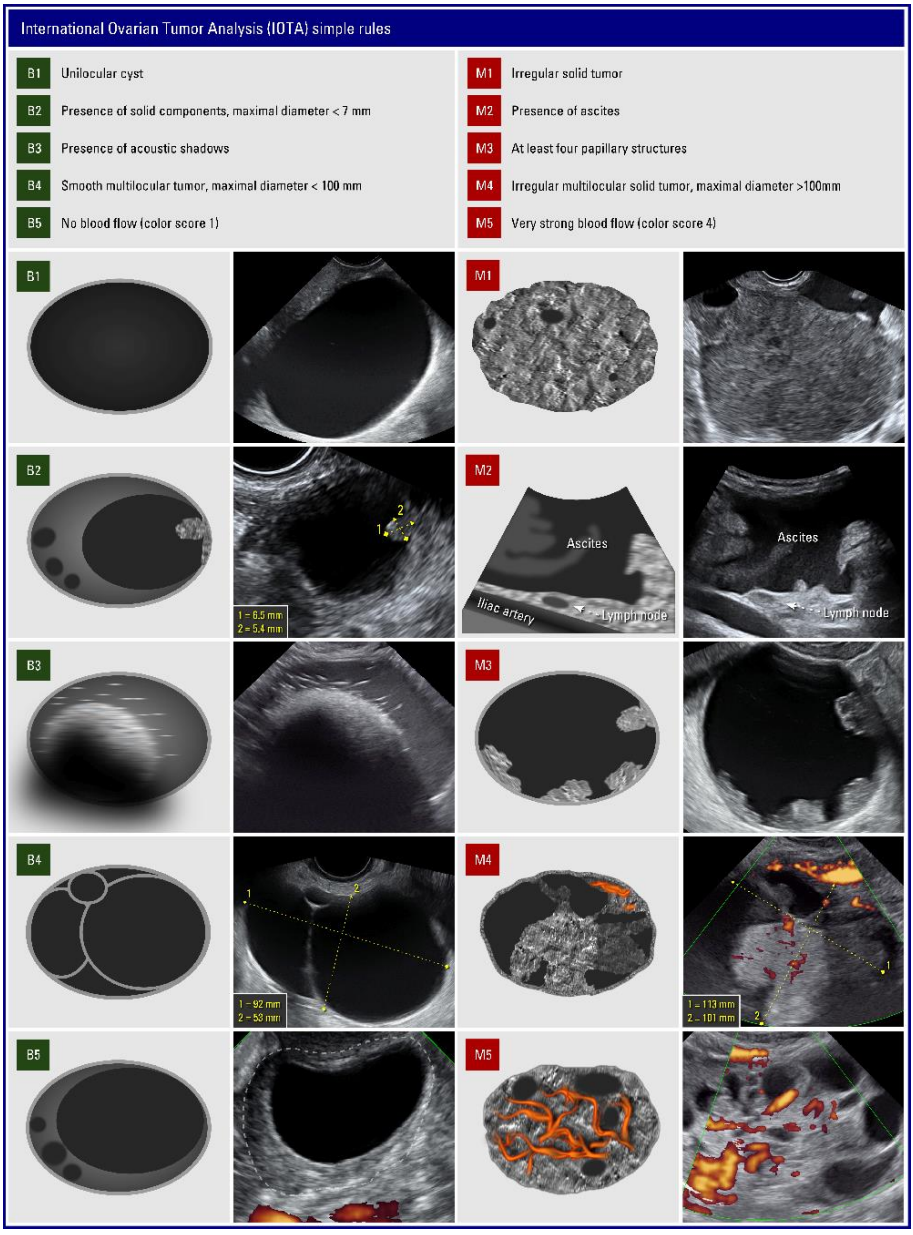


Figure 11. IOTA Simple Rules. Reproduced from http://www.iotagroup.org/index.php/educational-material.

LR1 had a discriminatory performance similar to that of LR2. [7, 41, 180, 181, 183, 185]

The IOTA Simple Rules are based on a number of clearly defined ultrasound features and have been developed on phase I data to guide ultrasound examiners without requiring a computer to perform complex calculations.[191] There are five M-features and five B-features (Figure 11). If at least one M-feature and none of the B-features are present, the mass is classified as malignant. If at least one B feature and none of the M-features are present, the mass is classified as benign. Otherwise, the Simple Rules are inconclusive and another diagnostic method should be used (such as subjective assessment by an expert). The Simple Rules were validated temporally and in new centers by the IOTA group,[183, 184, 192, 193] and by other research groups.[194-204] The Simple Rules were applicable in 76 to 89% of patients in the studies identified in a meta-analysis by Nunes and colleagues. They found that the sensitivity for malignancy is about 93% while the specificity is about 95% when the Simple

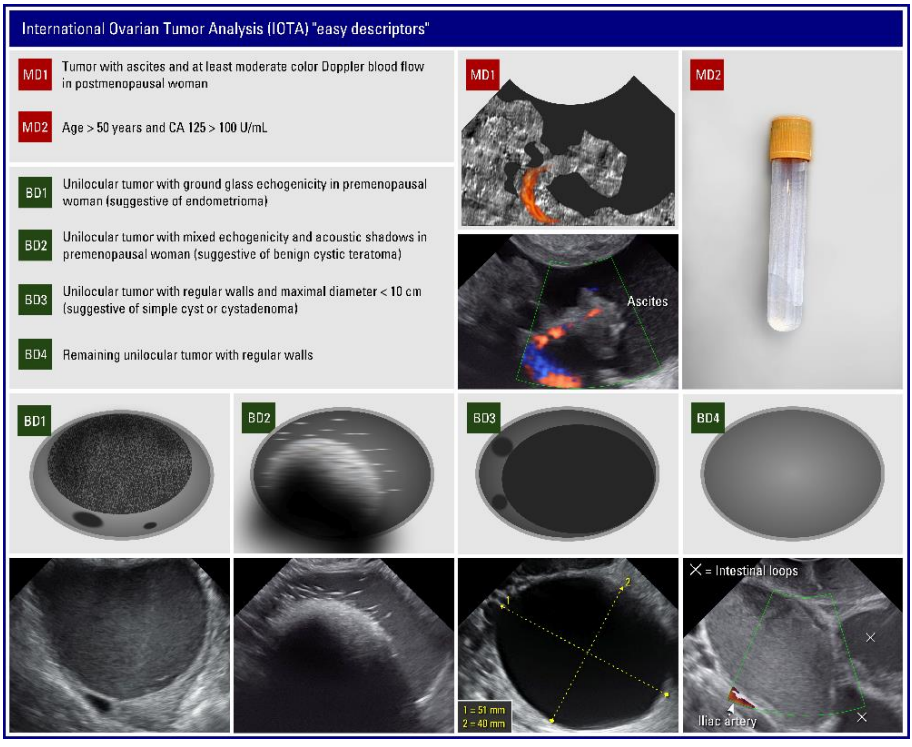


Figure 12. IOTA Easy Descriptors. Reproduced from http://www.iotagroup.org/index.php/educational-material.

Rules are applicable.[194] In a more recent meta-analysis by Meys and colleagues,[187] the Simple Rules, classifying inconclusive cases as malignant, had a sensitivity of 93% and a specificity of 80%. A strategy including the Simple Rules and the referral of inconclusive cases to experts for subjective assessment yielded a sensitivity of 91% and a specificity of 91%. The IOTA Simple Rules are better able to detect cancer than RMI, which has a sensitivity of 75% and a specificity of 92% at cut-off 200.[15, 187]

The IOTA Easy Descriptors were formulated to facilitate an instant classification for masses with a very recognizable ultrasound appearance (Figure 12). In the original study, using data from phases I, Ib and II retrospectively, 43% of masses could be instantly classified with a sensitivity of 98% and a specificity of 97%.[205] In validation studies by the IOTA group [183, 193, 198] and others [196] the Easy Descriptors were applicable in 42 to 67% of the masses, with sensitivities between 93% and 99.5% and specificities between 69% and 97%.

The ADNEX model enables a more elaborate classification of tumors, as it distinguishes between benign tumors and four classes of malignant tumors, namely borderline, stage I primary invasive, stage II-IV primary invasive and secondary metastatic ovarian tumors.[8] ADNEX was fitted on data from IOTA phases I, Ib and II as a multinomial logistic regression model with random center intercepts, and refitted on data from phases I, Ib, II and III after external validation:

(16)

$$\log\left(\frac{p_{ij\ \text{borderline}}}{p_{ij\ \text{benign}}}\right) = -7.5777 + \hat{a}_{j\ \text{borderline}} + 0.0045 \times age_{ij} + 0.1116 \times \log2\,(CA125_{ij}) + 0.3720 \times \log2\,(lesdmax_{ij}) + 6.9679 \times \left(\frac{soldmax_{ij}}{lesdmax_{ij}}\right) - 5.6559 \times \left(\frac{soldmax_{ij}}{lesdmax_{ij}}\right)^2 + 1.3751 \times loc10_{ij} + 0.6042 \times papnr_{ij} - 2.0416 \times shadows_{ij} + 0.9711 \times ascites_{ij} + 0.9530 \times oncocenter_{ij}$$

$$\log\left(\frac{p_{ij\ \text{stage I}}}{p_{ij\ \text{benign}}}\right) = -12.2760 + \hat{a}_{j\ \text{stage I}} + 0.0173 \times age_{ij} + 0.1972 \times \log2\,(CA125_{ij}) + 0.8735 \times \log2\,(lesdmax_{ij}) + 9.5831 \times \left(\frac{soldmax_{ij}}{lesdmax_{ij}}\right) - 5.8332 \times \left(\frac{soldmax_{ij}}{lesdmax_{ij}}\right)^2 + 0.7919 \times loc10_{ij} + 0.4004 \times papnr_{ij} - 1.8776 \times shadows_{ij} + 0.4527 \times ascites_{ij} + 0.4525 \times oncocenter_{ij}$$

$$\log\left(\frac{p_{ij\ \text{stage II-IV}}}{p_{ij\ \text{benign}}}\right) = -14.9158 + \hat{a}_{j\ \text{stage II-IV}} + 0.0512 \times age_{ij} + 0.7655 \times \log2\,(CA125_{ij}) + 0.4305 \times \log2\,(lesdmax_{ij}) + 10.3770 \times \left(\frac{soldmax_{ij}}{lesdmax_{ij}}\right) - 5.7098 \times \left(\frac{soldmax_{ij}}{lesdmax_{ij}}\right)^2 + 0.3737 \times loc10_{ij} + 0.3899 \times papnr_{ij} - 2.3552 \times shadows_{ij} + 1.3484 \times ascites_{ij} + 0.4590 \times oncocenter_{ij}$$

$$\log\left(\frac{p_{ij\ \text{metastatic}}}{p_{ij\ \text{benign}}}\right) = -11.9093 + \hat{a}_{j\ \text{metastatic}} + 0.0336 \times age_{ij} + 0.2762 \times \log2\,(CA125_{ij}) + 0.4490 \times \log2\,(lesdmax_{ij}) + 6.6449 \times \left(\frac{soldmax_{ij}}{lesdmax_{ij}}\right) - 2.3033 \times \left(\frac{soldmax_{ij}}{lesdmax_{ij}}\right)^2 + 0.9000 \times loc10_{ij} + 0.2156 \times papnr_{ij} - 2.4984 \times shadows_{ij} + 1.6364 \times ascites_{ij} + 0.8089 \times oncocenter_{ij}$$

$$\begin{bmatrix} \hat{a}_{j\ \text{borderline}} \\ \hat{a}_{j\ \text{stage I}} \\ \hat{a}_{j\ \text{stageII-IV}} \\ \hat{a}_{j\ \text{metastatic}} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{\tau}^2_{\text{borderline}} & 0 & 0 & 0 \\ 0 & \hat{\tau}^2_{\text{stage 1}} & 0 & 0 \\ 0 & 0 & \hat{\tau}^2_{\text{stage II-IV}} & 0 \\ 0 & 0 & 0 & \hat{\tau}^2_{\text{metastatic}} \end{bmatrix} \right)$$

with loc10 indicating whether the number of cyst locules is larger than 10, papnr the number of papillary structures (0, 1, 2, 3, 4, with 4 indicating more than 3), soldmax the maximal diameter of the largest solid component (in mm, not truncated), oncocenter an indicator with value 1 if the center is a specialized oncology referral center and 0 otherwise, and other variables defined as above. The inclusion of center type as a variable adjusts risks to reflect differences in patient populations that are hard to capture with patient and tumor characteristics. Validation of ADNEX on IOTA III data yielded a c-statistic of 0.94 to distinguish between benign and malignant tumors. Pairwise c-statistics between subtypes of tumors varied between 0.71 and 0.99.[8] Moreover, the model is well calibrated, whereas the calibration of earlier IOTA risk models LR1 and LR2 is suboptimal. The model maintained good results at external validation by other researchers, yielding AUCs between 0.91 and 0.96 to distinguish between benign and malignant tumors,[206-210] and an AUC of 0.90 to distinguish between secondary metastatic ovarian tumors and malignant cancer originating in the ovaries.[211]

3.3.4 The Simple Rules risk scoring system

Of all IOTA models and classification rules, the IOTA Simple Rules were perhaps best received by clinicians, but they are not able to classify all adnexal

Table 9. Estimated regression coefficients, random intercept variance, residual intraclass correlation and proportion of explained variance for regression models based on the Simple Rules ultrasound features fitted on phase Ib and II data (n=2445).

Effect	Empty RI	US	US + center type	US + RI	US + center type + RI
Intercept	-1.07 (0.19)	-0.53 (0.09)	-1.23 (0.15)	-0.50 (0.20)	-1.10 (0.26)
b1		-3.27 (0.33)	-3.14 (0.33)	-3.12 (0.34)	-3.10 (0.34)
b2		-1.81 (0.57)	-1.65 (0.59)	-1.58 (0.59)	-1.55 (0.59)
b3		-1.56 (0.26)	-1.55 (0.26)	-1.61 (0.27)	-1.58 (0.27)
b4		-3.53 (0.59)	-3.46 (0.60)	-3.60 (0.60)	-3.59 (0.60)
b5		-1.72 (0.23)	-1.83 (0.23)	-1.97 (0.24)	-1.96 (0.24)
m1		2.35 (0.37)	2.36 (0.37)	2.37 (0.40)	2.38 (0.39)
m2		2.75 (0.28)	2.70 (0.28)	2.91 (0.29)	2.87 (0.29)
m3		1.65 (0.27)	1.66 (0.27)	1.73 (0.28)	1.72 (0.28)
m4		1.06 (0.22)	1.10 (0.22)	1.11 (0.23)	1.12 (0.23)
m5		1.42 (0.23)	1.43 (0.23)	1.54 (0.24)	1.53 (0.24)
oncocenter			1.00 (0.16)		0.95 (0.31)
τ^2	0.59 (0.23)			0.50 (0.14)	0.26 (0.14)
RICC	15.15%			13.09%	7.24%
R^2	15.15%	75.46%	76.74%	77.38%	78.04%

b1: unilocular cyst; b2: solid components present, but < 7 mm; b3: acoustic shadows; b4: smooth multilocular tumor with largest diameter < 100 mm; b5: no blood flow, color score 1; m1: irregular solid tumor; m2: ascites; m3: at least 4 papillary structures; m4: irregular multilocular-solid tumor with largest diameter ≥ 100 mm; m5: very strong flow; color score 4; RI: random intercept; US: ultrasound features; τ^2: random intercept variance; RICC: residual intraclass correlation; R^2: proportion of explained variance. Effect estimate (standard error).

Table 10. Updated model coefficients using all data (n=4848).

Effect	Coefficient (standard error)
Intercept	-0.97 (0.24)
b1 (unilocular cyst)	-3.41 (0.27)
b2 (solid components present, but < 7 mm)	-2.25 (0.46)
b3 (acoustic shadows)	-1.66 (0.18)
b4 (smooth multilocular tumor with largest diameter < 100 mm)	-2.75 (0.27)
b5 (no blood flow; color score 1)	-1.86 (0.17)
m1 (irregular solid tumor)	2.19 (0.24)
m2 (ascites)	2.65 (0.21)
m3 (at least 4 papillary structures)	1.53 (0.20)
m4 (irregular multilocular-solid tumor with largest diameter ≥ 100 mm)	0.98 (0.16)
m5 (very strong flow; color score 4)	1.55 (0.16)
Oncology center	0.92 (0.27)

masses and they do not provide risk estimates. The Simple Rules risk scoring system was recently developed to overcome these shortcomings.[163]

The model was developed using data from phases Ib and II. A mixed effects logistic regression model was used to predict the risk of malignancy of an adnexal mass. The predictors are the ten ultrasound features from the Simple Rules. Interactions and nonlinear effects were not considered. It was a priori decided to include a random center intercept and an indicator variable for center type in the risk model. Table 9 presents various alternative model formulations (fitted in phase Ib and II data) for illustrative purposes.

The between-center heterogeneity in the development dataset was moderately large: $\hat{\tau}^2$ was 0.59 in an empty model with only a random intercept. 15.15% ($0.59/(0.59+\pi^2/3)$, with $\pi^2/3$ the random variance at the patient level) of the variance in the outcome is due to differences in malignancy rates between centers rather than variability at the patient level. After accounting for the ten ultrasound features of tumors, still 13.09% of the unexplained variance is due to differences between centers. A part of the differences between centers can be explained by taking into account the center type: the malignancy rate was 36% in oncology centers and 15% in all other centers. The proportion of

unexplained variance at the center level drops from 13.09% to 7.24% after the inclusion of an effect for center type. A model with ultrasound features, center type and a random center intercept explains most of the variance in the data. The R^2 is 78.04%, which is higher than the R^2 of a model with only ultrasound features (75.46%) or a model with ultrasound features and a random center intercept (76.74%).

Data from phase III was used for model validation. Predictions substituting random intercepts with 0 were used for model validation, to mimic the realistic situation where no center-specific intercepts are available. Advanced techniques beyond those discussed in chapter 2 were used to account for the multicenter nature of the validation data. Chapter 6 discusses this in more detail. For now, it suffices to say that the summary c-statistic based on random effects meta-analysis of center-specific results was 0.92 (95% CI 0.90-0.93), and plots showed good calibration in oncology and non-oncology centers. After validation proved satisfactory, the model was refitted using phase Ib, II

Table 11. Simple Rules feature combinations and the associated risk of malignancy (in %).

<table>
<tr><td>Oncology centers</td><td colspan="8">Number of M features</td></tr>
<tr><td>Number of B features</td><td>0</td><td>1 (M4)</td><td>1 (M3)</td><td>1 (M5)</td><td>1 (M1)</td><td>1 (M2)</td><td>2</td><td>>2</td></tr>
<tr><td>0</td><td>48.7</td><td>71.7</td><td>81.4</td><td>81.7</td><td>89.5</td><td>93.1</td><td>92.1-99.2</td><td>98.2-≥99.9</td></tr>
<tr><td>1 (B3)</td><td>15.2</td><td colspan="5" rowspan="4">Specific combinations are rare, consider suspicious
(risks estimated to be between 12.9 and 71.9% depending on which B- and M-feature)</td><td colspan="2" rowspan="4">Rare finding, consider highly suspicious</td></tr>
<tr><td>1 (B5)</td><td>12.8</td></tr>
<tr><td>1 (B2)</td><td>9.1</td></tr>
<tr><td>1 (B4)</td><td>5.7</td></tr>
<tr><td>1 (B1)</td><td>3.1</td><td colspan="5" rowspan="2">Rare finding, consider suspicious</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>2</td><td>0.49-2.7</td></tr>
<tr><td>>2</td><td>0.09-0.29</td><td colspan="5"></td><td colspan="2"></td></tr>
<tr><td>Other centers</td><td colspan="8">Number of M features</td></tr>
<tr><td>Number of B features</td><td>0</td><td>1 (M4)</td><td>1 (M3)</td><td>1 (M5)</td><td>1 (M1)</td><td>1 (M2)</td><td>2</td><td>>2</td></tr>
<tr><td>0</td><td>27.5</td><td>50.2</td><td>63.6</td><td>64</td><td>77.2</td><td>84.3</td><td>82.3-98.0</td><td>95.6-99.7</td></tr>
<tr><td>1 (B3)</td><td>6.7</td><td colspan="5" rowspan="4">Specific combinations are rare, consider suspicious
(risks estimated to be between 5.6 and 50.5% depending on which B- and M-feature)</td><td colspan="2" rowspan="4">Rare finding, consider highly suspicious</td></tr>
<tr><td>1 (B5)</td><td>5.6</td></tr>
<tr><td>1 (B2)</td><td>3.8</td></tr>
<tr><td>1 (B4)</td><td>2.4</td></tr>
<tr><td>1 (B1)</td><td>1.2</td><td colspan="5" rowspan="2">Rare finding, consider suspicious</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>2</td><td>0.19-1.1</td></tr>
<tr><td>>2</td><td>≤0.01-0.12</td><td colspan="5"></td><td colspan="2"></td></tr>
</table>

and III data to fully exploit all available information (Table 10). The risks are summarized in Table 11, which can be used in clinical practice if the model proves to have a good predictive performance in further external validation studies. For example, if no B-features and no M-features are present, and the patient was seen in an oncology center, the predicted risk of malignancy is 48.7%. If she was seen in a regional center or a referral center not specialized in gynecological oncology, the predicted risk of malignancy is 27.5%. The difference can be explained by complex referral patterns that cause heterogeneity in patient populations, which is hard to capture with patient-level characteristics.

3.4 Conclusion

Multicenter datasets are clustered: patients from within centers may be more similar than patients from different centers and independence between observations cannot be assumed. The International Ovarian Tumor Analysis (IOTA) dataset is an example of a large multicenter dataset. The IOTA group collects data on patients with suspicious ovarian masses in a standardized way. Many risk models and classification rules have been developed to support the diagnosis of ovarian cancer. Two recently developed IOTA models, ADNEX and the Simple Rules risk scoring system, have been developed using mixed effects logistic regression. A mixed effects model is able to yield center-specific predictions for centers included in the model development dataset and takes the clustering into account for the computation of standard errors. ADNEX and the Simple Rules risk scoring system include center type as a predictor to reflect that, due to complex referral patterns, the prevalence of ovarian cancer in specialized gynecological oncology centers is higher than in other centers.

4 Sample size for multicenter studies

This chapter investigates the number of events per variable (EPV) required to develop a reliable risk model. Section 4.1 recapitulates previous sample size research. Chapter 4.2 describes the set-up of a simulation study to investigate the effect of EPV on the predictive performance, and chapter 4.3 presents the results. Section 4.4 discusses the findings. Section 4.5 concludes.

This chapter is based on the following research paper:

Wynants L, Bouwmeester W, Moons KG, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *J Clin Epidemiol.* 2015; 68: 8.

4.1 Background

Researchers have proposed to use a sample including at least ten events per variable or candidate predictor (EPV) for the development of a risk model.[50, 51] The number of events is the number of observations in the smallest outcome category of the binary outcome. The number of variables, henceforward referred to as predictors, should be interpreted more broadly as the number of parameters to be considered. For example, more than one parameter per predictor must be estimated when polynomial terms are used to model a non-linear effect or when dummy coding is used to model the effect of a qualitative predictor with more than two categories. Hence, in a dataset of 500 observations with 50 events in total, only five parameters should be estimated in order to obtain an EPV of 10. Some researchers have proposed upward[52, 54, 212] and downward[213] adjustments for the EPV guideline, stating that the required EPV is influenced by the size of the predictor effects, the correlations among predictors, the prevalence of dichotomous predictors, the purpose of

the model, and the predictor selection strategy. Data-driven predictor selection in particular may result in strongly biased estimated regression coefficients in small samples, which decreases the predictive performance of the risk model when used in individuals other than those from which the model was developed.[52]

Multicenter consortia are gaining popularity: recruiting patients from different sites produces a representative sample and reduces recruitment times,[135, 214] but yields clustered data. Simulation studies have shown that the amount of clustering, the number of clusters, and cluster size influence the accuracy and precision of the parameter estimates, especially the random effect variances and the predictor effects at the cluster level.[146, 147, 215] These studies, however, considered neither the predictive performance nor the required number of EPV.

This study uses simulated data and an empirical example of the classification of ovarian tumors to study the effect of the number of EPV for risk modeling with clustered data, when using mixed effects logistic regression and predictions assuming an average center effect. We hypothesized that the number of EPV will influence both parameter estimates and the performance of the developed risk model. The effective sample size is smaller than the number of participants, because they have not been sampled independently.[134] The amount of clustering may therefore have a negative effect on the model's performance. However, this impact may be limited, since only the estimated predictor effects are used in prediction, and previous research has shown that these are usually unbiased.[140, 141, 146, 147, 149, 215]

4.2 Design of the simulation study

We studied the influence of the number of EPV on the parameter estimates and the predictive performance of the mixed effects logistic regression model. We built models in samples with varying numbers of EPV, drawn from a source population with a certain degree of clustering. The predictive performance of the models was tested in the source population. In what follows, we describe the design of the simulation study (technical details are in Appendix 2 and R code in Appendix 3) and the performance measures employed for validation.

4.2.1 The source populations

We created source populations (of approximately 100,000 observations across 200 clusters) with an intraclass correlation (ICC) of 0%, 5% or 20%. These values were chosen to reflect situations without clustering, with moderate clustering, and with extreme but realistic clustering in multicenter prediction research.[216] The source populations were generated according to a model with four uncorrelated continuous predictors (X_1~N(0,1), X_2~N(0,0.6), X_3~N(0,0.4), X_4~N(0,0.2)), four uncorrelated dichotomous predictors (X_5 to X_8, with prevalence 0.2, 0.3, 0.3 and 0.4, respectively), and a random intercept of which the variance was determined by the ICC. All regression coefficients were set at 0.8 to achieve a level of discrimination that occurs regularly in the clinical literature (C=0.8). The overall intercept was set at -2.1 to obtain an outcome event rate of 0.3, which is common for health outcomes in prediction research. For one population (with ICC=20%), we introduced a correlation between the random intercepts and X_1 and X_5, such that predictors were unequally distributed across clusters. The mean of X_1 ranged from -1.13 in the cluster with the lowest random intercept to 1.44 in the cluster with the highest random intercept, while the prevalence of X_5 ranged from 12% to 29%.

4.2.2 Sampling

We sampled datasets from the source populations. The number of EPV was set at 5, 10, 20 or 50. These values reflect popular choices and guidelines for prediction research. In clustered data, the EPV is determined as $EPV=N\times p/(K+1)$, where the total sample size N equals the sum of all individual cluster sizes n_j (j=1 to J), p is the sample's event rate, and K+1 is the number of parameters in the model to be estimated, including the random intercept variance. We defined several simulation conditions, grouped into sets of simulations which are characterized by the parameters that are varied (Table 12): the average number of observations per cluster ($9\leq\bar{n}_j\leq89$) in set 1 to 3 and set 6, the number of clusters (J=5, 10, 20, 30 or 50) in set 4, and the sample's event rate (p=0.05, 0.1, 0.2, 0.3 or 0.5) in set 5. The event rate in the sample determines the total sample size when EPV and the number of predictors are fixed, since $N=EPV\times(K+1)/p$. We drew 500 datasets per simulation condition.

Table 12. Simulation conditions.

Set	Condition	Population		Sample				Model		EPV
		ICC (%)	Corr (X,a_j)	J	$\bar{n}_j$	N	p	K+1	Backward selection	
1	1.1	**0**	0	30	**5**	**150**	0.3	8+1	No	**5**
1	1.2	**5**	0	30	**5**	**150**	0.3	8+1	No	**5**
1	1.3	**20**	0	30	**5**	**150**	0.3	8+1	No	**5**
1	1.4	**0**	0	30	**10**	**300**	0.3	8+1	No	**10**
1	1.5	**5**	0	30	**10**	**300**	0.3	8+1	No	**10**
1	1.6	**20**	0	30	**10**	**300**	0.3	8+1	No	**10**
1	1.7	**0**	0	30	**20**	**600**	0.3	8+1	No	**20**
1	1.8	**5**	0	30	**20**	**600**	0.3	8+1	No	**20**
1	1.9	**20**	0	30	**20**	**600**	0.3	8+1	No	**20**
1	1.10	**0**	0	30	**50**	**1500**	0.3	8+1	No	**50**
1	1.11	**5**	0	30	**50**	**1500**	0.3	8+1	No	**50**
1	1.12	**20**	0	30	**50**	**1500**	0.3	8+1	No	**50**
2	2.1	20	0	30	**9**	**270**	0.3	8+8 noise+1	No	**5**
2	2.2	20	0	30	**18**	**540**	0.3	8+8 noise+1	No	**10**
2	2.3	20	0	30	**36**	**1080**	0.3	8+8 noise+1	No	**20**
2	2.4	20	0	30	**89**	**2670**	0.3	8+8 noise+1	No	**50**
3	3.1	20	0	30	**9**	**270**	0.3	8+8 noise+1	Yes	**5**
3	3.2	20	0	30	**18**	**540**	0.3	8+8 noise+1	Yes	**10**
3	3.3	20	0	30	**36**	**1080**	0.3	8+8 noise+1	Yes	**20**
3	3.4	20	0	30	**89**	**2670**	0.3	8+8 noise+1	Yes	**50**
4	4.1	20	0	**5**	30	**150**	0.3	8+1	No	**5**
4	4.2	20	0	**10**	30	**300**	0.3	8+1	No	**10**
4	4.3	20	0	**20**	30	**600**	0.3	8+1	No	**20**
4	4.4	20	0	**50**	30	**1500**	0.3	8+1	No	**50**
5	5.1	20	0	30	30	900	**0.05**	8+1	No	**5**
5	5.2	20	0	30	30	900	**0.1**	8+1	No	**10**
5	5.3	20	0	30	30	900	**0.2**	8+1	No	**20**
5	5.4	20	0	30	30	900	**0.5**	8+1	No	**50**
6	6.1	20	>0	30	**5**	**150**	0.3	8+1	No	**5**
6	6.2	20	>0	30	**10**	**300**	0.3	8+1	No	**10**
6	6.3	20	>0	30	**20**	**600**	0.3	8+1	No	**20**
6	6.4	20	>0	30	**50**	**1500**	0.3	8+1	No	**50**

The varying parameters within each set of conditions are indicated in bold; ICC: intraclass correlation; corr(X, aj): correlation between predictors (X1 and X5) and the random intercept; J: the number of clusters; $\bar{n}_j$: the average number of observations per cluster; N: the total number of observations; p: the event rate of the outcome in the sample; K+1: the number of parameters to be estimated, including the random intercept variance; EPV: the number of events per variable.

4.2.3 Model building

Random intercept models (see equation (8)) were fitted in each sample. They were predefined models using all eight true predictors (sets 1, 4, 5, 6), full models including the eight true predictors and eight noise variables (set 2), or reduced models determined through backward variable selection (α=0.1) starting from all sixteen candidate predictors (set 3) (Table 12). The noise variables had the same distributions as the true predictors. To limit the amount of simulations, the presence of noise predictors and variable selection (sets 2 and 3), the effects of the ICC (set 1), and the correlations between random intercepts and predictors (set 6) were only studied in samples with J=30 and p=0.3.

We used the following criteria for model convergence: 10 to 100 iterations to fit the model, a change of less than 10^{-5} in deviances of the models fitted in the last two iterations, and no outlying estimated regression coefficients and standard errors (visual inspection). All models met these criteria and no samples needed to be deleted.

4.2.4 Model evaluation

All models were evaluated in terms of the accuracy of estimated regression parameters and the predictive performance.

Bias in the estimated regression coefficients

We compared each estimated regression coefficient $\hat{\beta}$ to β_{sp}, the regression coefficient when the model was built in the source population. The percentage of relative bias in the estimated regression coefficients was defined as $100\times(\hat{\beta}-\beta_{sp})/\beta_{sp}$. The use of β_{sp} ensures that random error originating from the generation of a source population is not included in the computation of the bias in the estimated regression coefficients. β_{sp} was between 0.770 and 0.857 for all predictors in all source populations. The relative bias of the estimated random intercept variance was computed analogously.

Predictive performance

The predictive performance of the resulting risk models was tested in the corresponding source population. We used predictions for the average center, omitting the random intercept, to be able to make predictions in clusters that were not represented in the sample.[142, 151]

The c-statistic or concordance probability (C) measured the discriminatory performance. It expresses the probability that for a randomly selected pair of an event and a non-event, the event has a higher predicted probability.[5] For the computation of the standard c-statistic, pairs of events and non-events belonging to different clusters as well as pairs from the same cluster are compared:[5]

(17)

$$\hat{C}=\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(\hat{p}_{ij}>\hat{p}_{i'j'} \text{ and } y_{ij}=1 \text{ and } y_{i'j'}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(y_{ij}=1 \text{ and } y_{i'j'}=0\right)}.$$

The calibration slope (β_{cal})[6, 65] was used to evaluate overfitting in the source population. It is obtained through logistic regression of the event indicator against the linear predictor of the risk model:

(18)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+\beta_{cal}LP_{ij}.$$

If β_{cal} is smaller than one, there is overfitting, in other words, the predicted probabilities are too extreme (high risks are overestimated and too close to 1, low risks are underestimated and too close to 0); if β_{cal} is larger than one, there is underfitting, that is, the predicted probabilities are not extreme enough (too close to the prevalence).

We also computed the within-cluster c-statistic (C_w) and the within-cluster calibration slope ($\beta_{cal\,w}$) to evaluate the performance at the cluster level instead of the population level. C_w is computed by only comparing pairs of events and non-events within the J centers:[217]

(19)

$$\hat{C}_w=\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(\hat{p}_{ij}>\hat{p}_{i'j} \text{ and } y_{ij}=1 \text{ and } y_{i'j}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(y_{ij}=1 \text{ and } y_{i'j}=0\right)}.$$

This corresponds to the average center-specific c-statistic, weighted by the number of pairs of events and non-events per center. Other weights may be used as well (see Chapter 6).[218]

$\beta_{cal\ w}$ is estimated using logistic regression with random cluster intercepts and slopes:[152]

(20)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\ w}+a_{j\ cal}+\beta_{cal\ w}LP_{ij}+b_{j\ cal}LP_{ij}$$

$$\begin{pmatrix}a_{j\ cal}\\ b_{j\ cal}\end{pmatrix}\sim N\begin{pmatrix}0 & \tau_a^2 & \tau_{ab}\\ 0' & \tau_{ab} & \tau_b^2\end{pmatrix}.$$

$\beta_{cal\ w}$ now is the average within-center calibration slope, while τ_b^2 is the variance of the within-center calibration slopes $b_{j\ cal}$. τ_{ab} is the covariance between the intercepts and calibration slopes.

The obtained c-statistics and calibration slopes were compared to those of a model developed and evaluated in the source population (henceforward C_{sp} and $\beta_{cal\ sp}$), which serve as upper limits for discriminatory power and calibration in the given population. The relative c-statistic and the relative calibration slope were computed as $100 \times \widehat{C}/C_{sp}$ and $100 \times \hat{\beta}_{cal}/\beta_{cal\ sp}$ (or $100 \times \widehat{C}_w/C_{w\ sp}$ and $100 \times \hat{\beta}_{cal\ w}/\beta_{cal\ w\ sp}$ for the within-cluster measures). Note that $\beta_{cal\ sp}$ will deviate from one because predictions for the average center, omitting the random intercepts, are used (see chapter 7).

All simulations and calculations were performed in R (http://www.r-project.org/) version 2.14.0.[133] The lmer function from the lme4 package[219] was used to fit mixed effects logistic regression models using Laplace approximation and the rms package was used for model evaluation.[5]

4.3 Results of the simulation study

4.3.1 Data clustering and the number of events per variable

The amount of clustering (ICC) did not influence the relative bias of the estimated regression coefficients (Figure 13A, representing results from simulation set 1). The median bias was close to 0% at each ICC. The bias of the estimated regression coefficients related to the EPV: the largest interquartile range (IQR) of the relative bias was obtained with the lowest EPV values. At ICC=20%, the median relative bias of $\hat{\beta}_4$ was -10% at EPV=5 and

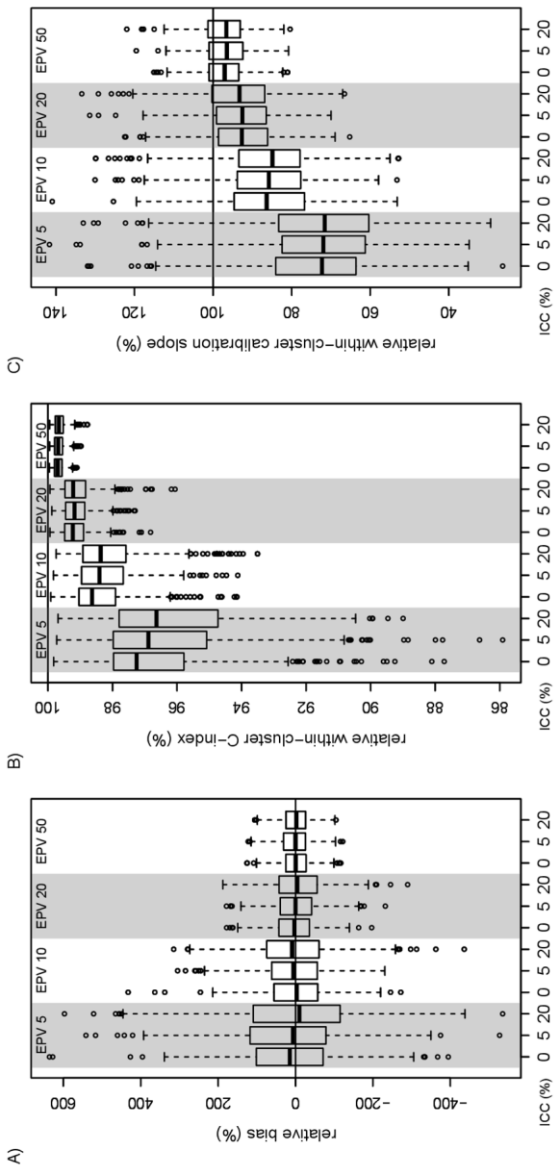


Figure 13. The relative bias in estimated regression coefficients and the predictive performance in relation to the number of events per variable (EPV) and the amount of clustering (ICC). A) the relative bias (%) in the estimated regression coefficient $\hat{\beta}_4$; B) the relative within-cluster c-statistics (%); C) the relative within-cluster calibration slope (%). The box indicates the interquartile range (IQR), the bold horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

-2% at EPV=50. Similar patterns were observed for the other regression coefficients (Table A 1). The random intercept variances were often underestimated, but a large number of EPV improved the estimation (median relative bias -15.2% at EPV=5 to -5.2% at EPV=50 for ICC=20%).

Model performance also related to the number of EPV. The within-cluster c-statistic of the model fitted and evaluated in the source population was 0.78 in each source population (ICC=0%, 5% and 20%). The relative within-cluster c-statistic of models fitted in the samples was the lowest for the samples with EPV=5, with median values of around 97% (Figure 13B). The calibration slopes of the models fitted in the source populations were 1.00 at each ICC and the relative performance of the models fitted in the samples were similar for varying ICCs (Figure 13C). At ICC=20%, the median relative within-cluster calibration slope increased from 71.6% at EPV=5 to 96.6% at EPV=50. The IQR of the calibration slope decreased with increasing EPV as well.

The same patterns were observed for the relative standard c-statistic and the relative standard calibration slope (Figure A 1, panels A and B), but the standard c-statistics and calibration slopes of the models fitted in the source populations decreased as ICC increased (C_{sp} 0.78, 0.78 and 0.76, and $\beta_{cal\ sp}$ 1.00, 0.97 and 0.88, for ICC=0%, 5% and 20% respectively). We will revisit this issue in chapter 7.

4.3.2 Variable selection

A high EPV was required to ensure the inclusion of important predictors and to prevent bias in the estimates when using backward variable selection (α=0.1). At EPV=5, X_4 was selected in only 24% percent of the samples. To ensure the selection of X_4 in >90% of the samples, 50 EPV were required (Table 13). The median relative bias of $\hat{\beta}_4$ (conditional on selection) was 118% at EPV=5 and disappeared at EPV=50 (0.05% relative bias) (Figure 14A, representing results from simulation conditions 1.3, 1.6, 1.9 and 1.12, and sets 2-3). The selection bias for other predictors was negligible. Predefined models that included all candidate predictors showed biases similar to predefined models that included only the eight true predictors. The smaller IQRs in models with sixteen variables reflect the larger sample sizes required to obtain the EPV values with sixteen rather than eight predictors.

Table 13. The selection frequency of predictors.

	EPV=5	EPV=10	EPV=20	EPV=50
X_1	100%	100%	100%	100%
X_2	88%	99%	100%	100%
X_3	60%	85%	99%	100%
X_4	24%	43%	65%	93%
X_5	64%	87%	100%	100%
X_6	72%	95%	100%	100%
X_7	76%	95%	100%	100%
X_8	79%	95%	100%	100%

EPV: events per variable.

Models containing all sixteen candidate predictors had a slightly better median relative within-cluster c-statistic and a lower median relative within-cluster calibration slope than models after variable selection at EPV=5 (95.9% versus 95.3% of $C_{w\ sp}$=0.78, and 67.3% versus 72.1% of $\beta_{cal\ w\ sp}$=1.00, Figure 14, panels B and C). This is in accordance with earlier findings.[220] The differences in performance decreased as EPV increased. At EPV=10, 20 and 50, the predictive performance of the models after variable selection was comparable to the model including only true predictors. The lower relative within-cluster c-statistic after variable selection at EPV=5 may be explained by the frequent exclusion of relevant predictors (Table 13). The same pattern was observed for the standard c-statistic and calibration slope (Figure A 1, panels C and D).

4.3.3 Sample size

The IQR of the relative bias in the estimated regression coefficients and the model performance related to the total sample size. The variability of the relative bias was smaller in samples with EPV=5 and a total sample size of 900 than in samples with EPV=5 and a total sample size of 150 (Figure 15A, representing results of simulation conditions 1.3, 1.6, 1.9, 1.12, and sets 4-5). A large total sample size (for the same EPV and number of centers) and a large number of clusters (for the same EPV and total sample size) reduced the relative bias in the estimated random intercept variance (Figure A 2). The relative within-cluster c-statistics and calibration slopes of models fitted in the larger samples were higher (Figure 15, panels B and C). For a given total sample size and number of EPV, samples with many small clusters yielded a predictive performance comparable to samples with a few large clusters. Note

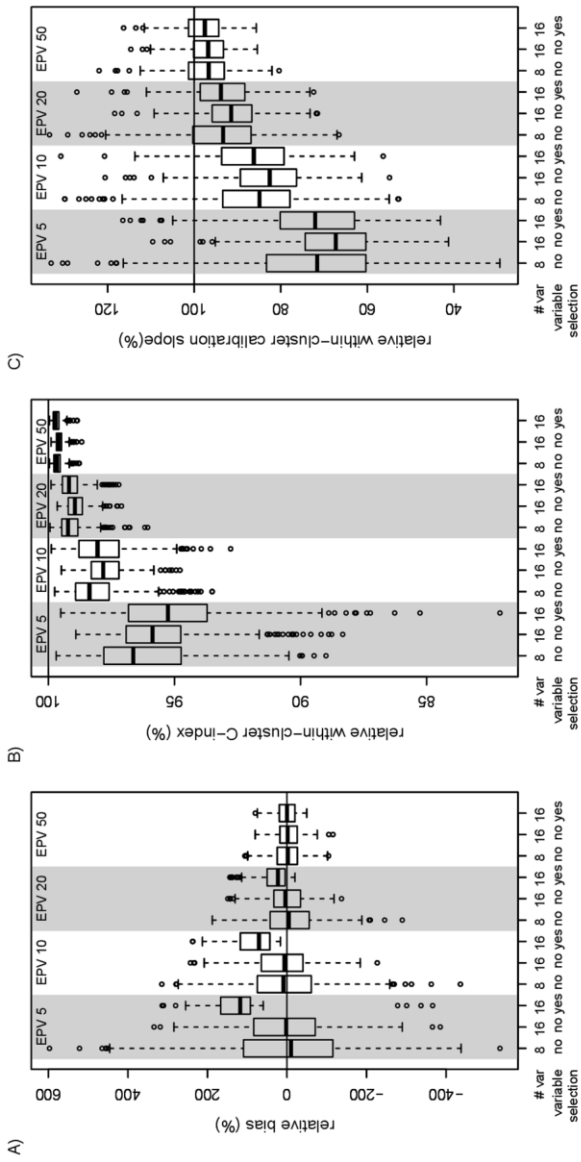


Figure 14. The relative bias in estimated regression coefficients and the predictive performance in relation to the number of events per variable (EPV). #var: number of candidate predictors. A) the relative bias (%) in the estimated regression coefficient $\hat{\beta}_4$; B) the relative within-cluster c-statistic; C) the relative within-cluster calibration slope (%). The box indicates the interquartile range (IQR), the bold horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

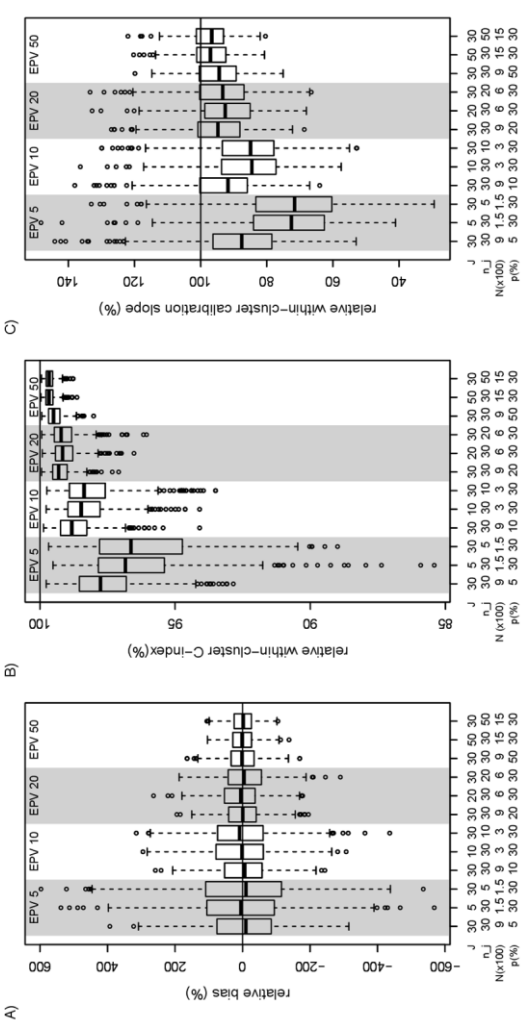


Figure 15. The relative bias in estimated regression coefficients and the predictive performance in relation to the number of events per variable (EPV) and sample characteristics (ICC=20%). J: the number of clusters, n_j: the average number of observations per cluster; N (x100): the total number of observations, to be multiplied by 100; p: prevalence. A) the relative bias (%) in the estimated regression coefficient $\hat{\beta}_4$; B) the relative within-cluster c-index (%); C) the relative within-cluster calibration slope (%). The box indicates the interquartile range (IQR), the bold horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

that at EPV=5, the IQRs of the bias in the estimated regression coefficients, the within-cluster C-indices and the within-cluster calibration slopes were slightly smaller in samples with a few large clusters. The standard ("overall" or population-level) performance measures showed similar patterns (Figure A 1, panels E and F).

4.3.4 Random cluster effects correlated with predictors

When the assumption of independence between random intercepts and predictors was violated, $\hat{\beta}_1$ was biased (Figure 16A, representing the results of simulation conditions 1.3, 1.6, 1.9, 1.12, and set 6). The same holds for $\hat{\beta}_5$. These regression coefficients accounted for the cluster-level association of X_1 and X_5 with the random intercept, yielding more severely underestimated random intercept variances (Figure 16B). Since $\hat{\beta}_1$ and $\hat{\beta}_5$ contributed to explaining differences between clusters, the relative standard ("overall" or population-level) c-statistic was increased compared to the situation in which predictors and random intercepts were independent (Figure 16D). This could not be observed for the within-cluster c-statistic, as $\hat{\beta}_1$ and $\hat{\beta}_5$ did not enable a better discrimination within clusters (Figure 16C). Finally, because $\hat{\beta}_1$ and $\hat{\beta}_5$ were positively biased, overfitting was more problematic (Figure 16, panels E and F).

4.4 Empirical example

We developed risk models to diagnose ovarian cancer using data from the International Ovarian Tumor Analysis (IOTA) group.[176, 178] We analyzed clinical and ultrasound information on 5912 patients with ovarian masses from 24 hospitals, collected between 1999 and 2012 (phases I to III). The data collected up until 2005 (n=1571, phases I and Ib, 9 centers, 409 (26%) malignant tumors) was used for the development of risk models for tumor malignancy that included random intercepts for centers. We drew one hundred samples of 409 events and 1162 non-events from the training set, with replacement. Seven predictors were considered (Table 14). Together with the random intercept variance, this yielded eight parameters to estimate (EPV=51). We further drew one hundred random subsets of 154 patients (40 malignancies) to obtain development samples with EPV=5. The most extreme estimates were obtained when variable selection was performed with EPV=5 (Table 14). Note that the random intercept variances estimated by the models

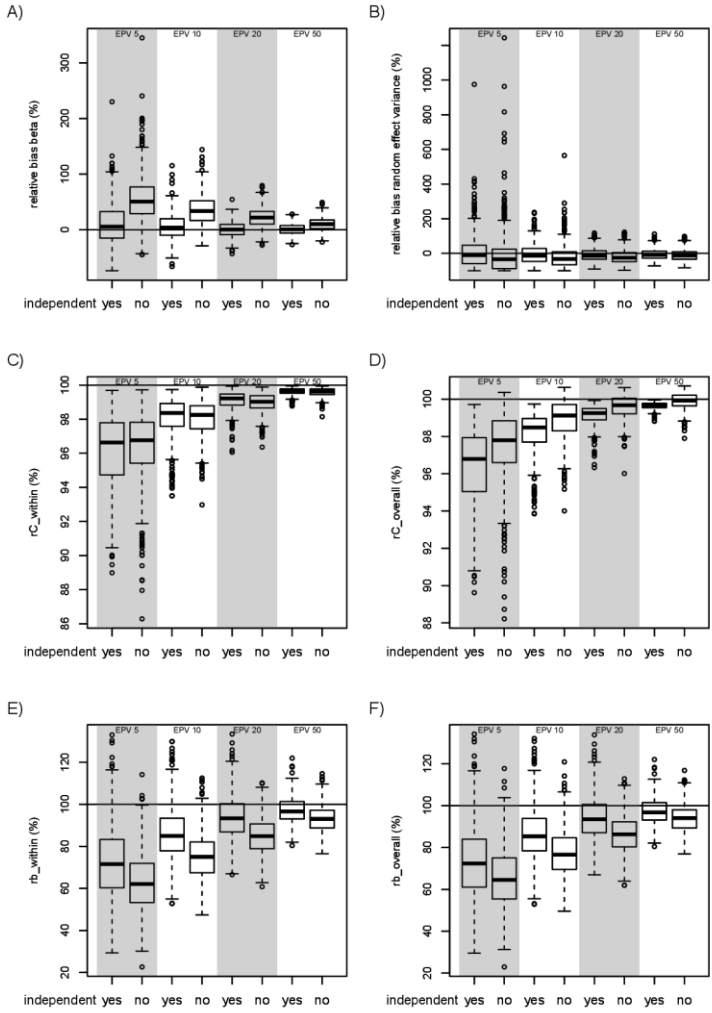


Figure 16. The relative bias in the estimated regression coefficients and the predictive performance in relation to the number of events per variable (EPV) when predictors are dependent versus independent of the random intercept (ICC=20%). ICC: intraclass correlation; rC_{within}: relative within-cluster C-index; $rC_{overall}$: standard relative C-index; rb_{within}: relative within-cluster calibration slope; $rb_{overall}$: standard calibration slope (%). A) relative bias in the estimated regression coefficient $\hat{\beta}_1$; B) estimated random intercept variance; C) relative within-cluster c-statistic; D) relative standard c-statistic (%); E) relative within-cluster calibration (%); F) relative standard calibration. The box indicates the interquartile range (IQR), the bold horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

fitted with EPV=5 were lower than the estimates obtained with EPV=51. This is in line with the findings of the simulation study, where the random intercept variance was underestimated in samples with EPV=5 (Figure A 2).

The data collected after 2005 (n=4341, phases II and III, 22 centers, 1522 (35%) malignancies) was used for the models' validation. The models' performance depended on the number of EPV: models developed on a dataset with a low EPV (5) had a similar discriminative ability to models developed on a dataset with high EPV (51), but were overfitted (Table 15). The full model fitted with EPV=51 gave a median validated within-cluster calibration slope of 0.966 (IQR 0.934 to 1.011) while the full model fitted with EPV=5 gave a much lower median calibration slope of 0.763 (IQR 0.509 to 0.855). The effect of EPV was similar when backward variable selection (α=0.10) was used: the median within-cluster calibration slopes were 0.970 (IQR 0.939 to 1.015) and 0.815 (IQR 0.735 to 0.933) for the models fitted with EPV=51 and

Table 14. Estimated coefficients of the fitted models for ovarian tumor diagnosis in 100 bootstrap samples, based on EPV=5 versus EPV=51, without (model 1 and model 2) and with (model 3 and model 4) backward variable selection (α=0.10).

Predictor	**Model 1 (EPV=51) Median (IQR)**	**Model 2 (EPV=5) Median (IQR)**	**Model 3 (EPV=51) Median (IQR) [selection frequency]**	**Model 4 (EPV=5) Median (IQR) [selection frequency]**
Age (per 10 years)	0.38 (0.35 to 0.42)	0.40 (0.27 to 0.55)	0.39 (0.36 to 0.42) [100]	0.48 (0.39 to 0.61) [76]
Maximum diameter of the lesion (per 10 mm)	0.12 (0.11 to 0.13)	0.13 (0.08 to 0.18)	0.12 (0.11 to 0.13) [100]	0.16 (0.11 to 0.19) [82]
Presence of solid tissue in the lesion (yes versus no)	3.25 (3.13 to 3.40)	3.69 (2.98 to 4.25)	3.23 (3.08 to 3.39) [100]	3.49 (2.91 to 4.08) [91]
Family history of ovarian cancer (yes versus no)	0.40 (0.20 to 61)	0.40 (-0.08 to 1.01)	0.69 (0.62 to 0.82) [30]	2.17 (1.99 to 2.26) [4]
Current use of hormonal therapy (yes versus no)	-0.35 (-0.46 to -0.24)	-0.47 (-0.87 to -0.07)	-0.43 (-0.54 to 0.38) [60]	-1.24 (-1.39 to -1.15) [15].
Pelvic pain during examination (yes versus no)	-0.13 (-0.25 to 0.01)	-0.32 (-0.62 to 0.03)	-0.39 (-0.43 to -0.33) [21]	-1.35 (-1.63 to -1.23) [8]
Presence of papillary structures (yes versus no)	-0.07 (-0.20 to 0.08)	-0.15 (-0.57 to 0.28)	-0.28 (-0.32 to -0.26) [14]	1.03 (-1.14 to 1.17) [24]
Intercept	-6.50 (-6.80 to -6.25)	-7.20 (-8.49 to –5.79)	-6.56 (-6.81 to -6.28)	-6.68 (-7.85 to -5.04)
Random intercept variance	0.59 (0.45 to 0.75)	0.20 (0.00 to 0.67)	00.61 (0.45 to 0.75)	0.19 (0.00 to 0.67)

EPV: events per variable.

Table 15. The performance of the fitted models for ovarian tumor diagnosis, based on EPV=5 versus EPV=51 in 100 bootstrap samples, without (model 1 and model 2) and with (model 3 and model 4) backward variable selection (α=0.10).

	Model 1 Predefined EPV=51	**Model 2 Predefined EPV=5**	**Model 3 Variable selection EPV=51**	**Model 4 Variable selection EPV=5**
Within-cluster C-index (IQR)	0.862 (0.860 to 0.864)	0.854 (0.844 to 0.859)	0.862 (0.861 to 0.863)	0.851 (0.842 to 0.859)
Within-cluster calibration slope (IQR)	0.966 (0.934 to 1.011)	0.763 (0.509 to 0.855)	0.970 (0.939 to 1.015)	0.815 (0.735 to 0.933)
Standard C-index (IQR)	0.882 (0.881 to 0.884)	0.873 (0.865 to 0.878)	0.882 (0.881 to 0.883)	0.868 (0.858 to 0.878)
Standard calibration slope (IQR)	0.955 (0.923 to 1.002)	0.751 (0.477 to 0.850)	0.959 (0.925 to 1.00)	0.809 (0.732 to 0.925)

EPV: events per variable.

EPV=5, respectively. The median validated within-cluster c-statistics were only slightly lower for models developed with EPV=5 (0.854 (IQR 0.844 to 0.859) without variable selection and 0.851 (IQR 0.842 to 0.859) with variable selection), than for models developed with EPV=51 (0.862 (IQR 0.860 to 0.864) without variable selection and 0.862 (IQR 0.861 to 0.863) with variable selection).

4.5 Discussion

Our research showed that the number of EPV determines the bias in parameter estimates of risk models developed in clustered data using mixed effects logistic regression, as well as the resulting models' predictive performance (discrimination and calibration) in external data. In contrast, the amount of clustering was not meaningfully associated with the models' predictive performance. The results further suggest that larger samples provide better models for a given number of EPV. This means that non-events contribute to the stability of the risk model, provided the number of events is sufficient. When variable selection was performed, a high number of EPV was needed to ensure the inclusion of relevant predictors and reduce estimation bias in the estimated predictor effects.

In accordance with earlier proposals for non-clustered data,[50, 51] we recommend using at least ten EPV to fit a predefined risk model in clustered data, although up to fifty EPV may be needed when stepwise variable selection is applied.[52] There were no negative effects of clustering in the simulation conditions we have considered if the number of EPV was

sufficiently large. However, it must be noted that it is impossible to obtain an optimal standard ("overall" or population-level) calibration slope for a random intercept model when predictions for the average center (omitting random intercept estimates) are used. Even if the model was fitted on the data of the entire source population, the median overall calibration slope was 0.88 rather than 1 when the ICC was 20%. The within-cluster calibration slope does not suffer from this issue. We will elaborate on this topic in chapter 7.

Common formulas for power calculations for the design of experiments take into account the ICC, because clustering reduces the effective sample size and necessitates larger samples.[134] Nonetheless, the estimation of the regression coefficients used for prediction purposes is little influenced by clustering,[140, 141, 146, 147, 149, 215] provided that the assumption of independence between predictors and random effects is not violated, and there is no interaction between predictor effects and cluster.[137] Hence, the existing EPV guidelines apply in clustered data if a random cluster intercept is added to the risk model and the estimation of this additional parameter is accounted for in the EPV calculation.

Mixed effects models are very useful tools when analyzing clustered datasets. Random slopes can be used to investigate interactions between predictor effects and cluster, and random intercepts can be used to model differences in outcome prevalence across clusters. It is difficult to reliably estimate the variance of the random intercept, but our results show that estimation improves when the number of clusters increases. This confirms earlier findings.[146, 147, 215] It has been recommended to collect data in at least fifty clusters,[147] although random effect variances may still be slightly underestimated with hundreds of clusters.[146] In reality, however, the number of centers in multicenter research is most often smaller than fifty, and determined by weighing benefits, such as the reduction of recruitment times, against practical concerns, to maintain the manageability of the study. Our findings also demonstrate that a high EPV benefits the estimation of the ICC, regardless of the total sample size or the number of clusters. This has not been studied previously.

The focus on the predictive performance of models, alongside the estimation of regression parameters, is a strength of this study, because the purpose of risk modeling is to make reliable predictions for new subjects.[6] For the same reason, other evaluation criteria such as confidence interval coverage, type I error rates, and the power of statistical tests are of lesser importance. The uncertainty inherent in model building was acknowledged by studying variable selection. We have used within-cluster performance measures to

acknowledge the use of risk models in separate centers.[218] Our simulation study, like all simulation studies, is restricted by the choice of simulation parameter settings. However, we have chosen practically relevant settings for the simulation parameters, such as the ICC, the number of EPV, and the number of clusters. Furthermore, we did not assume equal numbers of observations in clusters, as this hardly ever occurs in multicenter research.[214] The parameters that were not varied, such as the estimation method (Laplace approximation)[134] and the variable selection method (backward variable selection),[5, 6] were set at generally accepted or recommended choices. A final strength of our study is that we studied the effect of violating the assumption of independence between predictors and random effects.[134] Because two predictors were dependent on the random intercepts, they were also correlated with each other. We did not consider random predictor effects, assuming homogeneity of predictor effects across clusters. Center-level predictors were not included in the study. Reliable estimation of center-level effects will depend on the number of clusters.

4.6 Conclusion

This chapter showed that the number of events per variable is useful in guiding sample size decisions when the aim is to develop risk models using clustered data. The number of events per variable determines the bias in estimated parameters and the predictive performance of the model. The degree of clustering does not meaningfully increase the bias or decrease the performance for a given number of EPV.

5 Predictor selection for multicenter studies

This chapter examines techniques to select predictors in a multicenter study. Section 5.1 proposes a method to screen for clustering in predictor variables. Section 5.2 illustrates how the results of a naive backward variable selection procedure that ignores clustering differ from the results of backward variable selection that accounts for clustering. Section 5.3 concludes.

This chapter is based on the following research:

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 Med Res Methodol*. 2013; 13.

Wynants L, Timmerman D, Van Huffel S and Van Calster B. Variable selection for prediction models based on multicenter data. Annual Conference of the International Society of Clinical Biostatistics Bergen, Norway 2012.

5.1 Screening for data clustering in multicenter studies: the residual intraclass correlation

In multicenter studies, center-specific variations in measurements may arise for various reasons, such as low interrater reliability, differences in equipment, deviations from the protocol, sociocultural characteristics, and differences in patient populations, for instance due to local referral patterns. This section proposes a method to detect heavily clustered variables and identify physicians with outlying measurements.

5.1.1 Background

Multicenter consortia are rapidly gaining popularity in clinical research. Recruiting patients from a broad range of settings yields a representative

sample and simultaneously reduces patient recruitment times, which is especially appealing when studying rare diseases.[135, 214] A complication that arises from multicenter data collection is data clustering. In many studies it is useful to investigate clustering in detail to understand the data structure or to detect problems. Important causes of clustering are non-random differences in the measurements taken by different physicians recruiting patients in the contributing centers. Intraclass correlations (ICC) have traditionally been used to study interrater reliability. The ICC is commonly defined as the correlation between quantitative measurements made by different judges or raters for the same measurement object.[221, 222] However, in the context of multicenter studies, it has some drawbacks. First, it requires at least two raters measuring variables for the same patient. Since most multicenter studies are not designed to study interrater reliability, this would require additional efforts when designing the study and collecting the data. It is often impractical and cost-intensive to have physicians from different centers examining the same patients. Second, interrater reliability studies do not uncover all relevant forms of data clustering in multicenter studies. Besides the potential subjectivity of measurements, there are various other reasons for non-random differences across physicians. Differences in equipment or equipment settings, as well as deviations from the study protocol may give rise to clustering. Regional differences across centers may cause data clustering via sociocultural characteristics of patients or procedures. It is, for example, well documented that the perception and expression of pain is related to ethnicity.[223] Differences in measurements across physicians may also occur because of differences in their patient populations, for example due to local referral patterns.

An alternative formulation of the ICC (introduced in section 3.2.2) does not require measurements by multiple physicians for the same patient. Instead, it requires the assessment of multiple patients per physician, which is more in line with typical data collection in multicenter studies. In this setup, the ICC is defined as the correlation between any two measurements made by the same physician, or equivalently the amount of variance in the measurements that is attributable to the level of the physician. Consequently, the ICC is able to detect various forms of data clustering. It can be used for data quality control, as was recently demonstrated by Guthrie and colleagues,[224] or as a screening step before variable selection to develop a risk model. A drawback of this formulation of the ICC is that it does not acknowledge that physicians may systematically record lower or higher measurements because of different patient populations. In this work we present mixed effects models to study clustering while simultaneously taking case-mix differences across physicians into account. We use the residual intraclass correlation (RICC), an extension

of the ICC, to quantify the degree of clustering. We also relate the RICC to the proportion of explained variance by partitioning the total variance into explained and residual variance at the patient and physician level. We demonstrate these concepts on a multicenter dataset collected for the development and validation of risk models for preoperative ovarian tumor diagnosis.

5.1.2 Methods

A mixed effects model

Suppose a variable S_{ij} is a continuous variable measured in N patients by J physicians (clusters), $i = 1,\ldots, n_j$ and $j = 1,\ldots, J$. S_{ij} could be, for example, the size of a tumor. To account for data clustering, we consider a mixed effects model of S_{ij} by incorporating random intercepts a_j. This recognizes that the average measured value may systematically vary from physician to physician:

(21)

$$S_{ij}=\alpha+a_j+e_{ij}$$

$$a_j \sim N(0,\tau^2)$$

$$e_{ij} \sim N(0,\sigma^2_{error}).$$

The random intercept at the physician level and the random error term at the patient level (e_{ij}) are assumed to have normal distributions with mean zero and variance τ^2 and σ^2_{error}, respectively. When S_{ij} is a nominal or ordinal variable, model (21) can easily be reformulated as a generalized linear mixed model. Throughout this research, we use the logit link function in logistic regression models for dichotomous variables and proportional odds models for ordinal variables. Consequently, the error terms have a logistic distribution with $\sigma^2_{error} = \pi^2/3 \approx 3.29$.[134]

The intraclass correlation coefficient

The crude amount of clustering in S_{ij} can be expressed as the proportion of variance at the cluster level. The total variance in S_{ij} can be split up into the variance at the patient level (σ^2_{error}) and variance at the physician level (τ^2). The intraclass correlation (ICC) is defined as:

(22)

$$ICC=\frac{\tau^2}{\sigma^2_{error}+\tau^2}.$$

The patient populations often vary between clusters. In this case a more refined metric is required, as described below.

Patient case-mix

The patient case-mix can account for a substantial part of the between-physician variance. The prevalence of the disease under study may differ across physicians in, for example, a diagnostic accuracy study. In that case, the true health status of patient i seen by physician j (H_{ij}), as established by the golden standard for diagnosis, can be included as an explanatory variable in the model of S_{ij}:

(23)

$$S_{ij}=\alpha+a_j+\beta_1\times H_{ij}+e_{ij}$$

$$a_j\sim N(0,\tau^2)$$

$$e_{ij}\sim N(0,\sigma^2_{error}).$$

The health status will usually be represented by a dummy variable indicating presence or absence of the disease of interest, although an extension to multiple dummies to describe the true health status in more than two categories is straightforward. In what follows, we use one dummy for the health status (H_{ij}) as the only case-mix factor of interest, for the sake of simplicity. In other applications, other patient-level variables may be more relevant, and can be added to the model as necessary. Patient-level variables that are related to S_{ij} and have unequal distributions across clusters potentially explain clustering, and are thus the most relevant. Patient-level factors influencing S_{ij} that are not included in the model are regarded as disturbances captured by the error term.

The residual intraclass correlation coefficient

The total variance of S_{ij} can now be split up into a part that is accounted for by the explanatory variables (σ^2_{LP}), an unexplained part at the patient level (σ^2_{error}) and an unexplained part at the physician level (τ^2). σ^2_{LP} is the variance of the linear predictor, excluding the random intercepts. In model (23) this is

var($\beta_1 \times H_{ij}$). R^2 is the proportion of variance in S_{ij} accounted for by the explanatory variables:

(24)

$$R^2 = \frac{\sigma^2_{LP}}{\sigma^2_{LP} + \sigma^2_{error} + \tau^2}.$$

A high R^2 indicates that S_{ij} is little influenced by unexplained variance at the physician and patient level. The proportion of total variance at the physician level is labelled as a 'variance partitioning coefficient' (VPC):

(25)

$$VPC = \frac{\tau^2}{\sigma^2_{LP} + \sigma^2_{error} + \tau^2}.$$

Theoretically, this VPC can take values between 0 and 1, but 1 will only be reached if there is no residual variance at the patient level ($\sigma^2_{error} = 0$) while the explanatory variables have the same value for all patients or all have regression coefficients of zero ($\sigma^2_{LP} = 0$). The VPC is difficult to interpret as a measure of the degree of clustering because its magnitude depends on σ^2_{LP}.

The residual intraclass correlation (RICC) excludes σ^2_{LP} from the denominator:

(26)

$$RICC = \frac{\tau^2}{\sigma^2_{error} + \tau^2}.$$

Hence, the RICC is the proportion of unexplained variance situated at the physician level. Equivalently, it can be interpreted as the residual correlation between the measurements on any two patients seen by the same physician, after the effects of the explanatory variables have been taken into account.[134] The RICC reaches a maximum value of 1 when all of the variance in S_{ij} that is not explained by the explanatory variables is situated at the physician level. If the unexplained variance is only situated at the level of the patient, the RICC reaches its minimum value of 0, and there are no between-physician differences. This makes the RICC a pure measure of clustering and an easy to interpret screening tool. In certain applications, however, it may be useful to include R^2 in the evaluation of the RICC. Variables with a high R^2 may be

strongly associated with the health status, and the strength of the association may outweigh the disadvantages of clustering.

Elaborating the mixed effects model

An often overlooked assumption underlying mixed effects models is that random terms are uncorrelated with explanatory variables.[134, 137] It is unlikely that this assumption always holds in multicenter studies. For example, when we are interested in clustering of patients with a history of cancer within physicians, the history of cancer is the dependent variable (S_{ij}) in model (23) above. Complex referral patterns exist in clinical practice, which may cause patients with suspected recurrent cancers and other highly suspicious masses to be referred to specialized physicians. Physicians specialized in cancer treatment may therefore regularly encounter patients with a history of cancer, causing them to have high random intercepts (a_j) compared to other physicians. At the same time these physicians may generally encounter malignant tumors more often than other physicians, introducing a correlation between the random intercept and the health status (H_{ij}). The regression coefficient of H_{ij} (β_1), then, is a joint estimate reflecting not only the association of cancer history (S_{ij}) with tumor type (H_{ij}) at the patient level, but also the association between the prevalence of patients with a cancer history ($\bar{S}_{.j}$) with the prevalence of malignant tumors ($\bar{H}_{.j}$) that is observed when comparing physicians' practices.[134] These within and between physician associations cannot be separated, which is problematic since they can differ in strength and even have opposite signs.

The solution is to add the physician-level information on their patients' true health status ($\bar{H}_{.j}$) to the model, in order to obtain an unbiased estimate of the within-physician effect of health status.[134] When the health status is a dichotomous variable with categories indicating presence or absence of the disease, this is simply the prevalence of the disease for physician j. When health status is a variable with h categories, h-1 variables need to be constructed, each indicating the prevalence of one of the categories for physician j. When health status is a continuous variable, the physician-specific average should be included. Model (23) can be extended as follows:

(27)

$$S_{ij}=\alpha+a_j+\beta_1\times H_{ij}+\beta_2\times \bar{H}_{.j}+e_{ij}$$

$$a_j\sim N(0,\tau^2)$$

$$e_{ij} \sim N(0, \sigma^2_{error}).$$

The constant term for physician j now equals $\alpha^*_j = \alpha + a_j + \beta_2 \times \bar{H}_{.j}$ and model (27) can be rewritten as

(28)

$$S_{ij} = \alpha^*_j + \beta_1 \times H_{ij} + e_{ij}$$

$$\alpha^*_j \sim N(\bar{\alpha}^*_j, \tau^2)$$

$$e_{ij} \sim N(0, \sigma^2_{error}).$$

$\bar{\alpha}^*_j = \alpha + \beta_2 \times 1/J \times \sum_{j=1}^{J} \bar{H}_{.j}$ is the overall intercept, around which the α^*_j vary with $\tau^2 = var(\alpha^*_j) = var(a_j) + \beta_2^2 \times var(\bar{H}_{.j}) + 2\beta_2 \times cov(a_j, \bar{H}_{.j})$.

Finally, just like patient information can explain a part of the between-physician variance (i.e., τ^2), so can information at the physician level. It is possible to include physician-level explanatory variables (Z_j), such as their workplace (e.g., regional hospital versus tertiary center):

(29)

$$S_{ij} = \alpha^*_j + \beta_1 \times H_{ij} + \beta_3 \times Z_j + e_{ij}$$

$$\alpha^*_j \sim N(\bar{\alpha}^*_j, \tau^2)$$

$$e_{ij} \sim N(0, \sigma^2_{error}).$$

Note that the variance of the linear predictor, σ^2_{LP}, now equals $var(\beta_1 \times H_{ij} + \beta_3 \times Z_j)$.

Estimation

In practice, the variance terms τ^2, σ^2_{error} and σ^2_{LP} can be estimated using any standard statistical package that allows to fit mixed effects models. We have used SAS® software (version 9.3, SAS Institute, Cary, NC, USA) for all computations. A SAS macro for the computation of the RICC, based on the mixed and glimmix procedures, has been included in Appendix 5. τ^2 and σ^2_{error} error can be estimated by first estimating full model (27) to obtain $\hat{\beta}_1$, and subsequently fitting an empty random intercept model for S_{ij} (i.e., without explanatory variables) with $\hat{\beta}_1 \times H_{ij}$ as an offset variable. This way, the within-cluster effect of H_{ij} is separated from the between-cluster effect of $\bar{H}_{.j}$ in the

first step. The between-cluster effect of $\bar{H}_{.j}$ contributes to $\hat{\tau}^2$, the between-cluster variance estimated in the second step. If a physician-level explanatory variable is included, a full mixed effects model including H_{ij}, $\bar{H}_{.j}$ and the physician-level variable Z_{ij} is fitted in the first step. Next, a random intercept model for S_{ij} is fitted including $\hat{\beta}_1 \times H_{ij}$ as an offset variable, while the effect of the physician-level explanatory variable Z_{ij} is re-estimated. The resulting variance of the random intercept is the estimated variance of the random intercepts α_j^*, while the residual variance is the estimated variance of the error term, σ^2_{error}. If S_{ij} is a dichotomous or ordinal variable and generalized linear models are fitted using the logit transformation as the link function, the error terms have a logistic distribution with $\hat{\sigma}^2_{error} = \pi^2/3$.[134]

The estimation of σ^2_{LP} was chosen to correspond to the estimation of R^2 in linear and logistic models as proposed by Snijders and Bosker.[134] For a continuous S_{ij}, an empty random intercept model, that is, without fixed patient or physician level effects, was fitted first. The resulting estimated variance at physician and patient level was added together to obtain an estimate of the total variance. The explained variance (σ^2_{LP}) can subsequently be estimated by subtracting $\hat{\tau}^2$ and $\hat{\sigma}^2_{error}$ (obtained from estimating a model with explanatory variables) from the estimated total variance. In the case of an ordinal or dichotomous predictor, the linear predictor was obtained by multiplying the explanatory variables with the regression coefficient estimates from full model (27) or (29), as appropriate, and σ^2_{LP} was computed directly from the data.

Confidence statements

90% bootstrap confidence intervals were computed with the percentile method. Bootstrap resampling was performed at the physician level, including all patients for each sampled physician.[225]

Random intercept plots to identify outlying physicians

The random intercepts obtained when fitting model (27) or (29) can be plotted to identify physicians with measurements for S_{ij} that are unusually large or small, given the explanatory variables and compared to measurements by the average physician. Note that estimated random intercepts are shrunken towards zero, and that shrinkage increases as the number of patients seen by physician j decreases. Comparative standard errors are used to test whether the physician-specific random intercept is significantly different from zero.[151]

We used the false discovery rate method to account for multiple testing.[226] This methods orders p-values for multiple tests from small to large and rejects the null hypotheses for all tests with a p-value below a certain threshold, which is determined in such a way that the expected proportion of type I errors is controlled at α.

Data

We illustrate the described techniques on data from the International Ovarian Tumor Analysis (IOTA) group containing clinical and ultrasound information on patients with ovarian or tubal tumors, prospectively collected between 2002 and 2007 (phases Ib and II). We only used the data from patients for whom the physician performing the ultrasound scan was known. This resulted in a dataset of 2407 patients scanned by 40 physicians from 19 hospitals in 8 countries.

We focused on clustering at the physician level. Additional clustering of physicians in hospitals was not taken into account during the analysis because in 10 of the hospitals data was collected by only 1 physician, whereas for the other 9 hospitals there was 1 principal investigator collecting the vast majority of the data. The RICC was used to study between-physician variance for 18 variables collected during an ultrasound examination, 4 patient characteristics and the level of serum marker CA125. To account for differences in case-mix, mixed effects models were developed, with tumor histology (benign, borderline, primary invasive or metastatic invasive) as the explanatory variable. Ultrasound machine quality (high end, medium end and low end machines) was included in a second step as a physician-level characteristic to further explain differences between physicians in ultrasound measurements. Discussions with clinical experts were held to reveal the likely causes of the high interphysician variability of certain measurements. Additionally, a survey was conducted among physicians with unusually high or low random intercepts for highly clustered variables. The survey included questions on the measurement or registration of each of these variables.

5.1.3 Results

The median number of patients per physician was 9 (mean 60, IQR 2 to 82.5, range 1 to 509). 18 of 40 physicians have seen less than 5 patients, whereas 16 have seen more than 30. In total 72.1% of patients had a benign tumor, while in 5.4% of patients the tumor was borderline, in 19.4% it was primary invasive and in 3.2% it was a tumor that had metastasized to the ovary.

However, among physicians with more than 5 patients, the prevalence of benign, borderline, invasive and metastatic tumors per physician varied from 28.6 to 91.9%, 0 to 20.0%, 0 to 71.4% and 0 to 7.5%, respectively, indicating considerable case-mix differences.

The RICC showed much variation between variables (Table 16, Figure 17). Patient age had the lowest level of clustering (RICC 2.2%, 90% CI 0.9 to 3.7%). Six other variables had an RICC below 5% (the number of locules, the maximum diameter of the solid component, serum CA125, acoustic shadows, the number of papillations and the maximum lesion diameter). On the other hand, for the amount of fluid in the pouch of Douglas, 25.1% of the residual variance was due to between-physician differences (90% CI 6.7 to 32.1%). For current use of hormonal therapy this was 20.0% (90% CI 10.6 to 31.2%). Five other variables had an RICC above 15% (personal history of ovarian cancer, pelvic pain during examination, color score of intratumoral blood flow, the presence of papillations with detectable flow and the resistance index). For these seven variables with an RICC above 15%, the amount of variability accounted for by tumor type varied from 1.5% (90% CI 0.9 to 9.4%) for pelvic pain and 2.6% (90% CI 1.3 to 9.0%) for current use of hormonal therapy to 30.4% (90% CI 25.6 to 41.2%) for the presence of papillations with detectable flow and 30.9% (90% CI 24.9 to 36.3%) for the color score of intratumoral flow.

The VPC varied from 1.9% (90% CI 0.8% to 3.3%) for patient age to 21.2% (90% CI 5.1% to 28.4%) for the amount of fluid in the pouch of Douglas. The VPC and RICC were similar for most variables. However, when the tumor type accounts for a large proportion of the total variance (R^2 is high), the VPC was by definition considerably smaller than the RICC. For example, 12.6% of the residual variance of the perceived presence of metastases was due to between-physician differences, but the VPC was only 5.3%, because 57.9% of the total variance in the perceived presence metastases was accounted for by the histology of the tumor.

Taking the ultrasound machine quality into account yielded a considerable reduction in the RICC of Doppler blood flow indices such as the peak systolic velocity (reduction in RICC 4.0%, 90% CI -0.5% to 9.4%) and the resistance index (reduction in RICC 5.6%, 90% CI -1.3% to 19.2%) (Table 17). Higher quality machines gave more sensitive measurements. The explanatory power of ultrasound machine quality was moderate at best for the other variables.

Table 16. Variance partitioning of ultrasound measurements and patient characteristics.

Variable	N	CV or P	VPC	R^2	RICC
Patient age (years)	2407	0.4	1.9% [0.8% to 3.3%]	13.9% [10.6% to 17.4%]	2.2% [0.9% to 3.7%]
Number of locules (ordinal)	1997[c]	0.8	2.6% [0.5% to 4.4%]	10.6% [6.7% to 14.2%]	2.9% [0.6% to 4.9%]
Maximum diameter of the solid component (mm, log transformed)	1160[b]	0.2	2.4% [0.0% to 5.7%]	22.4% [17.4% to 28.9%]	3.1% [0.0% to 7.3%]
Serum CA125 (IU/L, log transformed)	1827[a]	0.4	2.0% [0.8% to 4.3%]	37.8% [32.8% to 41.7%]	3.3% [1.3% to 6.8%]
Acoustic shadows (yes/no)	2407	13.7%	3.0% [0.0% to 5.0%]	14.2% [8.0% to 72.1%]	3.4% [0.0% to 5.9%]
Number of papillations (ordinal)	468[d]	0.6	2.9% [0.0% to 5.5%]	15.7% [13.6% to 23.8%]	3.5% [0.0% to 6.6%]
Maximum lesion diameter (mm, log transformed)	2406[e]	0.1	4.5% [1.0% to 10.1%]	9.2% [5.9% to 11.4%]	4.9% [1.1% to 11.0%]
Presence of solid components (yes/no)	2407	48.2%	3.2% [1.0% to 4.4%]	43.5% [39.8% to 73.5%]	5.7% [2.3% to 8.4%]
Bilateral (yes/no)	2407	16.6%	7.8% [0.0% to 30.0%]	6.1% [3.6% to 8.2%]	8.3% [0.0% to 31.6%]
Ascites (yes/no)	2407	10.0%	4.9% [0.0% to 9.5%]	43.0% [34.5% to 49.3%]	8.5% [0.0% to 15.8%]
Free fluid in pouch of Douglas (yes/no)	2407	25.4%.	8.3% [2.2% to 13.3%]	14.0% [11.1% to 17.5%]	9.6% [2.6% to 15.2%]
Presence of papillations (yes/no)	2407	19.6%	8.7% [3.7% to 12.9%]	11% [7.4% to 14.9%]	9.8% [4.1% to 14.8%]
Irregular internal wall (yes/no)	2407	38.5%	10.0% [2.7% to 16.7%]	13.8% [11.6% to 16.7%]	11.6% [3.1% to 19.5%]
Peak systolic velocity (cm/s, log transformed)	1432[g]	0.3	11.2% [5.3% to 16.3%]	5.9% [2.4% to 10.8%]	11.9% [6.0% to 16.9%]
Metastases (yes/no)	1457[f]	10.7%	5.3% [0.2% to 8.4%]	57.9% [52.9% to 95.3%]	12.6% [1.3% to 21.1%]
Height of papillation (mm, log transformed)	468[d]	0.3	10.9% [0.0% to 28.2%]	16.7% [6.3% to 25.5%]	13.1% [0.0% to 31.7%]
Resistance index	1432[g]	0.3	14.5% [1.7% to 23.6%]	8.4% [4.6% to 13.5%]	15.9% [1.9% to 26.5%]
Papillations with detectable blood flow (yes/no)	468[d]	47.4%	11.3% [2.8% to 17.1%]	30.4% [25.6% to 41.2%]	16.2% [4.4% to 23.8%]
Color score of intratumoral blood flow (ordinal)	2407	0.5	12.2% [3.6% to 22.6%]	30.9% [24.9% to 36.3%]	17.6% [5.6% to 30.1%]
Pelvic pain during examination (yes/no)	2407	19.1%	18.6% [6.6% to 26.5%]	1.5% [0.9% to 9.4%]	18.9% [9.8% to 27.1%]
Personal history of ovarian cancer (yes/no)	2407	1.6%	16.5% [0.0% to 25.3%]	13.1% [10.3% to 72.8%]	19.0% [0.0% to 30.3%]
Current use of hormonal therapy (yes/no)	2407	12.7%	19.5% [10.0% to 30.4%]	2.6% [1.3% to 9.0%]	20.0% [10.6% to 31.2%]
Amount of free fluid in pouch of Douglas (mm, log transformed)	616[h]	0.3	21.2% [5.1% to 28.4%]	15.5% [9.8% to 21.7%]	25.1% [6.7% to 32.1%]

N: number of patients; CV: coefficient of variation; P: prevalence; VPC: percentage of the total variance at the sonographer level [90 CI]; R^2: percentage of the total variance explained by tumor type [90 CI]; RICC: the residual intraclass correlation [90 CI]. [a] Measurement of CA125 was not obligatory; [b] If ≥1 solid component is present; [c] If ≥1 locul; [d] In case of the presence of a papillary structure; [e] One influential outlier removed (measurement error); [f] Assessment of presence of metastases was not obligatory; [g] If there is intratumoural blood flow; [h] If there is fluid in the pouch of Douglas.

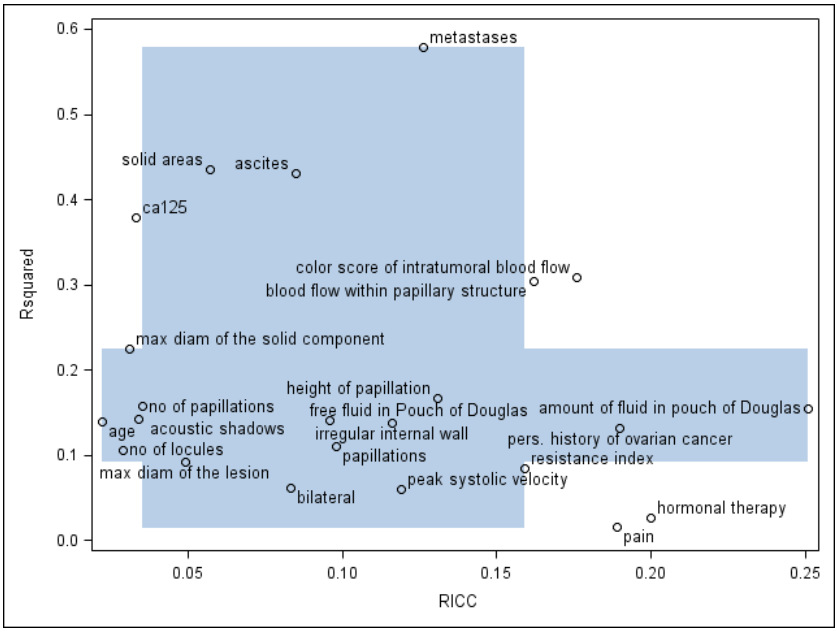


Figure 17. The residual intraclass correlation (RICC) versus the proportion of explained variance (R^2). Dark bands indicate the interquartile ranges.

To detect which physicians have systematically higher or lower measurements than the average physician, random intercept plots were constructed showing the physician's central tendency (mean or proportion) on the x-axis and the random intercept on the y-axis (Figure 18). There was considerable interphysician variability in the measurements of the amount of fluid in the pouch of Douglas, with average measurements per physician ranging from 10 mm to 45 mm. Taking patient case-mix into account, two physicians were identified as outliers with lower measurements than the average physician. Physician 5 positioned his patients horizontally on a flat bench while measuring the fluid, which could explain the low measurements. Physician 31 used an examination table with stirrups, on which the patient was placed in the supine position. However, the gradient of the patient's trunk, which is most likely to influence the measured amount of fluid in the pelvis, is unknown. Hence, we could not determine why this physician obtained remarkably low measurements.

Table 17. Effect of ultrasound machine quality on the residual intraclass correlation (RICC).

Variable	N	Reduction in RICC
Number of locules (ordinal)	1997[c]	0.3% [-0.1% to 1.2%]
Maximum diameter of the solid component (mm, log transformed)	1160[b]	-0.5% [-1.1% to 1.1%]
Acoustic shadows (yes/no)	2407	-0.4% [-3.6% to 0.9%]
Number of papillations (ordinal)	468[d]	3.1% [-0.0% to 6.2%]
Maximum lesion diameter (mm, log transformed)	2406[e]	0.2% [-0.6% to 2.8%]
Presence of solid components (yes/no)	2407	0.7% [-0.3% to 2.6%]
Bilateral (yes/no)	2407	-1.2% [-3.7% to 2.1%]
Ascites (yes/no)	2407	3.6% [-0.8% to 10.8%]
Free fluid in pouch of Douglas (yes/no)	2407	0.6% [-0.5% to 3.1%]
Presence of papillations (yes/no)	2407	2.5% [0.2% to 6.3%]
Irregular internal wall (yes/no)	2407	1.0% [-0.5% to 4.1%]
Peak systolic velocity (cm/s, log transformed)	1432[g]	4.0% [-0.5% to 9.4%]
Metastases (yes/no)	1457[f]	0.9% [-1.3% to 6.5%]
Height of papillation (mm, log transformed)	468[d]	0.1% [-2.9% to 8.8%]
Resistance index	1432[g]	5.6% [-1.3% to 19.2%]
Papillations with detectable blood flow (yes/no)	468[d]	2.7% [-1.5% to 12.8%]
Color score of intratumoral blood flow (ordinal)	2407	-0.4% [-2.2% to 1.5%]
Amount of free fluid in pouch of Douglas (mm, log transformed)	616[h]	0.8% [-0.9% to 3.9%]

Difference in the residual intraclass correlation (RICC [90% CI]. [b] If at least one solid component is present; [c] If there is at least one locul; [d] In case of the presence of a papillary structure; [e] One influential outlier removed (measurement error); [f] Assessment of presence of metastases was not obligatory; [g] If there is intratumoral blood flow; [h] If there is fluid in the pouch of Douglas.

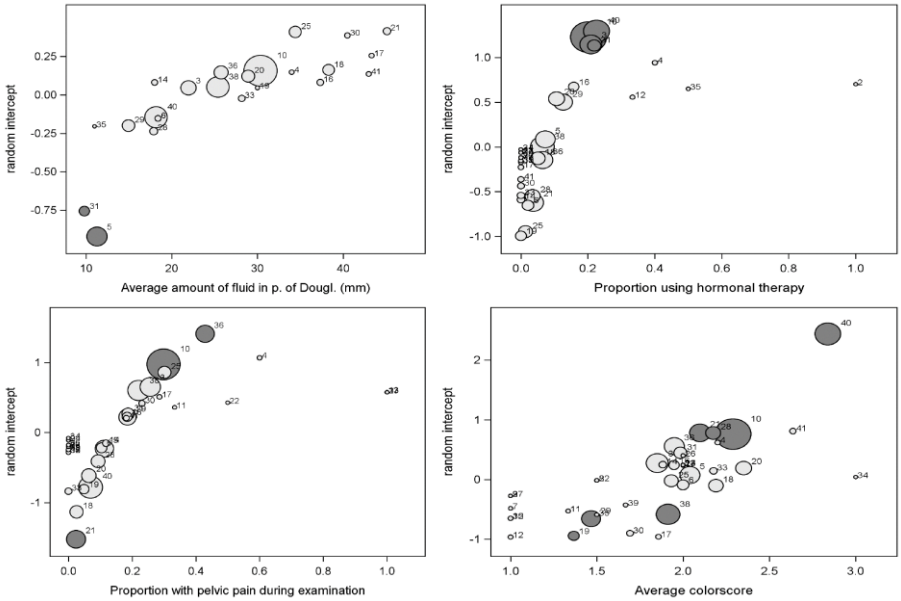


Figure 18. Average measurement or prevalence per physician versus the random intercept. From left to right and from top to bottom: amount of free fluid in the pouch of Douglas (mm), current use of hormonal therapy (yes/no), pelvic pain during examination (yes/no), color score of intratumoral blood flow (ordinal, 1 no blood flow, 2 minimal blood flow, 3 moderate blood flow, 4 very strong blood flow). X-axis: average measurement or prevalence per physician. Bubble size represents the number of patients per physician. Dark bubbles: physicians with a random intercept different from zero (adjusted p-value < 0.05 after correction for multiple testing using the false discovery rate method).

The percentage of patients using hormonal therapy at the time of the ultrasound examination was between zero and ten percent for most physicians (Figure 18). Given the patient case-mix, four physicians were identified as outliers with higher rates of hormonal therapy use among their patients than among the average physician's, as indicated by the positive random intercepts. Three of them were from centers in Belgium and Sweden, both countries in which hormonal therapy is relatively frequently prescribed. The survey among physicians revealed that the outliers had given their patients various examples when asking them about hormonal therapy use, which could have contributed to the large numbers of patients with hormonal therapy use that they registered. The survey further revealed that physicians do not fully agree on what they consider to be hormonal therapy.

The interphysician variability in patients' experience of pelvic pain during examination could not be explained by the physicians' examination style, the type of probe that was used, or the prevalence of endometriomas or abscesses, which are known to be more painful than other tumors, especially if pressure is applied on these masses. It is likely that the registration of pain was subject to the empathy of the physician on the one hand, and the pain threshold and inclination of the patient to express pain on the other. The latter may vary from person to person and may also be dependent om the country.[223]

Eight physicians were detected as outliers with respect to their measurements of the color score of intratumoral blood flow, five with high and three with low values, which may partly be explained by the use of color or power Doppler ultrasonography by different examiners. The survey among outlying physicians included five images of ovarian masses, which had to be rated. People with a tendency to give high scores to the survey images also tended to have higher random intercepts, and vice versa, indicating the subjectivity of color scores.

5.1.4 Discussion

This study described an innovative use of the RICC as a practical tool to screen variables for clustering in multicenter studies. The RICC expresses the degree of clustering as a percentage of the residual variance. Mixed models, which allow control for patient- and cluster-level fixed effects, are used to estimate three parts of the total variance: the variance explained by the fixed effects (R^2), the residual variance at the patient level, and the residual variance at the cluster level.

It may be useful to take R^2 into account when evaluating the degree of clustering, in a fashion that is comparable to a cost-benefit analysis (Figure 17). The 'cost' is the degree of clustering (i.e., the RICC) which complicates statistical analysis and may necessitate efforts to detect and remove causes of clustering. The 'benefit' is the extent to which a variable is related to the fixed effects, for example, the disease of interest. A variable with a high RICC but low R^2 may not be worth the effort and may be excluded from further data collection or analysis. For a variable with a high RICC and R^2, such as the color score of intratumoral blood flow in our case study, it can be worthwhile to investigate the cause of the clustering.

Applications

The proposed methods may be used for quality control when data collection in a large multicenter study is ongoing. In this way, problems with data clustering can be identified and remedied in a timely fashion. Another application is in model development, for example, of risk models.[6] Some researchers state that the interrater reliability of potential predictors should be assessed before considering their inclusion in a risk model.[27] Others argue that this is superfluous because the effects of unreliable measurements in multivariable models will be diluted. In our opinion, it is preferable to screen for various forms of data clustering as part of the data analysis in a multicenter study. This is discussed in more detail in the next section. The case study on clustering in predictors of ovarian cancer malignancy showed that clustering may occur even if efforts were made to clearly define variables and measurement protocols.[178] The results were used to inform variable selection for the subsequent development of the ADNEX model for the preoperative diagnosis of ovarian cancer.[8]

Note that inspecting R^2 while using the RICC to screen for clustering can provoke univariable predictor selection. This is generally not recommended, especially when the dataset available for model building is limited in size.[5] In our case study, there were 672 malignant tumors for 23 potential predictors, resulting in nearly 30 events per potential predictor, whereas guidelines propose 10 to 50 events per variable depending on the situation.[50, 52, 53, 212] We advise to rely mainly on prior expert knowledge and multivariate selection procedures. Nonetheless, R^2 can play a role if the sample size is sufficiently large and if it is incorporated in a carefully designed selection procedure. When no data-driven variable selection is performed, inspecting R^2 is not problematic. In randomized trials, the treatment effect should be taken into account when studying clustering, unless clustering of baseline measurements is investigated. Therefore, the proposed methodology can be applied during an interim analysis of the treatment effect. Recent studies have stressed the importance of acknowledging clustering in randomized controlled trials.[227]

Strengths and weaknesses

A strength of the proposed approach is that it can be applied to any regular multicenter dataset, as it does not require measurements by multiple physicians for the same patient. In addition, various forms of data clustering are captured, including subjectivity of measurements, differences in measurement equipment or settings, deviations from the measurement protocol, sociocultural characteristics, and differences in patient populations

due to, for example, local referral patterns. The result is that the cause of the clustering is not immediately clear. In our survey, the amount of time that has passed between data collection and the survey on measurements (six years) complicated the problem further and potentially yielded recall bias. Nonetheless, in order to alleviate clustering it is imperative to investigate the causes of clustering. For example, when variables are subjective or the measurement protocol is unclear, providing training or protocol adjustments may help. It is also useful to detect clustering caused by differences in the populations seen by physicians (e.g., due to local referral patterns), since it may not be possible to build a generalizable risk model when such heterogeneity is too large.

Note that RICC only detects systematic differences in measurements between physicians. In the case of subjectivity of measurements, random intercepts will only indicate which physicians give consistently higher or lower scores than the average physician. Random (within physician) variability will contribute to the residual variance at the patient level. To study non-systematic differences, an interrater reliability study is needed in which two or more physicians investigate the same patients, in order to detect whether physicians would give the same scores to the same patients.[221, 222]

There are two methodological difficulties associated with our approach. First, a relatively large amount of data is needed to reliably compute the variance at the cluster level. The amount of variance at the cluster level will often be underestimated, and this bias increases as the number of clusters decreases.[146, 147, 215] The problem is most pronounced for categorical variables. This also explains the observed width of the bootstrap confidence intervals in the case study. Second, as in provider profiling studies,[228] it is not always clear how patient case-mix should be taken into account. In this study, tumor type was categorized into four groups (benign, borderline, invasive cancer, metastatic cancer), but a more general or more detailed categorization could have been chosen as well. This decision should involve experts' opinions. Relevant categories should be taken into account, but the number of categories must be low enough to guarantee reliable estimates. For our case study, it is known that pelvic pain during examination might be worse for endometriomas and abscesses. Therefore, we could have used endometriomas and abscesses as a fifth category. Doing so marginally increased the RICC from 18.9% to 19.4%. This implies that caution must be taken when comparing the RICC across studies, as results may vary depending on variable definitions or the choice of fixed effect factors. For time-to-event outcomes, the health status may be accounted for by using the cumulative baseline hazard at the event or censoring time and an event indicator as fixed effects in the mixed model,

analogous to a suggested approach for missing data imputation in the time-to-event setting.[229]

Conclusions

Although performing multicenter studies enhances the generalizability of results, we recommend that the clustered nature of collected data is acknowledged and investigated. The RICC is a useful tool that expresses the degree of clustering as a percentage. An advantage of the RICC is that it does not require repeated measurements on the same patients by various physicians. The observed degree of clustering may be decreased by adjusting the measurement protocol and providing training to physicians. Potential predictors with severe clustering may be excluded when building a risk model.

5.2 Statistical variable selection in clustered data

Stepwise variable selection procedures such as backward and forward selection rely on statistical testing to retain only significant candidate predictors. Backward selection starts with a full model, removing insignificant predictors one by one, beginning with the least significant predictor. Alternatively, forward selection starts with an empty model and enters significant predictors one by one, starting with the most significant predictor. Programs to perform variable selection for mixed effects models are not available in standard statistical software packages. Researchers might therefore use the programs for non-clustered data to perform variable selection and test after variable selection whether random effects are needed. This study considers the effect of ignoring clustering during stepwise variable selection when building a logistic regression to predict the malignancy of ovarian tumors.

5.2.1 Background

Standard and mixed effects logistic regression yield different estimated models. Regression coefficients from the standard model are marginal estimates that represent the average effects in the population of patients, while coefficients from the random intercept model are conditional estimates that represent the effects in an average center. Marginal effect are typically closer to zero than conditional effect estimates.[138, 139, 230] Using a cumulative

Gaussian approximation to the logistic function leads to the following approximation, with τ^2 the variance of the random intercept distribution:[139]

(30)

$$\alpha_{marginal} \approx \alpha_{conditional}/f$$

$$\beta_{marginal} \approx \beta_{conditional}/f,$$

$$\text{with } f = \sqrt{1+\tau^2 c^2}$$

$$\text{and } c = \frac{16\sqrt{3}}{15\pi}.$$

Moreover, standard logistic regression models assume independent observations, but this assumption does not hold in clustered data. As a consequence, standard errors and p-values cannot be trusted.[134, 136, 137] This may be especially problematic when clustering is ignored during statistical variable selection, as variables are omitted or included in each step of the procedure based on significance tests. We expect that a different variable subset is selected if clustering is taken into account in every step of the selection procedure by adding a random cluster intercept.

More advanced variable selection procedures based on statistical testing, such as the multivariable fractional polynomial (MFP) algorithm, are expected to suffer from the same problem. The MFP ra2 algorithm starts with a full model, assuming linear effects for all continuous predictors. A closed test procedure based on deviance tests is built into an iterative backward variable selection algorithm to select the best transformation of continuous variables.[57] The considered transformations are of the form $x^{p1}+x^{p2}$, where p1 and p2 are chosen from the set -2, -1, -0.5, 0, 0.5, 1, 2 and 3, with x^0 defined as the logarithmic transformation. Within each cycle of the algorithm, the significance of excluded predictors is re-evaluated. The algorithm stops when convergence has been reached.

5.2.2 Methods

This case study is about the prediction of the probability of malignancy of ovarian masses, using data on 19 predictor variables from 3511 patients recruited at 21 centers during IOTA phases I, Ib and II. Standard logistic regression and random intercept logistic regression were compared. We first fitted the full models for both methods to compare model coefficients. Then,

backward variable elimination was performed (with α=0.01), using either standard or random intercept logistic regression in each step of the procedure of the MFP ra2 algorithm.

Additionally, bootstrap resampling (100 samples) was used to compare selection frequencies of variables with and without the use of random intercepts. The bootstrap samples were used to compute the optimism-corrected (standard or "population level") c-statistic, sensitivity, specificity, calibration intercept and calibration slope. The calibration slope was computed using equation (18), and indicates whether there is over- (β_{cal}<1) or underfitting (β_{cal}>1). The calibration intercept was assessed by including the linear predictor (LP_{ij}) as an offset in equation (18), instead of estimating its effect:

(31)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+LP_{ij}.$$

If the calibration intercept (α_{cal}) is larger than zero, the predicted probabilities are on average too low; if calibration intercept (α_{cal}) is smaller than zero, the predicted probabilities are on average too high.

We used predictions for the average center, omitting the random intercept (see equation (11)), to be able to make predictions in clusters that were not represented in the bootstrap sample.[142, 151]

5.2.3 Results

The model coefficients of the full models fitted by standard and random intercept logistic regression differed (Table 18), while the standard errors of regression coefficients were larger when random intercepts were used. For 11 predictors, the coefficients from the random intercept model were further from zero than the coefficients of the standard logistic regression model. Multicollinearity among predictors is the reason that the approximations given by equation (30) do not hold exactly. Differences between effects are smaller when population-averaged effects are considered for the random intercepts logistic regression model (see Figure 19 for an example). These are obtained by averaging over center-specific random intercepts drawn from the estimated random effects distribution.

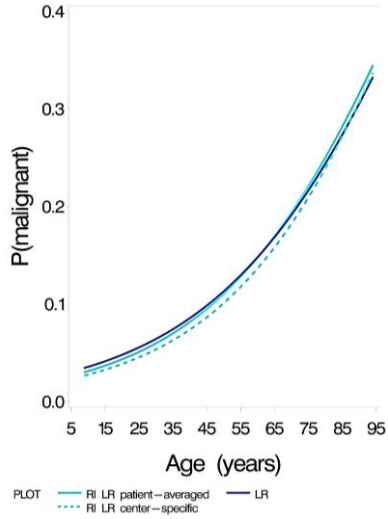


Figure 19. The marginal effect of age estimated by the standard logistic regression model, the marginalized effect of the random intercept model and effect in the average center estimated by the random intercept model.

The MFP method selected the same transformations when a random intercept was included as when clustering was ignored. The current use of hormonal therapy was not included as a predictor in the random intercept model, while it was included in the standard model. Both procedures selected a logarithmic transformation of the maximum diameter of the largest solid component.

Variable selection frequencies in the backward selection procedure were higher for a number of variables when clustering was not accounted for (Figure 20). The relative selection frequency differed by more than 10% for five variables, with the highest difference being 64% for the current use of hormonal therapy. The optimism-corrected performance was similar for the final standard and random intercept logistic regression models. The c-statistic was 0.94 (95% CI 0.93 to 0.95) for both. The standard logistic regression model had a slightly better sensitivity (95% versus 94%) while the random intercept logistic regression model had a slightly better specificity (75% versus 74%). The calibration slope was 0.97 for the standard logistic regression model and 0.95 for the random intercept logistic regression model. The reason for this discrepancy will be studied in chapter 7. The calibration intercept was <0.001 for both models.

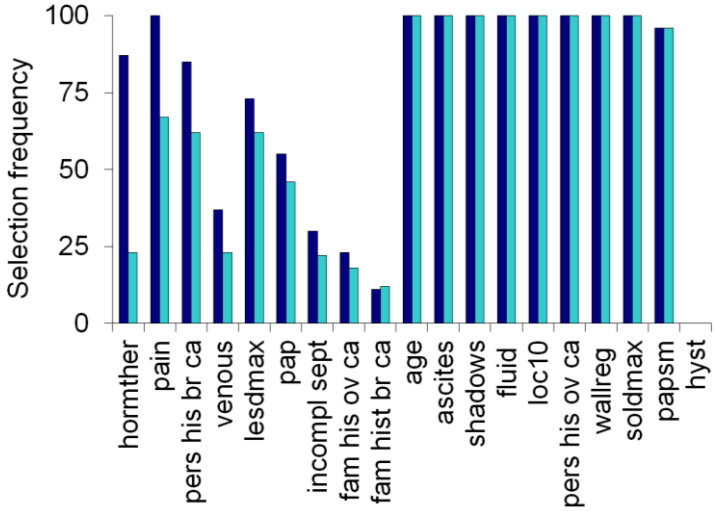


Figure 20. Variable selection frequencies in 100 bootstrap samples, using backward standard logistic regression (dark bars) and backward random intercept logistic regression (light bars).

5.2.4 Discussion

Using random intercepts in each step of the stepwise variable selection procedure yielded conservative results, with higher standard errors and lower selection frequencies than when clustering was ignored. Whereas multicenter data are beneficial for model generalizability, the clustering in centers must be accounted for when developing risk models. The predictive performance of the random intercept and standard logistic regression model were similar. The predictive performance of random intercept models will be further investigated in the next chapters.

Stepwise procedures have been criticized before. The set of chosen predictors is unstable, changing when another sample is considered. The regression coefficients are biased away from zero, due to the fact that only the effects that withstood statistical testing are retained, and standard errors and p-values are estimated as if the model was specified a priori.[5, 6, 44, 45, 52, 55] Nonetheless, stepwise selection remains widespread. This case study demonstrates that

Table 18. Estimated regression coefficients for the full standard and random intercept (RI) logistic regression (LR) models.

Model	Full Standard		Full Random intercept	
Effect	**Estimate**	**s.e.**	**Estimate**	**s.e.**
Intercept	-6.434	0.346	-6.656	0.384
Age (in years)	0.031	0.004	0.034	0.004
Fluid in the pouch of Douglas*	0.173	0.031	0.178	0.033
Max diam of largest solid component*	0.384	0.020	0.385	0.020
Max diameter of the lesion*	0.086	0.025	0.077	0.026
Family history of ovarian cancer	0.572	0.362	0.592	0.364
Family history of breast cancer	0.194	0.192	0.225	0.198
Personal history of ovarian cancer	2.333	0.357	2.322	0.365
Personal history of breast cancer	0.977	0.249	0.820	0.255
Hysterectomy	0.032	0.217	0.056	0.224
Current use of hormonal therapy	-0.620	0.180	-0.357	0.189
Pain during examination	-0.708	0.165	-0.549	0.171
Ascites	1.761	0.228	1.769	0.237
Incomplete septum	-0.408	0.223	-0.371	0.236
Irregular papillary projection	0.986	0.215	1.023	0.222
Irregular internal cyst wall	1.165	0.161	1.165	0.167
Acoustic shadows	-2.038	0.250	-2.038	0.260
Venous blood flow only	-0.464	0.228	-0.416	0.241
Presence of papillary structures	-0.621	0.223	-0.574	0.229
More than 10 locules	1.198	0.226	1.315	0.232
τ^2	/	/	0.325	0.144

*in mm, transformed as $\sqrt{\text{predictor}+1}$; s.e.: standard error.

stepwise variable selection in clustered data is even more problematic when clustering is ignored.

In general, the use of expert knowledge for predictor selection is preferable to the use of statistical variable selection. In some cases, expert knowledge may not be sufficient, and statistical selection may be needed. An important topic for future research is the inclusion of random slopes in risk models. It is generally advised to be parsimonious and careful when considering interactions in prediction research, as they are hard to replicate and do not tend to improve predictions much. [5, 6, 59] The question arises whether center-by-predictor interactions ("random slopes") should be given a different status. In any case, a strategy to deal with random slopes should be defined a priori. Random slopes may be considered only for predictors for which heterogeneous effects may reasonably expected, for instance when different measurement equipment is used or when the measurement is highly subjective. If unsure, one could select the variables that are expected to have the strongest predictive effects. The random slopes of various effects can be tested simultaneously.[231] A remaining question is where in the model building process random slopes should be investigated: before or after the selection of main effects, interactions between predictors and transformations of linear variables?

5.3 Conclusion

Measurements may systematically vary from one center to the next. The ICC is a tool to detect systematic differences in measured values in multicenter datasets. Predictors subject to such differences may be excluded from a risk model. If statistical variable selection is required, the clustered nature of the data should be taken into account at every step of the variable selection algorithm in order to avoid incorrect inferences.

6 Performance evaluation in multicenter studies

This chapter discusses methods to evaluate the predictive performance of risk and classification models in multicenter validation studies. Section 6.1 discusses common causes of heterogeneity in predictive performance. Section 6.2 discusses techniques that have been proposed to acknowledge heterogeneity for some well-known performance statistics. Section 6.3 offers examples on the evaluation of diagnostic models for ovarian cancer. Section 6.4 investigates novel methods to summarize clinical utility in multicenter studies. Section 6.5 concludes.

This chapter is based on the following research:

Testa A, Kaijser J, Wynants L, et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. *Br J Cancer*. 2014; 111: 680-8.

Timmerman D, Van Calster B, Testa A, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis (IOTA) group. *Am J Obstet Gynecol*. 2016; 214: 424-37.

Sayasneh A, Wynants L, Preisler J, et al. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. *Br J Cancer*. 2013; 108: 2448-54.

Wynants L, Timmerman D, Verbakel JY, et al. Benefit of ovarian cancer diagnostic models: external validation of five models using decision curve analysis. 2016, in preparation.

Wynants L, Van Huffel S and Van Calster B. Evaluating the clinical utility of prediction models in a multicenter population using decision-analytic measures. Conference of the International Society for Clinical Biostatics. Utrecht, the Netherlands 2015.

Wynants L, Timmerman D, Van Huffel S and Van Calster B. Evaluating the Clinical Utility of Prediction Models in a Heterogeneous Multicenter Population Using Decision-Analytic Measures: The Random Effect Weighted Net Benefit. 36th Annual North American Meeting of the Society of Medical Decision Making. Miami, Florida 2014.

6.1 Heterogeneity in predictive performance

The validity of a risk model in a new setting depends on various factors, including overfitting of the risk model, characteristics of the patient population, true differences in predictor effects between the development and validation setting, and differences in the definitions of predictors or the outcome. Simulation studies have investigated factors that systematically influence discrimination, calibration and clinical utility.[6, 98, 232]

The c-statistic is influenced by the heterogeneity in the patient case-mix in a particular setting. If predictors show less variability at model validation than at model development (e.g., all patients are approximately the same age), the c-statistic will be lower, while more heterogeneous predictors (e.g., large age differences) typically yield higher c-statistics. In addition, low calibration slopes because of overfitting and suboptimal discrimination often go hand in hand.[5, 79]

Calibration may be affected by case-mix differences and invalid regression coefficients. If a predictor that is missing from the risk model yields a more severe case-mix at validation than at model development (e.g., more patients with comorbidities), the predicted risks are on average too low. In case the missing predictor yields a less severe case-mix, the predicted risks are on average too high. The missing predictor may be a center-level characteristic that is related to the outcome, for example, the specialization of the center, as specialized oncology centers may have more patients with malignant tumors than regional centers. The missing predictor may also be an unobservable variable influencing the outcome, leading to under- or overpredicted risks. For example, some centers may have many difficult cases referred to them, but it may not be possible to observe what makes a case "difficult". Lastly, calibration slopes may be lower than one due to overfitting in the development set or because of truly different predictive effects.[5, 78]

Clinical utility (as measured by NB) is influenced by a complex interplay of calibration, discrimination, the prevalence of the outcome and the risk threshold at which the clinical utility is evaluated. In general, miscalibration always reduces the NB, but the impact of miscalibration is reduced when the model discriminates well between the outcomes.[6, 232] Overfitted models and models under- or overestimating risks may be harmful, that is, have a lower NB than the default strategies of treating all or none of the patients.[6, 232] Measures of clinical utility are also influenced by the prevalence of the outcome,[233] which is most pronounced at low risk thresholds. It may be intuitively clear that in a high-prevalence population, treating all patients may

be equivalent or preferable to using a risk model at low risk thresholds that imply that the benefit of a true positive outweighs the harm of a false positive.

The previous paragraphs explained why the performance of a model at external validation may differ from the performance at internal validation. For the same reasons, the performance of a model may differ between various external validation settings. In chapter 3, it was explained that multicenter studies are often used to cover a broad range of clinically interesting settings, in order to increase the generalizability of the results. Moreover, small centers give rise to uncertainty in center-specific performance estimations. Hence, in multicenter validation studies, discrimination, calibration and clinical utility may be expected to differ from one center to another.

6.2 Performance measures for multicenter validation

To exploit all information that is available in multicenter validation studies, performance measures should allow the user to inspect the heterogeneity in performance across centers, and deal in a proper way with between-center differences in performance and within-center sampling variability, for example due to small center sizes. The within-cluster c-statistic proposed by Van Oirbeek (see equation (19)) and the within-cluster calibration slope (see equation (20)) were already introduced in chapter 4. This section discusses other methods that have originally been proposed to summarize predictive performance across multiple validation studies in a meta-analysis, or to evaluate performance in relevant subgroups. The focus is restricted to the commonly used metrics for the validation of risk models that were introduced in section 2.1.6.

6.2.1 Sensitivity and specificity

Suppose J (j=1 to J) centers are available, including n_{1j} and n_{0j} patients with and without the disease of interest, respectively. In center j, the model under validation yields a positive classification for r_{11j} patients with the disease (true positives), and a negative classification for r_{00j} patients without the disease (true negatives). The observed sensitivity in each center is r_{11j}/n_{1j} and the observed specificity is r_{00j}/n_{0j}. A bivariate random effects meta-analysis can simultaneously model sensitivity and specificity:[234, 235]

(32)

$$r_{11j} \sim \text{bin}(n_{1j}, Se_j)$$

$$r_{00j} \sim \text{bin}(n_{0j}, Sp_j)$$

$$\begin{pmatrix} \text{logit}(Se_j) \\ \text{logit}(Sp_j) \end{pmatrix} \sim N\left[\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}, \boldsymbol{\Omega}\right], \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$

This model specifies a binomial distribution for the number of true positives and true negatives. The true center-specific logit sensitivities and specificities are assumed to be normally distributed with mean γ_1 and γ_2 and variance-covariance matrix $\boldsymbol{\Omega}$. Hence, there is between-center heterogeneity in the logit sensitivity (τ_1^2) and the logit specificity (τ_2^2), and a correlation (ρ_{12}) between the logit sensitivity and the logit specificity. This model can be fitted with frequentist or Bayesian methods. A summary sensitivity and specificity can be obtained by applying the inverse logit transformation to the estimated $\hat{\gamma}$ ($\hat{\gamma}_1$ or $\hat{\gamma}_2$, as appropriate): $\exp(\hat{\gamma})/(1+\exp(\hat{\gamma}))$. However, a single summary statistic does not reflect the heterogeneity in discrimination that may be expected in new centers. In the frequentist framework, a 95% prediction interval of the logit sensitivity or specificity in a new center can be obtained as:[234]

(33)

$$\hat{\gamma} - t_{J-2}\sqrt{\hat{\tau}^2 + \text{var}(\hat{\gamma})},\ \hat{\gamma} - t_{J-2}\sqrt{\hat{\tau}^2 + \text{var}(\hat{\gamma})},$$

where $\text{var}(\hat{\gamma})$ is the variance of $\hat{\gamma}$ and t_{J-2} is the $100(1-\alpha/2)^{th}$ percentile of the t-distribution with J-2 degrees of freedom, with $\alpha=0.05$ for a 95% prediction interval. In case a Bayesian approach is chosen, probability statements for the sensitivity and specificity in a new center, such as the probability that the true sensitivity and specificity will both be above some clinically acceptable values, can be deduced from their posterior distribution. This requires the specification of prior distributions for γ_1, γ_2 and $\boldsymbol{\Omega}$.[234]

Sensitivity and specificity are often thought of as independent of the prevalence of the disease, but research has shown that this does not always hold true.[63] Therefore, Chu and colleagues[236] have proposed a trivariate random effects model that includes the disease prevalence p_j in center j:

(34)

$$n_{1j} \sim \text{bin}(n_j, p_j)$$

$$r_{11j} \sim \text{bin}(n_{1j}, Se_j)$$

$$r_{00j} \sim \text{bin}(n_{0j}, Sp_j)$$

$$\begin{pmatrix} \text{logit}(Se_j) \\ \text{logit}(Sp_j) \\ \text{logit}(p_j) \end{pmatrix} \sim N\left[\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}, \boldsymbol{\Omega}\right], \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \rho_{13}\tau_1\tau_3 \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 & \rho_{23}\tau_2\tau_3 \\ \rho_{13}\tau_1\tau_3 & \rho_{23}\tau_2\tau_3 & \tau_3^2 \end{pmatrix}.$$

In practice, it may be difficult to estimate the correlations in this model reliably.[234] Another approach may be to include the center-specific prevalence as a covariate (see section 6.2.5). Alternative parametrizations of models (32) and (34) in terms of the positive and negative predictive values instead of sensitivity and specificity are possible as well.[234, 236]

6.2.2 The c-statistic

Van Oirbeek suggested to decompose Harrell's c-statistic into different parts to evaluate the predictive performance of a model on the between-cluster and the within-cluster level separately.[5, 217] The within-cluster c-statistic proposed by Van Oirbeek was introduced in chapter 4 (see equation (19)). This measure is actually the average cluster-specific c-statistic, with cluster-specific c-statistics weighted by the number of pairs of events and non-events per cluster. Several other weights can be used to obtain a weighted average, among which random effects weights have been identified as particularly interesting, as they allow to quantify and inspect the heterogeneity in discriminatory performance across centers.[218, 234, 237] Since the distribution of the c-statistic across centers may deviate from normality, it is advisable to perform the analysis on the logit scale.[218, 238] A univariate random effects meta-analysis model of center-specific c-statistics (C_j) on the logit scale looks as follows:

(35)

$$\text{logit}(C_j) \sim N(\theta_j, \sigma_j^2)$$

$$\theta_j \sim N(\gamma, \tau^2).$$

The logit of the estimated c-statistic in each center (C_j) is assumed to have a normal sampling distribution of which the variance (σ_j^2) is assumed to be known. θ_j represents the true logit c-statistic for center j, which is assumed to be drawn from a normal distribution with mean γ and between-center variance τ^2. Hence, the summary c-statistic can be obtained as $\exp(\hat{\gamma})/(1+\exp(\hat{\gamma}))$. In the

frequentist framework, the 95% prediction interval is obtained with equation (33). In a Bayesian framework, using prior distributions for γ and τ^2, probability statements about the c-statistic in a new center can be obtained from the posterior distribution.

6.2.3 Calibration

A one-stage random effects analysis of calibration in multiple centers was introduced in chapter 4 (see equation (20)).[152] In this case, the patient data from all individual centers is pooled in one large dataset, and a random effects logistic calibration model is estimated to calculate the calibration slope in the average center as well as the between-center variance in calibration slopes. Calibration-in-the-large can be assessed by computing the calibration intercept in the same framework, including LP_{ij} as an offset in the random effects logistic calibration model, and allowing the center-specific intercepts to vary:[152]

(36)

$$\log\left(\frac{p_{ij}}{1\text{-}p_{ij}}\right)=\alpha_{cal\,w}+a_{j\,cal}+LP_{ij}$$

$$\left(a_{j\,cal}\right)\sim N\left(0,\tau_a^2\right).$$

$\alpha_{cal\,w}$ now is the average within-center calibration intercept, while τ_a^2 is the variance of the within-center calibration intercepts $a_{j\,cal}$.

An alternative to the one-stage meta-analysis described above is a two-stage meta-analysis of center-specific calibration slopes (or intercepts).[237] In the first stage, calibration slopes (or intercepts) are obtained in each center separately (using logistic calibration, see equation (18)). In the second stage, a summary estimate and prediction interval are obtained by performing a univariate random effects meta-analysis of the calibration intercepts (or slopes). A model analogous to model (35) can be used for this purpose, but a logit transformation of calibration slopes (or intercepts) is not required.[238]

Snell and colleagues proposed a multivariate approach for the meta-analysis of performance measures. They suggest to jointly synthesize calibration and discrimination, while accounting for the correlation of performance measures.[143] Let $\text{logit}(C_j)$ and $\beta_{cal\,j}$ be the observed logit c-statistic and calibration slope in center j.

(37)

$$\begin{pmatrix} \text{logit}(C_j) \\ \beta_{\text{cal j}} \end{pmatrix} \sim N\left[\begin{pmatrix} \theta_{j1} \\ \theta_{j2} \end{pmatrix}, \mathbf{\Sigma_j}\right], \mathbf{\Sigma_j} = \begin{pmatrix} \sigma_{j1}^2 & \rho_{Wj12}\sigma_{j1}\sigma_{j2} \\ \rho_{Wj12}\sigma_{j1}\sigma_{j2} & \sigma_{j2}^2 \end{pmatrix}$$

$$\begin{pmatrix} \theta_{j1} \\ \theta_{j2} \end{pmatrix} \sim N\left[\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}, \mathbf{\Omega}\right], \mathbf{\Omega} = \begin{pmatrix} \tau_1^2 & \rho_{B12}\tau_1\tau_2 \\ \rho_{B12}\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$

The observed values come from a bivariate normal distribution with θ_{j1} and θ_{j2} the underlying true logit c-statistic and calibration slope for center j, and $\mathbf{\Sigma_j}$ the within-center variance-covariance matrix for center j, which can be obtained by bootstrapping. γ_1 and γ_2 are the means of the measures of interest, τ_1^2 and τ_2^2 the between-center variances and ρ_{B12} the between-center correlation induced by differences in populations and settings. Using a bivariate t-distribution with J-2 degrees of freedom instead of a univariate t-distribution in equation (33), a 95% joint prediction region can be obtained for the c-statistic and calibration slope. Alternatively, a Bayesian approach can be used to make probability statements for new centers.

An alternative approach is the graphical presentation of a calibration plot with separately drawn curves for large centers. The advantages are that a higher level of calibration can be assessed if flexible calibration curves are used (moderate calibration rather than weak calibration, see section 2.1.6),[64] and that the heterogeneity is visualized. A disadvantage of this approach is that it does not distinguish between random sampling variability within the centers and between-center heterogeneity. The flexible calibration curves commonly used in calibration plots are known to deviate from the diagonal due to random variance. Rules of thumb suggest to have at least 100 events for external validation,[95, 96] while 200 events and 200 non-events are needed to reliably assess a flexible calibration curve.[64] Therefore, differences in calibration between centers should be carefully interpreted and the focus needs to be on large centers. An alternative may be to visualize center-specific logistic calibration curves, as they may require less data.

6.2.4 Net benefit

Although heterogeneity in the NB is to be expected given heterogeneity in discrimination, calibration and prevalence, no summary measures for NB have been proposed so far. A one-stage approach combining the datasets of all centers and computing the NB as if no clustering was present is equivalent to

computing a weighted average of the center-specific NB, using the number of observations in each center as weights. Kerr and colleagues propose to draw separate NB curves for subpopulations, especially when they have different predicted risk distributions.[239] Following this reasoning, separate curves may be drawn for centers of a different type (e.g., tertiary referral centers versus regional centers), or center-specific curves for large centers may be investigated to grasp the heterogeneity between centers. Section 6.4 will explore methods for a meta-analysis of center-specific NB.

6.2.5 Explaining heterogeneity

One can attempt to explain differences in performance between centers using variables at the center level, such as center characteristics or aggregated patient-level variables (e.g., the average patient age or the variance in patient age). This may help to identify centers in which the model has a good predictive performance. Performance statistics and plots can be produced for the different subgroups defined by the explanatory variable. Alternatively, a meta-regression can be performed by introducing a covariate to the meta-analysis model.[143] For example, a meta-regression for sensitivity and specificity, accounting for a center-level characteristic Z_j would look as follows:

(38)

$$r_{11j} \sim \text{bin}(n_{1j}, Se_j)$$

$$r_{00j} \sim \text{bin}(n_{0j}, Sp_j)$$

$$\begin{pmatrix} \text{logit}(Se_j) \\ \text{logit}(Sp_j) \end{pmatrix} \sim N\left[\begin{pmatrix} \gamma_1 + \delta_1 Z_j \\ \gamma_2 + \delta_2 Z_j \end{pmatrix}, \boldsymbol{\Omega}\right], \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$

6.2.6 Leave-one-center-out cross-validation

If a model is tested on the same multicenter dataset that was used for model development, a cross-validation procedure can be used that mimics external validation.[92] This procedure is known as leave-one-center-out cross-validation (or internal-external cross-validation in the IPD meta-analysis literature)[142, 234, 240] and should be applied as follows:

1. Select a center j (j=1 to J) to be omitted.

2. Fit a risk model to the remaining J-1 centers (using one of the techniques considered in section 3.2 and adhering to an a priori defined model development strategy as explained in section 2.1.4).
3. Apply the model developed in step 2 in the omitted center j and compute the performance statistic of interest.
4. Repeat steps 1 to 3 for each center j=1,...,J.

Note that in case a mixed effects model was built in step 2, one can and should not employ conditional predictions estimates in step 3, as Empirical Bayes estimates are typically not available for new centers (see section 3.2.2). A summary statistic of performance can be obtained by computing the average of the J estimates obtained in step 3,[92] or by performing a random effects meta-analysis on them, as proposed in the previous sections.[234] The latter option gives an estimate of the between-center heterogeneity in performance. Alternatively, one could also compute J apparent performance statistics (see section 2.1.6) in step 3 by applying the model on data from the same centers that were used for model development. Next, the optimism in model performance can be computed by subtracting the performance of the model in center j from the apparent performance. If the model is externally valid, that is, generalizable to new centers, the optimism is zero. A weighted average of the estimated optimism can be presented along with its confidence interval, as proposed by Royston and colleagues.[240] Note that it is not useful to calculate the optimism for the calibration intercept and slope in this way, since logistic calibration is by definition perfect on the development dataset.

Debray and colleagues recommended inspecting the J performance estimates.[142] Provided the J models developed in step 2 all perform well in step 3, a risk model derived from the entire multicenter dataset is likely to be generalizable. In case some of the J models do not perform well, external validation is not guaranteed in all centers, and the researcher can attempt to explain why by inspecting differences in the characteristics of the centers, measurements or variable definitions. If the leave-one-center-out cross-validation shows limited external validity and the heterogeneity cannot be explained, the researcher might have to conclude that it is not possible to build one risk model (based on the considered predictors) that applies to all centers. Using center-specific predictions, as obtained by one of the techniques discussed in section 3.2.2 (among others, mixed effects models and model updating techniques) may improve predictions.

Note, finally, that the results of leave-one-center-out cross-validation are only informative if there is sufficient data.[142] In case only a few centers are available, or the centers all operate in a similar clinical setting and are not

representative of the broader target population, little insight in the external validity can be gained. In addition, if the center sizes are too small, the J performance statistics may not be reliably estimated.[95, 96]

6.3 Some examples

The validation studies of risk and classification models for the pre-operative diagnosis of ovarian masses in new data presented in this section serve as an illustration of some of the techniques discussed in the previous section. Four examples are discussed: the validation of IOTA strategies and RMI on IOTA phase III data in terms of discrimination and calibration, the validation of the newly developed Simple Rules risks scoring system in terms of discrimination and calibration, the validation of IOTA strategies in the hands of users with different levels of experience in IOTA IVb data, and the validation of the clinical utility of IOTA strategies, RMI and ROMA in IOTA phase III data. From these clinical studies, only the most important meta-analyses of predictive performance are presented. A complete overview of the results and details of these clinical studies can be found in the original publications.[163, 183, 184]

6.3.1 The validation of IOTA strategies on phase III data: a meta-analysis of discrimination and calibration

The purpose of this analysis is to compare different ultrasound-based IOTA strategies and RMI for ovarian cancer diagnosis using a meta-analysis of center-specific data from IOTA phase III, which includes 2403 patients with 1423 benign and 980 malignant adnexal masses. All patients underwent standardized transvaginal ultrasonography and the golden standard for diagnosis was histology after surgery. More details on the IOTA project, phase III data and risk and classification models for the pre-operative diagnosis of ovarian masses were given in section 3.3, and information on the multiple imputation of missing CA125 values is given in Appendix 1.

We evaluated five one-stage strategies, five two-stage strategies and two three-stage strategies for the pre-operative classification of ovarian tumors. The one-stage strategies are the use of either LR1, LR2, subjective assessment (SA), RMI or the IOTA Simple Rules (SRs, classifying all tumors where the SRs yield an inconclusive result as malignant). The two-step strategies are: the SRs as a first stage test and SA for tumors in which the SRs yield an inconclusive result; LR2 as a first stage test and SA for tumors in which LR2

yields an equivocal result (defined as a predicted risk of malignancy between 5% and 25%); IOTA Easy Descriptors (EDs) as a first stage test, SRs for tumors unclassifiable by the EDs and tumors unclassifiable by the SRs classified as malignant; EDs as a first stage test and LR2 for tumors unclassifiable by the EDs; EDs as a first stage test and SA for tumors unclassifiable by the EDs. The three-step strategies are: EDs as a first stage test, SRs for tumors unclassifiable by the EDs and SA for masses in which SRs are inconclusive; EDs as a first stage test, LR2 for tumors unclassifiable by the EDs and SA for masses in which LR2 yields a predicted risk between 5% and 25%.

Statistical analysis

All strategies were evaluated in terms of their ability to discriminate between benign and malignant masses. The c-statistic was computed for LR2, RMI and SA. Six levels of diagnostic confidence were used to compute the c-statistic for SA. We also calculated the sensitivity, the specificity, the positive and negative likelihood ratio (LR+ and LR–) and the diagnostic odds ratio (DOR). We used the cutoffs suggested in previous work for LR1, LR2 and RMI (i.e., a predicted probability of malignancy ≥10% for LR1 and LR2, and RMI >200 indicating malignancy). This section focusses on sensitivity and specificity. The results for LR+, LR- and the DOR can be found in the original publication.[183] Subgroup analyses for RMI and the most extensively validated IOTA methods (i.e., LR2 and SRs) were performed for pre- and postmenopausal women. The results of subgroup analyses can be found in the original publication as well.[183]

To acknowledge that performance may differ across centers, results were computed using the meta-analysis techniques described in the previous section. The calculations were performed using SAS software (version 9.3, SAS Institute, Cary, NC, USA). Proc mixed was used to estimate the meta-analysis model for the c-statistic (see equation (35)) with maximum likelihood. Proc glimmix was used for the bivariate model of logit sensitivity and logit specificity (see equation (32)), using maximum likelihood estimation with an adaptive Gauss-Hermite quadrature. Forest plots for LR2, SRs and RMI were created in R (www.r-project.org) using the rmeta package to present center-specific results and summary estimates.

For LR2, we also assessed calibration, that is, we assessed the extent to which the estimated risks of malignancy corresponded to the observed prevalence of malignancy, by constructing logistic calibration curves per center. RMI and SRs do not provide risk estimates but comparable center-specific curves were

Table 19. Predictive performance of subjective assessment by an expert, the IOTA strategies and the Risk of Malignancy Index (RMI), using a meta-analysis of center-specific data.

Approach	C (95% CI)	Sens, % (95% CI)	Spec, % (95% CI)
One step			
SA	0.914 (0.886 to 0.936)	92.5 (89.4 to 94.8)	87.7 (83.2 to 91.2)
LR1	0.930 (0.917 to 0.942)	93.7 (91.4 to 95.4)	77.6 (70.9 to 83.0)
LR2	0.918 (0.905 to 0.930)	90.2 (86.9 to 92.8)	78.9 (73.2 to 83.7)
SRsMal	N/A	95.3 (93.1 to 96.9)	74.1 (67.7 to 79.7)
RMI	0.875 (0.853 to 0.894)	67.1 (61.4 to 72.4)	90.6 (87.3 to 93.1)
Two steps			
EDs+SRsMal	N/A	95.7 (93.5 to 97.1)	73.6 (66.7 to 79.5)
EDs+LR2	N/A	91.1 (88.1 to 93.5)	78.1 (72.4 to 82.9)
EDs+SA	N/A	93.0 (90.0 to 95.1)	86.5 (81.8 to 90.1)
SRs+SA	N/A	91.8 (89.1 to 93.9)	89.0 (85.2 to 92.0)
LR2+SA	N/A	92.3 (89.5 to 94.5)	84.8 (80.4 to 88.3)
Three steps			
EDs+SRs+SA	N/A	92.5 (89.6 to 94.6)	87.6 (83.5 to 90.7)
EDs+LR2+SA	N/A	93.1 (90.5 to 95.0)	83.7 (79.2 to 87.4)

C: c-statistic; CI: confidence interval; sens: sensitivity; spec: specificity; SA: subjective assessment; SRsMal: assuming malignancy in case the Simple Rules are inconclusive; RMI: Risk of Malignancy Index; EDs: Easy Descriptors; SRs: Simple Rules.

obtained in the following manner: for RMI, logistic curves were constructed to link RMI values (based on log (RMI + 1)) to the observed risks; for the SRs, the proportion of malignant masses was calculated for each classification level (benign, inconclusive and malignant). In addition, one-step random effects logistic calibration was used for LR2. The calibration model (see equation (20)) was estimated with Laplace approximation in proc glimmix.

Results

The logistic regression models LR1 (c-statistic 0.930, 95% CI 0.917 to 0.942) and LR2 (c-statistic 0.918, 95% 0.905 to 0.930) had a c-statistic similar to SA by an expert (c-statistic 0.914; 95% 0.886 to 0.936). The c-statistic of RMI was lower (0.875, 95% CI 0.853 to 0.894). The IOTA risk models LR1 and LR2, and strategies using various combinations of SRs, EDs and LR2 achieved a sensitivity between 90% and 96% and a specificity between 74% and 79% (Table 19). When these strategies were combined with an expert's SA in case of an inconclusive or equivocal result, the specificity increased to 84% and higher, with a slightly reduced sensitivity in most instances (Table 19). The sensitivity of RMI was 67% and the specificity 91%, making it the least sensitive strategy to detect cancer, with a specificity comparable to that of multiple-step strategies including expert SA in the event of inconclusive results of the SRs, EDs or both.

There is considerable variation in the number of included masses and the observed prevalence of malignancy per center (Figure 21). The malignancy rate varied between 0 and 69%, while the number of enrolled cases per center ranged from 6 to 443. The observed c-statistic for LR2 showed little variability across centers, with a few exceptions for centers that enrolled a very small number of cases (Figure 21). In contrast, the observed c-statistic for RMI showed more variation between the centers (Figure 21). Intercenter differences in sensitivity and specificity were observed for LR2, RMI and SRs combined with SA (Figure 22). For LR2 and SRs, the observed differences between centers in sensitivity were smaller than the observed differences in specificity, whereas the inverse held true for RMI.

Note that the summary estimates of test performance of LR2, SRs and RMI were similar irrespective of whether they were estimated using meta-analysis or without taking clustering into account (i.e., pooling the data from all centers in one analysis) (Figure 21 and Figure 22). However, pooling all data underestimates the uncertainty originating from between-center differences, while this is appropriately addressed by adopting meta-analysis techniques, as reflected in the broader confidence intervals. Additionally, prediction intervals indicate the likely performance in new centers. The 95% prediction interval for the c-statistic is 0.904 to 0.930 for LR2 and 82.6% to 91.2% for RMI. The 95% prediction interval for the sensitivity is 85.3% to 93.6% for LR2, 37.6% to 87.4% for RMI and 85.4% to 94.6% for SRs followed by SA. The 95% prediction interval for the specificity is 70.5% to 85.4% for LR2, 80.1% to 95.9% for RMI and 83.3% to 93.0% for SRs followed by SA.

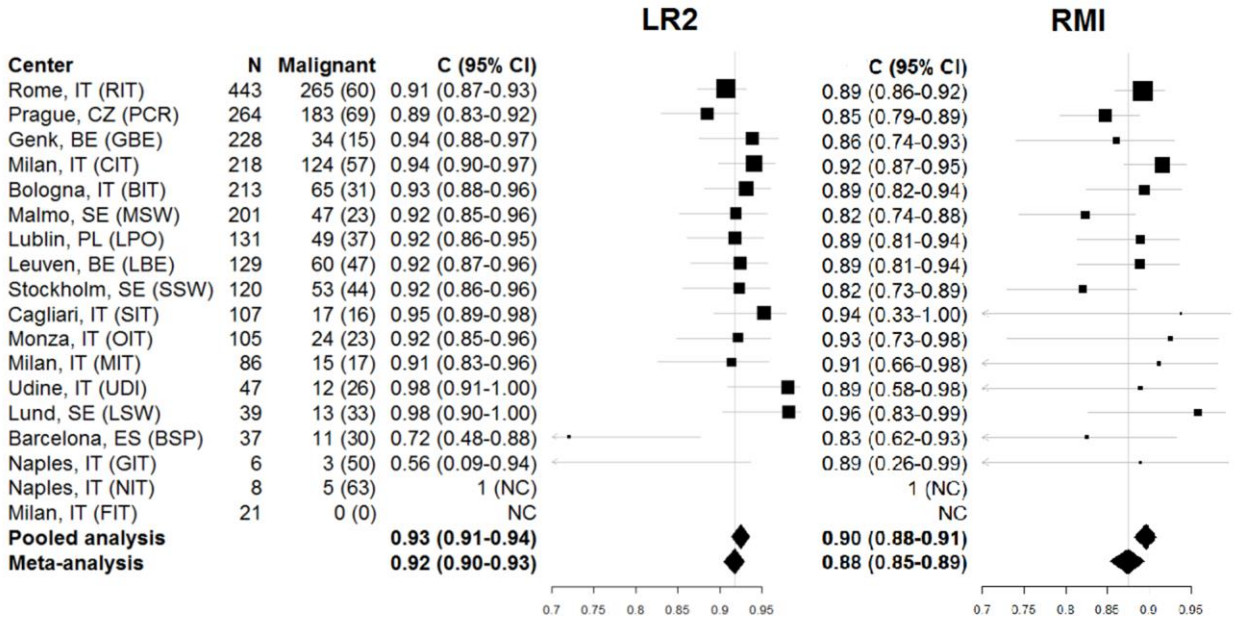
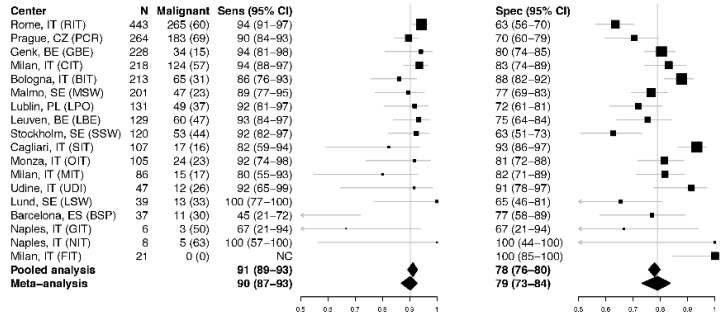
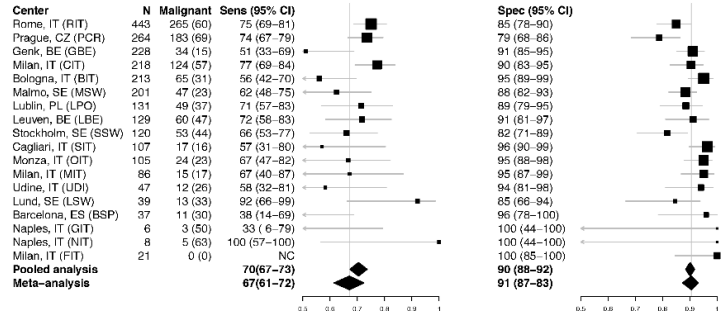


Figure 21. The c-statistic for LR2 and the Risk of Malignancy Index (RMI) per contributing center and for all centers combined using a meta-analysis approach and pooled data. NC: not computable. Numbers in brackets denote the prevalence (%) of malignant masses in each center.

A



B



C

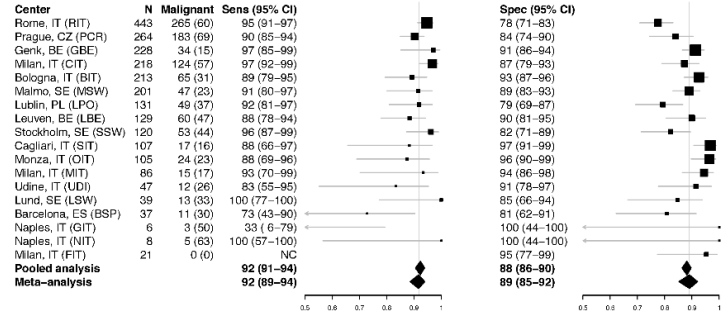


Figure 22. The sensitivity and specificity for LR2 (A), the Risk of Malignancy Index (B) and a two-stage strategy using the Simple Rules as a first stage test followed by subjective assessment by an expert if the Simple Rules are inconclusive (C). Results are presented per contributing center and for all centers combined using a meta-analysis approach and pooled data. NC: not computable. Numbers in brackets denote the prevalence (%) of malignant masses in each center.

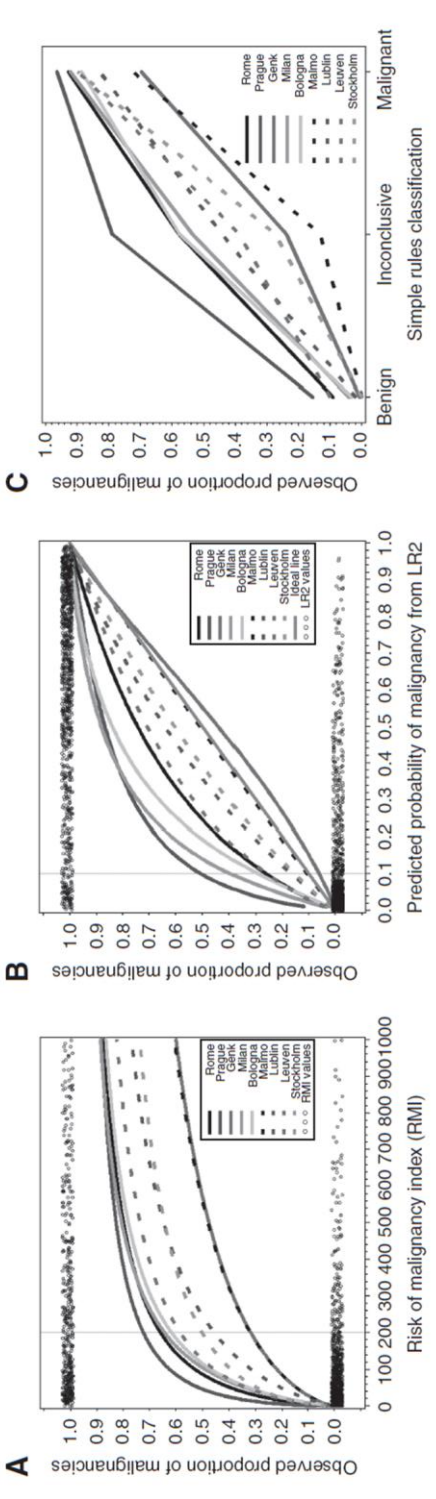


Figure 23. Center-specific logistic calibration curves for the Risk of Malignancy Index (A), LR2 (B) and the Simple Rules (C).

Calibration results differed between the nine largest centers for LR2, RMI and SRs (see Figure 23). This means that, given a specific prediction from a diagnostic test (LR2 risk, RMI value or SRs category), the observed prevalence of malignancy varies between centers. The prevalence of malignancy in women with an RMI score of 200 varied between centers from 30 to 70%, and RMI values of 200 or more were associated with high malignancy rates. For LR2, the risk of malignancy was underestimated in seven of the nine centers (as indicated by calibration curves above the diagonal), slightly overestimated in one center and perfectly calibrated in one center. The average calibration intercept was 0.83 (95% CI 0.45 to 1.20), indicating a clear underestimation of the predicted risks, while the average calibration slope was 1.02 (95% CI 0.93 to 1.11). Most of the observed variance in calibration intercepts was attributable to heterogeneity between centers (ICC=83%), whereas there was no heterogeneity in calibration slopes.

Discussion

This comparison of IOTA risk models and diagnostic strategies in different clinical environments using meta-analysis showed excellent predictive performance for all IOTA methods to characterize adnexal masses before surgery. All IOTA strategies manifested better discrimination than RMI. In addition, the results have demonstrated intercenter differences in predictive performance for LR2, SRs and RMI. The use of meta-analysis did not meaningfully change the summary estimates of performance from those obtained with a standard analysis of pooled data, but gave wider confidence intervals, properly reflecting the uncertainty caused by differences between centers. Furthermore, the random effects meta-analysis minimizes the risk that results are overly influenced by a single center recruiting many more patients than others.

The many participating centers and the different types of participating centers make it likely that the results are generalizable. A limitation of this study is that the tested strategies were not directly applied when scanning the patients. Instead, 440 clinical and ultrasound variables were prospectively collected from each patient, and predictions and classifications were obtained centrally at the time of data analysis. A second limitation is that information on CA125 was missing in 40% of cases, which was dealt with by multiple imputation.[28] Sensitivity analyses using only complete cases for CA125 confirmed the difference in test performance in favor of the IOTA methods (LR2 and SRs)[183]. A third limitation is that most of the patients in the IOTA phase III study were scanned by the same experienced examiners as in the centers where

the IOTA methods were developed, or by examiners that had already adopted the IOTA examination technique and terminology.

This multicenter study demonstrated differences in test performance between centers for LR2, SRs and RMI. However, in all centers, also in those with a low observed prevalence of malignancy, the sensitivity with regard to malignancy was much higher for the IOTA methods than for RMI. The overall discriminative capacity (c-statistic) for LR2 and RMI did not seem to be affected by cancer prevalence. In contrast, the results suggest important differences in calibration results for LR2, SRs and RMI. The type of center appeared to explain these differences: oncology centers have a higher prevalence of malignant tumors and suffered from underestimation of the predicted risk. The heterogeneity in calibration was noticeable for SRs and RMI as well, although these methods do not directly provide an estimated risk. For example, it can be derived that the implicit average risk at an RMI value of 200 is 54%. This implies that at this (commonly used) cutoff, patients with a risk of malignancy of ≥50% are classified as low risk patients in some centers.

We did not undertake a full meta-regression to explain in detail the intercenter differences in results for test performance and calibration as this was beyond the scope of the research. The most plausible explanations are differences in study populations (e.g., patients' age, body mass index or tumor mix), equipment and examiners' use of the IOTA terms. However, the variation between centers in calibration curves highlights that caution is needed when interpreting and using diagnostic test results (that is, predictions) for individual patient management within the context of a single center. Future studies should explore the reasons for differences in predictive performance between different centers, the ultimate goal being to improve risk prediction for ovarian cancer.

6.3.2 The validation of the IOTA Simple Rules risks scoring system on phase III data: a meta-analysis of discrimination and a graphical assessment of calibration in specialized oncology centers and other centers

The aim of this analysis is to evaluate the Simple Rules risk scoring system on IOTA phase III data. The development of the Simple Rules risk scoring system was discussed in section 3.3.4. Details on IOTA phase III data have been discussed in the previous section and in section 3.3.2.

Statistical analysis

The risk estimates were externally validated on IOTA phase III data. The c-statistic, sensitivity and specificity were calculated through a meta-analysis of center-specific results, as described in section 6.2. We used SAS software (SAS Institute, Cary, NC, USA) to estimate the meta-analysis models (proc mixed with maximum likelihood for the c-statistic (see equation (35)) and proc glimmix with an adaptive Gauss-Hermite quadrature for the bivariate model of logit sensitivity and logit specificity (see equation (32))). R (www.r-project.org) was used to create forest plots (with the rmeta package). The predictive values and positive and negative likelihood ratio were computed as well but will not be considered in this section. These results can be found in the original publication.[163] The risk thresholds considered to classify a mass as malignant were 1%, 3%, 5%, 10%, 15%, 20%, 25% and 30%. To assess the relationship between calculated risks and observed proportions we plotted flexible calibration curves (loess curves) using the val.prob.ci.2 function in R.[64] We explored differences in discrimination and calibration between oncology centers and other centers. After external validation, the risk calculation underlying the Simple Rules risk scoring system was updated using all available data (phases Ib, II and III) to fully exploit all available information. The re-estimated equation is presented above (section 3.3.4).

Results

The c-statistic was 0.917 (95% CI, 0.902 to 0.930) and was very similar in oncology centers (0.917; 95% CI, 0.901 to 0.931) and other centers (0.916; 95% CI 0.873 to 0.945). In all but 3 centers, the observed c-statistic was at least 0.90. Two centers had an observed c-statistic of 0.89 and one small center had an observed c-statistic <0.80 (Figure 24). The 95% prediction interval was 0.901 to 0.931. The estimated risks were well calibrated in the entire validation dataset, and when assessed for patients from oncology centers and other hospitals separately (Figure 25). In all, 22.8% of the patients in the validation set had a calculated risk of malignancy <1%, while 48.5% had a calculated risk ≥30%. The sensitivity varied between 99.7% at the 1% risk threshold and 89.0% at the 30% risk threshold. The specificity varied between 33.7% at the 1% risk threshold and 84.7% at the 30% risk threshold. The sensitivity and specificity for the same risk threshold differed between oncology centers and other centers (Table 20).

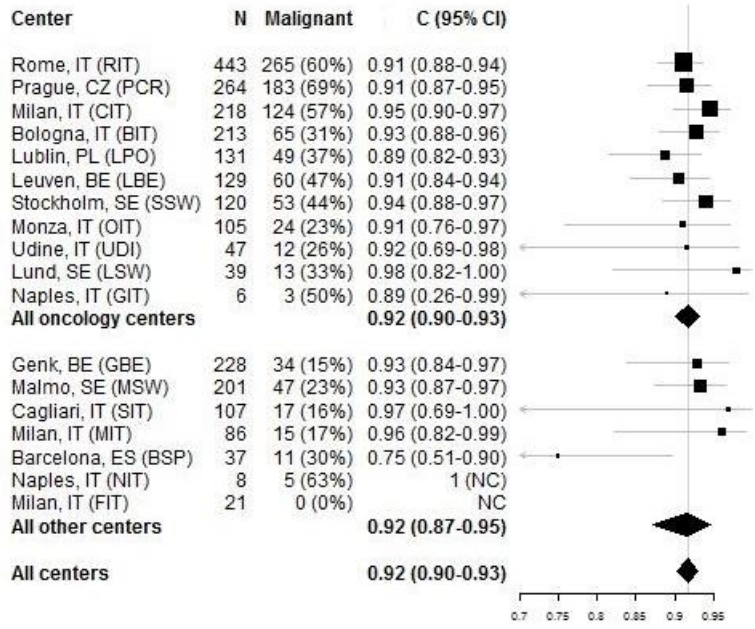


Figure 24. The c-statistic for the Simple Rules risk scoring system per contributing center, for all oncology centers, all other centers and all centers combined using a meta-analysis. NC: not computable. Numbers in brackets denote the prevalence (%) of malignant masses in each center.

Discussion

At prospective validation the Simple Rules risk scoring system showed its good ability to discriminate between benign and malignant tumors and its good agreement between the predicted risk of malignancy and the true prevalence of malignancy.

The IOTA Simple Rules are intuitively attractive because of their ease of use. However, when used as originally suggested they allow only a categorization of tumors into three groups: benign, malignant or inconclusive.[191] The results showed that the Simple Rules can also be used to estimate the risk of malignancy in every adnexal mass and can be used for individualized patient management. The type of center needed to be included in the risk estimation, because the risk of a malignant tumor is higher in oncology centers than in others. The predictive performance of the Simple Rules risk scoring system was good in both oncology and other centers. In future studies, the Simple

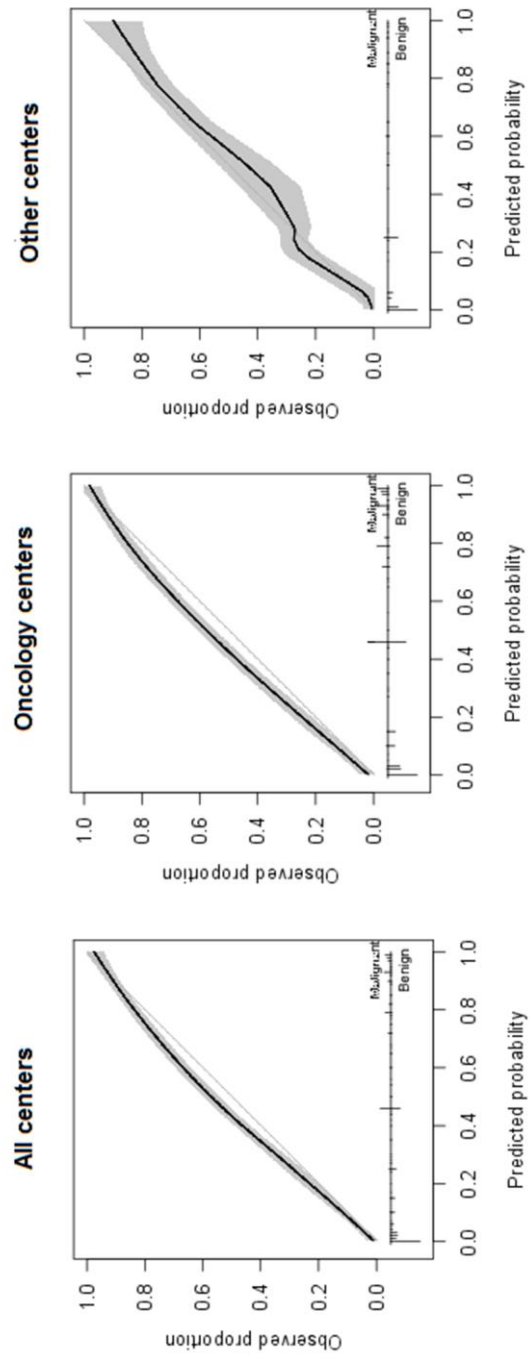


Figure 25. Calibration curves for all data, data from oncology centers and data from other centers. Gray line: ideal calibration; black line: flexible calibration curve, gray area: 95% confidence band. The distribution of predicted risks of malignancy is depicted in the histogram at the bottom, with positive bins showing the number of patients with malignant tumors and negative bins showing the number of patients with benign tumors.

Table 20. Sensitivity and specificity for the Simple Rules risk scoring system in oncology centers and other centers.

Risk threshold for malignancy	Center	Size of high risk group n (%)	Sensitivity (95% CI)	Specificity (95% CI)
1%	Oncology	1439 (83.9)	99.7 (99.0-99.9)	27.3 (20.3-35.5)
	Other	417 (60.6)	98.3 (84.5-99.8)	48.0 (37.4-58.8)
3%	Oncology	1312 (76.5)	98.4 (97.3-99.1)	41.3 (34.8-48.1)
	Other	325 (47.2)	98.5 (85.0-99.9)	66.4 (52.6-77.9)
5%	Oncology	1201 (70.0)	97.8 (96.3-98.7)	57.0 (46.9-66.5)
	Other	299 (43.5)	98.4 (84.9-99.9)	72.5 (57.5-83.7)
10%	Oncology	1199 (69.9)	97.8 (96.4-98.7)	57.2 (47.3-66.4)
	Other	255 (37.1)	96.7 (90.1-98.9)	80.1 (67.7-88.6)
15%	Oncology	1121 (65.4)	96.1 (93.3-97.7)	65.6 (56.6-73.7)
	Other	255 (37.1)	96.7 (90.1-98.9)	80.1 (67.7-88.6)
20%	Oncology	1045 (60.9)	94.9 (92.0-96.8)	73.4 (66.9-79.1)
	Other	254 (36.9)	96.7 (90.1-98.9)	80.2 (67.9-88.6)
25%	Oncology	1045 (60.9)	94.9 (92.0-96.8)	73.4 (66.9-79.1)
	Other	249 (36.2)	95.8 (90.1-98.3)	80.2 (67.9-88.6)
30%	Oncology	1042 (60.8)	94.9 (91.8-96.9)	73.7 (67.2-79.3)
	Other	123 (17.9)	63.3 (44.5-78.8)	94.8 (91.0-97.1)

Rules risk estimates need to be prospectively and externally validated, and their use in a classification system for clinical management has to be further investigated. We conclude that individual risk estimates can be derived from the 10 ultrasound features in the Simple Rules with a performance similar to that of the best previously published algorithms. A simple classification based on these risk estimates may form the basis of a clinical management system. This will hopefully facilitate choosing optimal treatment for all patients presenting with adnexal masses.

6.3.3 The validation of IOTA models and RMI in the hands of users with varied training on phase IVb data: a meta-regression of test accuracy

The correct characterization of ovarian tumors is critical to optimize patient care. The purpose of this study is to evaluate the diagnostic performance of LR2, the Simple Rules (SRs), the Risk of Malignancy Index (RMI) and subjective assessment (SA) for the preoperative characterization of adnexal masses, when ultrasonography is performed by examiners with different background training and experience. Data from IOTA phase IVb was used for this purpose. 35 level II ultrasound examiners from three UK hospitals contributed data. They performed transvaginal ultrasonography using a standardized approach. The final outcome was determined by the surgical findings and histological diagnosis. The final dataset contained information of 255 women with 74 malignant masses. More details on phase IVb data can be found in section 3.3.2.

Statistical analysis

The c-indexes were computed to summarize discrimination by LR2 and RMI. Since 70% of the patients were collected at one hospital, we report c-statistics computed on the whole sample instead of performing a random effects meta-analysis of hospital-specific c-statistics, the results of which were nearly identical. Diagnostic performance measures (sensitivity, specificity, LR+, LR- and DOR) were computed for the classification as benign or malignant based on RMI, LR2, the SRs and SA. We assumed a risk threshold of 10% for LR2 and a cut-off of 200 for RMI. In what follows, sensitivity and specificity are reported. Other results can be found in the original publication, as well as subgroup analyses for pre- and postmenopausal patients.[184]

We also conducted an exploratory analysis of the influence of experience on the performance of subjective assessment and the risk and classification

models. We fitted a regression model for accuracy using the number of ovarian mass scans (7 ordinal categories; <100, 100–200, 200–500, 500–1000, 1000–2000, 2000–5000 and 5000–10 000), background training (sonographer or MD) and tumor outcome (benign or malignant) as predictors. The outcome was added to adjust the effects of the predictors. We used a mixed effects model to account for the clustering of patients within operators. All analyses were performed using SAS software (version 9.3, SAS Institute, Cary, NC, USA).

Results

Of the 255 women in the study sample, 74 (29%) had a malignant tumor. The c-statistics for LR2 and RMI were 0.94 (95% CI 0.89 to 0.97) and 0.90 (95% CI 0.83 to 0.94), respectively. The SRs were able to classify 214 (83.9%) of the masses as benign or malignant. Of the 41 unclassified tumors, 20 were benign and 21 malignant. A strategy classifying all these tumors as malignant (SRsMal) yielded a higher sensitivity (91%) than using the RMI (72%). However, the specificity of this strategy was lower (87% versus 94%). When examiners used their own SA as a second-stage test when the SRs were inconclusive, sensitivity was higher than for RMI (86% versus 72%), with no difference in specificity (Table 21).

In all, 62.9% of the operators had performed <1000 ultrasound scans; 24% of the operators were MDs, whereas 76% were sonographers. The exploratory analysis of the influence of operator experience and training on diagnostic performance suggested that MDs were better able to subjectively assess the correct diagnosis than sonographers (odds ratio 2.59, 95% CI 0.77 to 8.74). When using SRsMal and LR2 to classify masses as benign or malignant, the odds ratios were 1.10 (95% CI 0.32 to 3.81) and 0.68 (95% CI 0.18 to 2.61), respectively, suggesting similar performance of these models in the hands of

Table 21. Sensitivity and specificity of LR2, RMI, the Simple Rules and subjective assessment by users with varied levels of experience and training.

	Sensitivity	**Specificity**
LR2	0.88 (0.78 to 0.93)	0.90 (0.84 to 0.93)
RMI	0.72 (0.60 to 0.81)	0.94 (0.90 to 0.97)
SRs+SA	0.91 (0.82 to 0.95)	0.87 (0.82 to 0.91)
SRsMal	0.86 (0.77 to 0.92)	0.94 (0.90 to 0.97)
SA	0.88 (0.78 to 0.93)	0.91 (0.85 to 0.94)

RMI: Risk of Malignancy Index; SRs: Simple Rules; SA: subjective assessment; SRsMal: assuming malignancy in case the Simple Rules are inconclusive.

MDs and sonographers. When using the RMI, the odds ratio was 0.32 (95% CI 0.06 to 1.70), suggesting slightly better performance for sonographers. The number of previous ovarian mass scans had little effect, with odds ratios between 0.85 and 1.01 for each category increase on the ordinal measurement scale. Adding hospital as a fixed effect in the mixed effects model had no influence on the final results.

Discussion

The results suggest that the test performance of IOTA models and RMI was maintained in examiners with varying levels of training and experience. Criticism of external validation studies of IOTA models has focused on the fact that the models were developed and tested by examiners with a specific expertise in the imaging of adnexal pathology. In this study, on the contrary, the ultrasound scans were performed by examiners with different training (sonographers and doctors) and level II experience. As most ovarian pathology is probably examined by sonographers or doctors who do not have a special interest in gynecologic ultrasonography (level II), it seems reasonable to suggest these findings offer clinicians a clearer idea on the performance of the different adnexal mass risk models in daily practice.

The performance of LR2 was in accordance with the results of Nunes and colleagues, who externally validated the LR2 model in the hands of a single relatively inexperienced gynecologist (level II).[186] Despite the limited sample size, the results suggest that subjective impression of the nature of an adnexal mass tended to be better by medically trained examiners than sonographers. However, this difference was not seen when doctors and sonographers were asked to enter ultrasound findings into the risk model LR2 or when they used SRs. This is likely to reflect variations in training, as sonographers are in general taught to identify and report the structures they see. Hence, they are skilled in accurately entering the presence or absence of the structures required for use in risk or classification models but are less likely to offer an opinion on the final diagnosis. This is an important observation, as the original aim of the IOTA study was to develop tools that could be used by all examiners to enhance their diagnostic performance.

Ideally, when the SRs are inconclusive, the patient should be referred to an expert in gynecological scanning for further assessment (level III) as an optimal second-stage test. In the absence of level III ultrasonography, the results suggest that if the SRs are inconclusive, an acceptable second-stage test for level II doctor ultrasound examiners is the subjective impression of the

scan findings. For sonographers, however, a reasonable strategy would be to classify all such lesions as malignant.

In conclusion, correctly classifying the nature of ovarian pathology is a common diagnostic problem in gynecology, and correctly identifying the presence of cancer in these cases is the key to ensure patients access appropriate treatment. This study shows that the LR2 model and the SRs perform well in the hands of both relatively inexperienced doctors and when used by sonographers.

6.3.4 The validation of the clinical utility of models on phase III data: net benefit in specialized oncology centers and other centers

Whereas previous sections focused on discrimination and calibration, model performance can also be evaluated in terms of appropriate decision-making: this transcends the statistical measures and approaches a more clinically relevant evaluation. The purpose of this analysis is to externally validate preoperative diagnostic models for ovarian cancer by means of decision curve analysis (see section 2.1.6).[72, 73, 239] This method is based on the direct link between risk threshold and the relative consequences of false positives and false negatives.[74] Using this link, the classification of benign and malignant tumors can be combined into a single figure (NB) that quantifies the utility of the model for clinical decision-making. IOTA phase III data (see section 3.3.2) is used to assess the utility of RMI, ROMA (cut-off 200), LR2, the SRs risk scoring system and ADNEX to decide which patients should be referred to a specialized gynecological oncologist (see section 3.3.1 for more clinical background information).

Statistical analysis

Clinical utility was assessed with NB at a sensible range of risk thresholds (5% to 50%) to select patients for referral to a gynecological oncologist. When a risk threshold of 5% (odds 5:95=1:19) was adopted to select patients for treatment, this would be equivalent to accepting up to 19 false positives per true positive, that is, referring at most 19 patients with a benign mass to a gynecological oncologist for each patient with a malignant mass. When a risk threshold of 50% (odds 50:50=1:1) was adopted to select patients for treatment, this would be equivalent to accepting up to 1 false positive per true positive, that is, referring one patient with a benign mass to a gynecological oncologist for each patient with a malignant mass.

The decision curves show the NB for different risk thresholds. The NB of the models was compared with the NB of two default strategies: referring all patients to a gynecological oncologist or referring no patients at all. If a model had a lower NB at a given threshold than either of the default strategies, the model was deemed harmful, and using the best default strategy to make referral decisions was clinically more useful than using the model.

In addition, we compared the clinical utility of the best performing model with that of all other models with a NB higher than that of the best default strategy (i.e., models that were not harmful). The difference in clinical utility between models was expressed in terms of gain in net sensitivity, keeping a constant specificity. This was computed as (NB best model−NB model)/prevalence.[81] Additionally, we used an equivalent expression of the difference in terms of the net amount of avoided false positives per hundred patients, keeping a constant number of true positives, computed as $100\times(1-t)/(t)\times$(NB best model−NB model), where t denotes the risk threshold. 95% confidence intervals were obtained using the percentile method with 1000 bootstrap samples. We investigated the clinical utility of models in subgroups of oncology centers and non-oncology centers.

The measurement of serum CA125 was optional and we used single imputation to deal with missing values. CA125 was estimated with predictive mean matching regression, as described in Appendix 1. Typically, multiple imputation is preferred over single imputation to get variance estimates that reflect uncertainty due to missingness. However, we noticed in previous studies that variance estimates were not meaningfully smaller if single imputation was used for the models we assess in this study. Hence, we use single imputation to reduce the computational burden.

Results

NB was high for the SRs risk scoring system and for ADNEX at all risk thresholds (Figure 26). The NB of RMI using a cut-off of 200 was considerably lower than the NB of all other models, across all risk thresholds. LR2 showed intermediate clinical utility. At risk threshold 10%, the NB of the SRs risk scoring system, ADNEX, LR2 and RMI were 0.376 (95% CI 0.355 to 0.398), 0.374 (95% CI 0.354 to 0.395), 0.357 (0.338 to 0.378) and 0.278 (0.258 to 0.296), respectively. Using RMI was harmful at risk thresholds below 20%, meaning that referring all patients is more clinically useful than using RMI if one is willing to risk referring more than four benign cases to a gynecological oncologist in order to detect one malignant case.

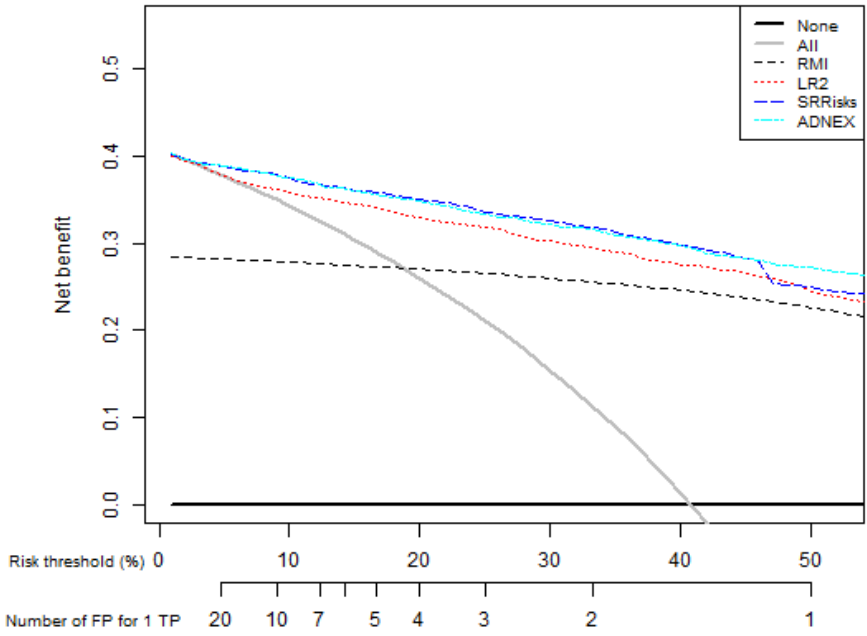


Figure 26. Decision curves of the Risk of Malignancy Index (RMI), the LR2 model, the IOTA Simple Rules Risk scoring system (SRRisks) and the ADNEX model.

In oncology and non-oncology centers, the SRs risk scoring system and ADNEX maintained their high NB, while RMI had no to little clinical utility (see Figure 27). In oncology centers, RMI was harmful at risk thresholds of 20% or less. The clinical utility of the SRs risks tended to be higher than the clinical utility of LR2 and RMI up to risk thresholds of 40%, but comparable to the clinical utility of ADNEX (see Appendix 6). For example, at a risk threshold of 10%, using the SRs risks instead of LR2 yielded an increase in net sensitivity of 4% (95% CI 2% to 6%), or equivalently, a net amount of 17 avoided false positive results per 100 patients (95% CI 10 to 25). Using the SRs risks rather than ADNEX at the risk threshold of 10% yielded an increase in net sensitivity of 0% (95% CI -1% to 1%), or equivalently, a net amount of 0 avoided false positive results per 100 patients (95% CI -4 to 5). In other centers, RMI was harmful at risk thresholds of 10% or less. The clinical utility of LR2 was comparable to that of the SRs risks at risk thresholds of 15% and above (see Appendix 6).

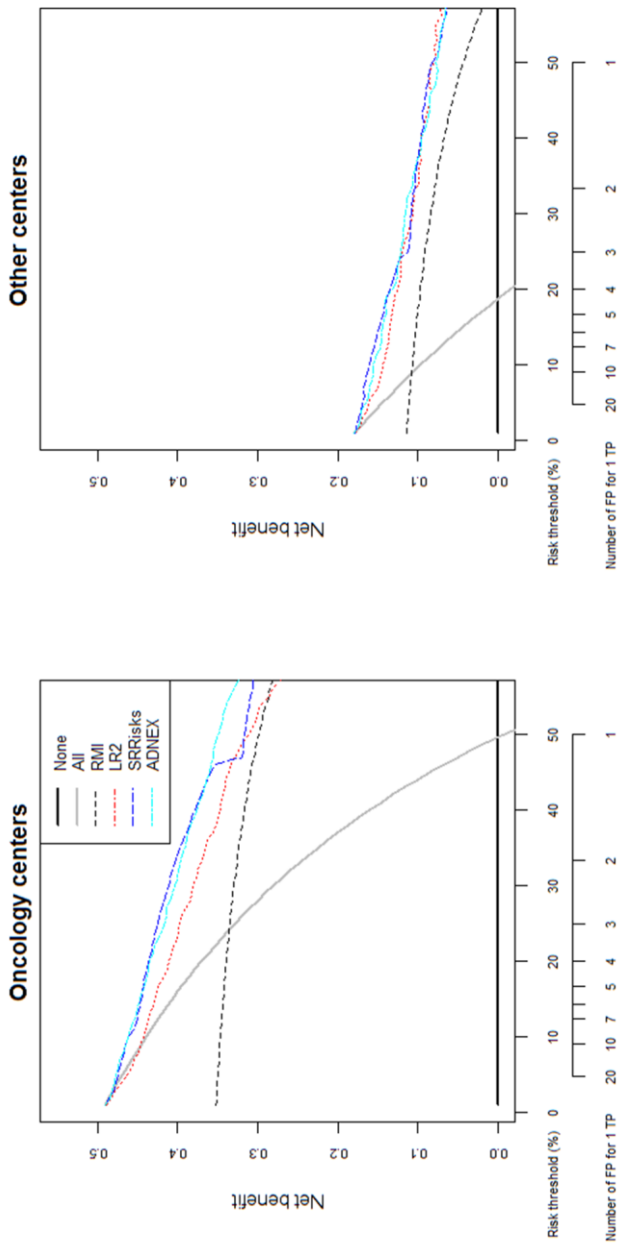


Figure 27. Decision curves of the Risk of Malignancy Index (RMI), the LR2 model, the IOTA Simple Rules Risk scoring system (SRRisks), and the ADNEX model in oncology centers and other centers.

Discussion

This study has shown that IOTA models (ADNEX, the SRs risk scoring system and LR2) have clinical utility at a broad range of risk thresholds to refer patients with ovarian masses to a gynecological oncologist. The ADNEX model and the SRs risk scoring system consistently had a very high NB. In contrast, RMI (with a cut-off of 200) had less clinical utility and was harmful at risk thresholds below 20%. A different prevalence and differential calibration of the models in tertiary referral centers with a specialized gynecological oncology unit and other centers resulted in different decision curves. Nonetheless, IOTA models also had more clinical utility than RMI in regional hospitals and tertiary referral centers without a gynecological oncology unit. These findings suggest that the preoperative characterization of adnexal masses and the subsequent decision to refer patients to a gynecological oncologist should incorporate the use of the IOTA Simple Rules or the ADNEX model, regardless of the clinical setting.

We chose to investigate the clinical utility of pre-operative diagnostic models to refer patients to a gynecological oncologist. We defined a relevant range of risk thresholds (5 to 50%) for the decision to refer to a gynecological oncologist. However, in the management and triage of patients with suspected ovarian masses, many other decisions must be considered, such as the decision to assign a patient to conservative follow-up instead of scheduling her for operation. For other decisions, different risk thresholds than those investigated here may be more appropriate. For example, for the decision to conservative follow-up, risk thresholds > 10% may not be appropriate.

A question a decision curve analysis is unable to answer, is which threshold to choose.[239] Although a risk threshold may increase the uptake of models and simplify patient management, choosing a threshold should not be taken lightly. In fact, no single risk threshold can be recommended, because the risk threshold depends on the clinical setting in which the model is applied. The threshold should be in alignment with a country's health care system and local referral patterns and guidelines. Risk thresholds should also reflect patients' preferences and characteristics. For example, different risk thresholds may be appropriate for women of reproductive age and postmenopausal women. Furthermore, classification rules based on risk thresholds could never substitute a physician's clinical judgement, which is essential to interpret an algorithm's outcome in the context of broader understanding of the patient's condition.

6.4 A meta-analysis of net benefit

NB is dependent on the prevalence in the population,[233] discrimination and calibration.[232] Using the risk model may even be harmful, that is, worse than treating all patients or treating none of the patients, if the model is badly calibrated. Differences in prevalence and model performance may be common in multicenter studies, as the clinical settings in which a model is to be applied are often heterogeneous. So far, no methods have been proposed for a meta-analysis of NB. This is crucial, however, if one wants to evaluate whether it will be useful to implement a model in a new center. This section aims to explore methods that capture the heterogeneity in the clinical utility of a model across centers. First, multiple ways to obtain a summary decision curve are investigated. Second, a random effects meta-analysis of center-specific NBs is applied to IOTA data. Third, some directions for future research into methods for meta-analysis of NB are proposed.

6.4.1 Various fixed and random effects weights

This section considers various fixed and random effects meta-analyses of NB to construct a summary decision curve in a simulated heterogeneous dataset.

Methods

We used the various fixed and random effects weights that have previously been proposed by van Klaveren for the meta-analysis of c-statistics in multicenter studies:[218]

1. The number of observations at each center (n_j). This is equivalent to pooling the data of all centers and computing the NB directly in an unweighted fashion.
2. Equal weights for each center, 1/J, where J is the number of centers.
3. The number of events in each center.
4. The inverse of the centers' estimated sampling variances of the NB, $1/\widehat{\sigma}_j^2$. The sampling variance was computed by bootstrap resampling.[73, 241]
5. Random effects weights, computed as $1/(\widehat{\sigma}_j^2 + \hat{\tau}^2)$, where $\hat{\tau}^2$ is the estimated between-center variance in NB.
6. Random effects weights, computed as above but performing the meta-analysis on the log(1-NB) scale.

If at a certain threshold there are no positive classifications, NB is zero, and the bootstrap estimate of the sampling variance is zero as well. In this case, the inverse variance weights are undefined. The proc mixed procedure to perform the random effects meta-analysis in SAS software (SAS Institute, Cary, NC, USA) also requires within-cluster sampling variance estimates larger than zero. We imputed an arbitrarily small sampling variance, defined as half of the minimum observed sampling variance of the NB in centers with a sampling variance larger than zero.

For option 5, we computed an approximate 95% prediction intervals using equation (33). For option 6, the prediction interval is computed on the log(1-NB) scale and then transformed back to the NB scale.

A simulated dataset was used to illustrate these techniques. By manipulating the case-mix, the outcome prevalence and the true association between the outcome and predictors, we were certain that the NB was truly heterogeneous across centers. A dataset with 50 centers and on average 100 patients per center was created, in which each center had its own true population model and case-mix (see Appendix 7). One standard logistic regression model (ignoring clustering) was applied to all centers in the dataset. This model was developed from a separate multicenter dataset generated from the same underlying center-specific populations.

The NB of the risk model was computed at thresholds from 0 to 0.9 by steps of 0.1, using each of the proposed weights for meta-analysis. The summary NBs for each of the risk thresholds were joined in a decision curve.

Results

The center-specific sample sizes in the simulated validation dataset ranged from 66 to 125 (median 100, IQR 93 to 105), while the prevalence ranged from 4% to 71% (median 25%, IQR 17% to 33%) (see Appendix 7). The observed c-statistics ranged from 0.58 to 0.95 (median 0.86, IQR 0.77 to 0.91). The observed calibration intercepts ranged from -1.8 to 1.5 (median 0.04, IQR -0.8 to 0.6), and the observed calibration slopes ranged from 0.3 to 3.2 (median 1.2, IQR 0.8 to 1.5). As a consequence of the large differences in the observed discrimination, calibration and prevalence, there were also large differences in the observed decision curves across centers (see Appendix 8).

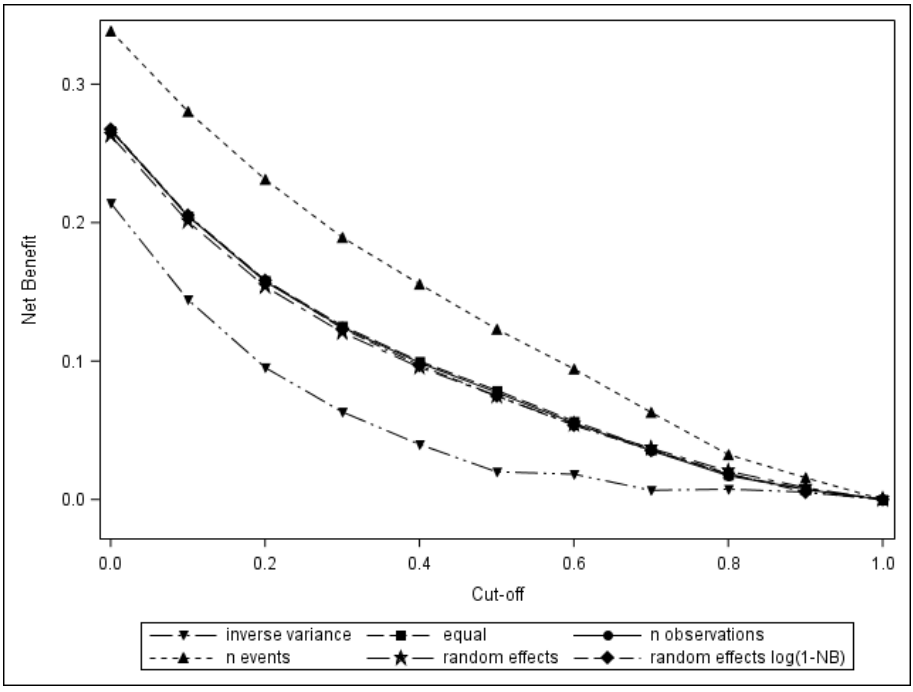


Figure 28. Decision curves using various weights for a meta-analysis to summarize center-specific net benefit.

Weighting by the number of events produced high summary NBs. NB tends to be high in populations with a high prevalence, and these centers tended to get high weights (see Figure 28 and Appendix 9). In contrast, the inverse variance weighted summary curve produced low summary NBs. Centers with a prevalence around 0.5 tend to have the highest sampling variability and hence the lowest weights. These centers also tended to have very high NBs. The random effect weighted summary curve based on the untransformed and the transformed NB, the equal weights summary curve and the curve weighted by the number of observations per center were very similar.

The summary estimates based on the random effects meta-analysis of NB and log(1-NB) are similar, although the point estimates based on the log(1-NB) transformation are slightly larger at low thresholds. The 95% prediction intervals based on the transformed NB are asymmetric and are shifted slightly downward in comparison to the prediction intervals based on the untransformed NB (see Table 22).

Table 22. A comparison of the summary estimates and prediction intervals of a random effects meta-analysis based on net benefit and a log(1-net benefit) transformation.

	Transformed			Untransformed		
t	**NB**	**lpl**	**upl**	**NB**	**lpl**	**upl**
0	0.27	-0.05	0.49	0.26	-0.01	0.53
0.1	0.21	-0.13	0.44	0.20	-0.09	0.49
0.2	0.16	-0.18	0.40	0.15	-0.13	0.44
0.3	0.12	-0.17	0.34	0.12	-0.14	0.38
0.4	0.10	-0.15	0.29	0.10	-0.13	0.32
0.5	0.07	-0.13	0.24	0.07	-0.11	0.26
0.6	0.05	-0.11	0.19	0.05	-0.10	0.21
0.7	0.04	-0.08	0.14	0.04	-0.07	0.12
0.8	0.02	-0.02	0.05	0.02	-0.02	0.06
0.9	0.01	0.00	0.01	0.01	0.01	0.01
1	0.00	0.00	0.00	0.00	0.00	0.00

t: risk threshold; NB: net benefit; lpl: lower prediction limit; upl: upper prediction limit.

Discussion

This study illustrates that the summary curves for NB depend on the weights chosen for meta-analysis. It seems impossible to define what the "true" summary curve should be. Nonetheless, two of the investigated options may be considered inappropriate to obtain summary curves. Weighting by the number of events systematically gives high weights to centers with a high NB, as the prevalence strongly influences both the number of events (and hence the weights) and the NB itself. Using inverse variance weights, by contrast, systematically gives low weights to centers with a high NB, as a high prevalence (around 0.5 can be considered high in many cases) is associated with a high sampling variance (and thus low weights) and a high NB. Since the random effects weights include the sampling variance in the denominator, it may be expected that they suffer from this issue as well.

An advantage of using a random effects approach to the meta-analysis of center-specific NB is that a prediction interval can be calculated to reflect the likely NB that may be expected in new centers. However, more research is needed to derive the coverage probabilities of the prediction intervals based on random effects meta-analyses of NB and log(1-NB). The normality

assumption on the distribution of center-specific random effects may be violated in practice. Especially at high risk thresholds, one may expect an asymmetric distribution that is strongly peaked near zero, but with many centers with a NB below 0. Specifically, in centers in which the model discriminates well but is badly calibrated such that it overestimates the risk of an event, a NB below zero occurs often at high risk thresholds.[232] Therefore, prediction intervals that are too concentrated at (or slightly above) zero at high risk thresholds may be misleading.

A drawback of the inverse variance weighted and random effects weighted NB is that estimates of the sampling variance are needed, for which no closed formula exists. Hence, they need to be estimated using bootstrap resampling.[73, 242] Moreover, the sampling variance obtained from the bootstrap is zero when there are no observations classified as positive in a particular center. As the risk threshold increases, this occurs more and more often. The inverse variance weights are undefined when the sampling variance is zero, and the random effects meta-analysis also requires within-cluster sampling variance estimates larger than zero. In this study, we imputed an arbitrarily small sampling variance, yielding an arbitrarily large weight. The rationale behind this is that the NB is bound to go to zero as the risk threshold increases. Imputing arbitrarily large sampling variances yields decision curves with bumps at high risk thresholds, before decreasing to zero at a risk threshold of one.

Note that clinical utility is defined as having a NB that exceeds the NB of the better of two default strategies, namely treating all patients and treating none of the patients. While the latter is always zero, the former depends on the prevalence and hence differs from center to center. Therefore, a 95% confidence interval for the difference between the NB of the model and the NB of the best default strategy may be assessed to investigate whether the model is likely to have clinical utility in the average center.

6.4.2 Random effects meta-analysis of the net benefit: an example

This section explores random effects meta-analysis of center-specific NBs using a case study of the clinical utility of the LR2 model.

Methods

All IOTA data collected in phases I to III was used in this analysis to maximize the number of available centers. 1066 (18%) of the 5914 observations used for

the validation of the clinical utility of LR2 were also involved in the development of the model. Of the new patients (82%), 2420 were recruited in the same centers that contributed data to the development dataset and 2428 were recruited in new centers. Hence, this validation of LR2 is a mix of apparent and external (that is, temporal and geographical) validation.

We calculated the summary decision curve using a random effects meta-analysis of untransformed center-specific NBs, as described above. SAS software (proc mixed) was used for this purpose (SAS Institute, Cary, NC, USA). This curve was compared to the summary decision curve obtained by pooling all center-specific data and ignoring clustering. The meta-analysis was repeated in subgroups of oncology centers and other centers. In addition, a random effects meta-analysis was performed for the difference of the NB of LR2 with the best default strategy. The LR2 model was refitted to recalibrate the predicted probabilities. We re-estimated the predictor effects, included an effect for center type, and included a random center intercept. The NB of the refitted model was compared to the NB of the original LR2 model. Finally, the NB of LR2 was compared to that of RMI, the SRs assuming malignancy for unclassified masses, subjective assessment by an expert, and the default strategies of treating all and treating none of the patients. The curves for all strategies were fitted using random effects meta-analysis.

In contrast to section 6.3.4, RMI was not dichotomized using a cut-off value for the outcome. Instead, the continuous outcomes of RMI were linked to predicted probabilities by performing a logistic regression in the training data (phase I, the same as for LR2 and SRs) of the true outcome (benign versus malignant tumor) on log(RMI+1). This regression model was then also used to obtain predicted risks based on RMI outcomes in phases Ib, II and III. Missing CA125 values were dealt with using multiple imputation, as described in 7.

Results

There were considerable differences between the observed center-specific decision curves (see Figure 29). These observed differences are largely due to between-center heterogeneity (I^2=97.4 at t=0.1, 95% CI 96.9 to 97.9). The decision curve that was constructed by random effects meta-analysis of NB was lower than the decision curve constructed by pooling all center-specific data (see Figure 29). Some large centers with a high NB (LBE, RIT, PCR, CIT and LPO) contributed a large portion of data to the pooled dataset (respectively 16%, 13%, 6%, 5% and 5%). Hence, they had a high impact on

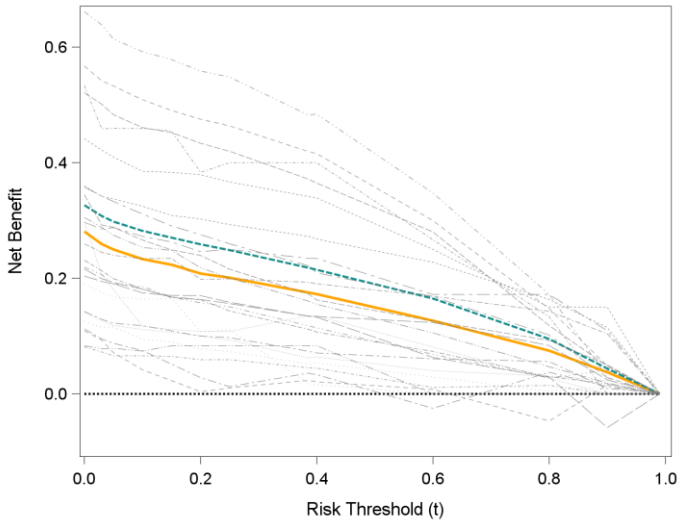


Figure 29. Observed center-specific decision curves (grey lines), the decision curve obtained by pooling all data (bold dashed line) and the decision curve obtained by random effects net benefit (bold full line).

the curve constructed from the pooled dataset, but this effect was moderated by the random effects weighting.

The heterogeneity in NB could partly be explained by differences between oncology and other centers. Malignant tumors were more prevalent in oncology centers, and NB typically increases with the prevalence of events. Separate meta-analyses for oncology centers and other centers revealed that NB was higher in oncology centers than in other centers, and there was more heterogeneity in NB between oncology centers, as also reflected in the broader prediction intervals (see Figure 30). The difference in the NB between LR2 and the best default strategy was 0.01 (95% CI -0.005 to 0.031) at risk threshold 0.1 in oncology centers. LR2 had clinical utility in oncology centers for risk thresholds >15%, and at all risk thresholds in other centers.

The NB was remarkably low in a number of oncology centers. This may have been caused by the miscalibration of LR2: the probability of malignant tumors was underestimated in these centers. Refitting the LR2 model using phase I to III data yielded more accurate predicted probabilities and a higher decision curve (see Appendix 10, Figure A 4). A comparison of summary curves constructed by meta-analysis suggests that the refitted LR2 model and SRsMal outperform the other models and default strategies (i.e., treat all and treat

none) at relevant risk thresholds for cancer detection ($0.05 \leq t \leq 0.5$) (see Appendix 10, Figure A 5).

Discussion

This empirical example illustrates that NB increases because of an increased prevalence of malignancy, and decreases due to insufficient calibration. Hence, NB can be highly heterogeneous in multicenter studies. This heterogeneity should be recognized and explored using appropriate techniques, such as random effects meta-analysis.

The decision curve constructed by random effects meta-analysis was lower than the decision curve computed directly from the pooled dataset. There are two potential explanations for this observation. First, in a random effects meta-analysis in heterogeneous populations, centers get weights closer to each other than when datasets of various sizes from different centers are pooled.[218] In this example, there were large centers with high decision curves that had less influence in the meta-analysis than in the analysis of the pooled dataset. Second, in section 6.4.1 it was shown that placing sampling variability in the denominator of weights may give pessimistic NB summary estimates. The

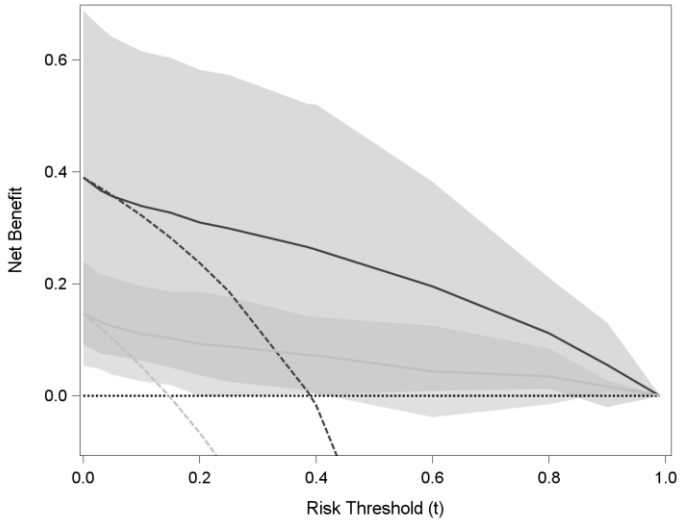


Figure 30. Decision curves obtained by random effects net benefit of oncology centers and other centers. Black full curve and dark prediction interval: LR2 in oncology centers; black dashed curve: treat all in oncology centers; grey curve and light prediction interval: LR2 in other centers; grey dashed curve: treat all in other centers; black dotted horizontal line: treat none.

closer NB is to 0.5, the larger the sampling variance. Hence, centers with a high NB systematically have a low impact on the summary curve. In a random effects-meta-analysis, this may be problematic when there is little to moderate between-center heterogeneity in comparison with the within-cluster sampling variance. The percentage of variability due to between-center heterogeneity rather than sampling error was very large in this example: I^2 was between 98.4% and 81.9% at risk thresholds between 0 and 0.8. Hence, it is not very likely that the summary estimates based on random effects meta-analysis are overly pessimistic in this example.

This case study has a few disadvantages that may hinder the clinical interpretability of the results. First, the validation dataset included the data that was used for model development. The rationale for including all data was to increase the number of centers available for analysis to study heterogeneity in NB. Consequently, this study is not an external validation study of clinical utility, as opposed to the study presented in section 6.3.4. The optimism of the estimated NB is nevertheless likely to be small, as only 18% of the observations were used for model development. Moreover, all models (LR2, SRsMal and the probability model for RMI) were developed on the same dataset, to ensure an honest comparison. Only the refitted LR2 model was developed on the entire dataset, and hence its estimated NB may be overoptimistic. Second, we derived risks for RMI, instead of using RMI with a cut-off for classification, as we did in section 6.3.4. By doing so, we showed the potential clinical utility RMI would have if it were used as a risk model. However, this is not how RMI is used in clinical practice. Moreover, even as a risk model, it had less clinical utility than the IOTA models at the relevant risk thresholds for cancer detection.

6.4.3 Future research: a Bayesian approach

The previous sections have shown that in case of between-center heterogeneity, a random effects meta-analysis of NB may be required to summarize clinical utility in a multicenter validation dataset. Nonetheless, more research is needed to determine how such a meta-analysis should be conducted.

First, research into transformations of NB that break the link between the NB estimate and its sampling variance is needed. For this purpose, the logit and arcsine transformations are often used as variance-stabilizing functions in meta-analyses of prevalence,[243] but these transformations are not defined in the entire range of values that NB can take ($-\infty$ to 1). The $\log(1-\text{NB})$

transformation that was used in section 6.4.1 permits asymmetric confidence and prediction intervals but does not break the correlation between NB estimates and their sampling variance. Hence, the summary estimate of NB may be pessimistic when the sampling variability is large compared to the between-center heterogeneity.

Second, the coverage probabilities of 95% prediction intervals should be studied, with a special focus on high threshold values. The normality assumption for the distribution of center-specific NBs may not hold. As the threshold increases, NB rapidly approaches 0 for most centers, yielding a distribution that is sharply peaked at zero. Nonetheless, the distribution may have heavy tails and be asymmetric. For example, for centers in which the risk model is badly calibrated but discriminates well, negative NBs may be expected at high risk thresholds.[232]

The summary NB can also be computed from meta-analysis summary estimates of sensitivity, specificity and prevalence, as an alternative to performing the meta-analysis directly on the NB. This has the advantage that one does not have to assume a normal distribution for the NB, nor does one need to use the sampling variability of the NB. One can instead specify the model in terms of the number of true positives (r_{11j}), true negatives (r_{00j}) and cases (n_{1j}) in each center j, by extending model (34) as follows:

(39)

$$n_{1j} \sim \text{bin}(n_j, p_j)$$

$$r_{11j} \sim \text{bin}(n_{1j}, Se_j)$$

$$r_{00j} \sim \text{bin}(n_{0j}, Sp_j)$$

$$\begin{pmatrix} \text{logit}(Se_j) \\ \text{logit}(Sp_j) \\ \text{logit}(p_j) \end{pmatrix} \sim N\left[\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}, \boldsymbol{\Omega}\right], \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \rho_{13}\tau_1\tau_3 \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 & \rho_{23}\tau_2\tau_3 \\ \rho_{13}\tau_1\tau_3 & \rho_{23}\tau_2\tau_3 & \tau_3^2 \end{pmatrix}$$

$$NB = \frac{\exp(\gamma_1)}{1+\exp(\gamma_1)} \times \frac{\exp(\gamma_3)}{1+\exp(\gamma_3)} - \left(1 - \frac{\exp(\gamma_2)}{1+\exp(\gamma_2)}\right) \times \left(1 - \frac{\exp(\gamma_3)}{1+\exp(\gamma_3)}\right) \times \frac{t}{1-t}.$$

The summary NB is estimated from γ_1, the summary logit sensitivity, γ_2, the summary logit specificity, γ_3, the summary logit prevalence and t, the risk threshold of interest. In a Bayesian framework, using vague prior distributions

Table 23. Summary sensitivity, specificity, net benefit and the probability that ear thermometry will be harmful in a new study, calculated using a Bayesian approach.

	t=0.2	t=0.5	t=0.8
Sensitivity	0.64	0.64	0.64
Specificity	0.98	0.98	0.98
NB ear	0.30	0.29	0.25
95% CrI	0.19 to 0.41	0.18 to 0.40	0.15 to 0.36
95% PI	0.05 to 0.64	0.04 to 0.63	0.01 to 0.59
NB treat all	0.33	-0.07	-1.68
95% CrI	0.23 to 0.44	-0.24 to 0.10	-2.10 to -1.25
95% PI	0.03 to 0.66	-0.56 to 0.44	-2.90 to -0.34
P(ear harmful)	0.70	0.004	0.02

NB: net benefit; CrI: credible interval; PI: prediction interval; t: risk threshold.

for γ_1, γ_2, γ_3 and their variance-covariance matrix (e.g., normal distributions and a Wishart distribution), sensitivity, specificity and prevalence can be sampled from their joint posterior distribution to obtain a 95% prediction interval for the NB in new centers. By calculating the NB of treat all, one can also obtain the probability that the model will be harmful in a new center. We applied these methods to an example of the accuracy of infrared ear thermometry for the diagnosis of fever in children, taking an electronic rectal thermometer as the reference standard. The data were taken from a study by Riley and colleagues[234] and comprise the number of true positives, true negatives, cases and non-cases for eleven studies. We investigated three risk thresholds (0.2, 0.5 and 0.8).

We were able to estimate the summary NB, a 95% credible interval, a 95% prediction interval, and the probability that using ear thermometry would be found harmful in a new study (see Table 23). Assuming that an undetected case of fever is four times more harmful than a false positive (t=0.2), ear thermometry does not have clinical utility to diagnose fever, as it would be harmful (i.e., worse than classifying none of the children or all children as having fever) in 70% of studies. In contrast, assuming that an undetected fever is equally harmful as a false positive (t=0.5), or assuming that a false positive is four times more harmful than an undetected fever (t=0.8), one must conclude that ear thermometry has clinical utility to diagnose fever, as it would be harmful in respectively only 0.4% and 2% of new studies.

This preliminary example shows that a Bayesian approach to a meta-analysis of NB is promising, but more research is needed. For example, it must be investigated whether it is feasible to estimate the three-by-three variance-covariance matrix of sensitivity, specificity and prevalence in other examples, and whether the Wishart distribution is appropriate as an uninformative prior.

6.5 Conclusion

Risk models are often developed to support decision-making in hospitals. The patient case-mix or definitions of predictors may differ from one hospital to another, and this may influence the predictive performance. For example, predicted probabilities may be too high in centers that have a low prevalence compared to others. To investigate the performance at the center and the heterogeneity in predictive performance across centers, this chapter proposed to use techniques that were originally developed for the evaluation of diagnostic tests in IPD meta-analyses. For example, a random effects meta-analysis of the c-statistic provides a summary measure of discrimination across centers, and a 95% prediction interval can reflect the heterogeneity in discrimination that may be expected in new centers. Sensitivity and specificity can be analyzed in a bivariate random effects meta-analysis model that accounts for the correlation between both measures. The heterogeneity in a model's discriminatory performance may be explained by factors such as the background training of the users applying the model, and these may be incorporated in the meta-analysis model. Calibration can be investigated by a one- or two-stage meta-analysis of calibration intercepts and slopes. Alternatively, one can visualize flexible calibration curves for large centers or for relevant subgroups of centers. The NB transcends measures of discrimination and calibration and evaluates the clinical utility of the model for decision-making. The NB may be evaluated separately for meaningful subgroups of centers. A random effects meta-analysis of NB is hampered by the correlation between the NB point estimate and its variance. Computing the summary NB and its prediction interval in a Bayesian trivariate meta-analysis of sensitivity, specificity and prevalence may be more appropriate than performing the meta-analysis directly on the NB, but more research is needed in this area.

7 Does ignoring clustering in multicenter data influence the predictive performance of risk models? A simulation study

This chapter presents a comprehensive framework for the evaluation of the predictive performance of risk models at the center level and the population level, considering population-averaged predictions, center-specific predictions and predictions assuming an average random center effect. Section 7.1 is an introduction. Section 7.2 presents the generalized framework for performance evaluation. Section 7.3 reviews what is known about the difference between marginal and conditional regression coefficients and deduces what this implies for model calibration. Section 7.4 presents a simulation study that investigates the performance of mixed effects and standard logistic regression models within the proposed framework, and the influence of sample size. Section 7.5 presents an example on the prediction of the risk of tumor malignancy. Section 7.6 discusses the main findings and section 7.7 concludes. Hence, this chapter provides a synthesis of elements treated in chapters 3, 4 and 6, and answers the remaining question: does accounting for clustering improve the predictive performance?

This chapter is based on the following research paper:

Wynants L, Vergouwe Y, Van Huffel S, et al. Does ignoring clustering in multicenter studies influence the performance of prediction models? *Stat Methods in Med Research*. 2016: e-pub ahead of print.

7.1 Introduction

An important aspect of a risk model is its performance in new individuals. To this date, the literature has not yet provided evidence that a mixed effects model's predictive performance is superior to a standard regression model's,

nor is it clear how exactly predictions for individuals from new centers should be obtained from mixed effects models. In previous research comparing a standard logistic regression model and a mixed effects logistic regression model, the random intercept was substituted with zero in order to make predictions for new centers which were not included in the dataset used for model development (see section 3.2.2).[152] This way, predictions for new individuals assume an average random center effect. The mixed effects model produced miscalibrated results at the population level, that is, the predicted probabilities of experiencing the event did not reflect the observed probabilities in the population. Calibration slopes deviated from one and the miscalibration was worse when the degree of clustering (the intraclass correlation) increased. Pavlou and colleagues recently pointed out that calibrated results can be obtained with the mixed effects logistic regression model if marginal predictions are used.[153] As explained in section section 3.2.2, these are obtained by integrating over the estimated random effects distribution, rather than substituting the random intercept by zero.[151] Pavlou, however, focused solely on the predictive performance of the model at the population level, while others have distinguished between performance at the population level and at the center level.[217, 218] In chapter 6, the relevance of the predictive performance at the center level was stressed. This chapter investigates whether a mixed effects logistic regression model has a better predictive performance in terms of calibration and discrimination than a standard logistic regression model in clustered data, by assessing performance at the center level and the population level.

7.2 A framework of performance evaluation of risk models in clustered data

Risk models can be developed with standard or mixed effects regression analysis. Chapter 3 described how the standard logistic regression model yields marginal predictions (equation (3)), which have a population-averaged interpretation. The mixed effects model yields three types of predictions: conditional, center-specific predictions (equation (10)), predictions for an individual in a center with an average random intercept (equation (11)), and marginal predictions (equation (12)). The predictive performance of a model is crucial and needs extensive evaluation, preferably using data from new clinical settings. Validation data can be obtained from a single center or from multiple centers; the latter may enhance the generalizability of the results.

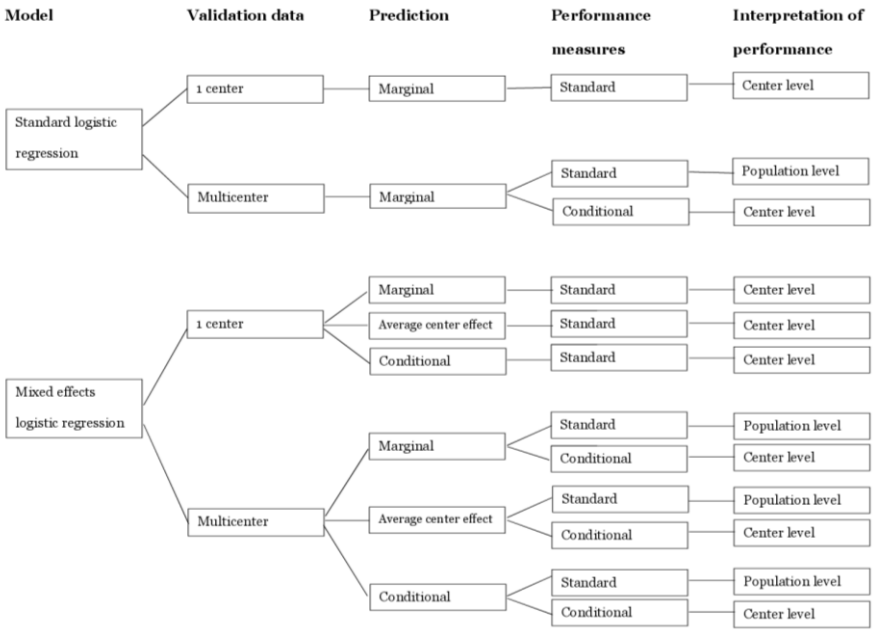


Figure 31. A comprehensive framework of options for model validation, subject to the type of risk model that is being evaluated (standard or mixed effects logistic regression) and the available validation dataset (one center or multicenter).

Chapter 6 has demonstrated the potential between-center heterogeneity of the predictive performance of a risk model and described methods that evaluate the performance within centers, as an alternative to standard methods that ignore the heterogeneity and describe only the performance in the entire population.

It may seem natural to use conditional (within-center) measures of predictive performance only for conditional predictions and standard (population level) performance measures for marginal predictions. However, there are many other options to evaluate the predictive performance of a model, and Figure 31 places them in a comprehensive framework. The choice of performance measure in multicenter validation data should depend on the use of the risk model and the research question. The conditional performance measures should be used to assess the performance within centers. Consider a model predicting the risk that an ovarian mass in a patient is malignant.[183] The treatment decision is made in the center the patient is treated in, requiring adequate conditional performance of the risk model. Conditional performance measures are not useful when the validation dataset contains data from a single

center. Therefore, that situation is not considered in this chapter, although Figure 31 includes the option for completeness. When a model is validated in a single center, the validation results may not be generalizable to other centers. When multicenter data is available, standard performance measures will quantify how well the model performs in the entire population of individuals, as an overall measure of performance. This is useful, for example, for the recommendation of a risk model in national guidelines. Sometimes, the risk model is used for decision support at a higher level than the cluster level. For example, a recently developed prediction model to screen for Chlamydia trachomatis infection was developed in a dataset that was clustered within neighborhoods.[244] Given that the intended user of the model screens individuals from various neighborhoods, population-level performance measures were of interest. Appendix 11 presents an overview of the formulas for the c-statistic and logistic calibration in this comprehensive framework.

7.3 Calibration slopes for marginal and center-specific logistic regression models

Marginal effect estimates from the standard logistic regression model are typically closer to zero than conditional effect estimates from the mixed effects model.[138, 139, 230] The estimates differ by a factor f, with $f = \sqrt{1+\tau^2 c^2}$ and $c=(16\times\sqrt{3})/(15\times\pi)$ (see equation (30)). This implies that, when a standard logistic regression model has an overall calibration slope β_{cal}, the overall calibration slope of the corresponding mixed effects model using the average random intercept could be approximated by β_{cal}/f:

(40)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+\beta_{cal}LP_{LR\,ij}$$

$$=\alpha_{cal}+\beta_{cal}(\hat{\alpha}_{marginal}+\sum_{k=1}^{l}\hat{\beta}_{k\,marginal}X_{ki})$$

$$\approx\alpha_{cal}+\frac{\beta_{cal}}{f}(\hat{\alpha}_{conditional}+\sum_{k=1}^{l}\hat{\beta}_{k\,conditional}X_{ki}).$$

Likewise, when a mixed effects model has a within-center calibration slope $\beta_{cal\,w}$ assuming an average random center effect ($a_{j\,cal} = b_{j\,cal} = a_j = b_j = 0$),

the calibration slope of the corresponding standard model would be approximated by $\beta_{cal\,w} \times f$:

(41)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_{cal\,w} + a_{j\,cal} + \beta_{cal\,w} LP_{MLR\,a\,ij} + b_{j\,cal} LP_{MLR\,a\,ij}$$

$$= \alpha_{cal\,w} + \beta_{cal\,w}(\hat{\alpha}_{conditional} + \sum_{k=1}^{l} \hat{\beta}_{k\,conditional} X_{kij})$$

$$\approx \alpha_{cal} + \beta_{cal\,w} f(\hat{\alpha}_{marginal} + \sum_{k=1}^{l} \hat{\beta}_{k\,marginal} X_{kij}).$$

This demonstrates how the calibration slope may deviate from one due to the choice of modeling technique. For example, if a risk model was fitted using mixed effects logistic regression and this model was perfectly calibrated in a center with an average effect, the corresponding standard model would have a within-center calibration slope larger than one in a center within an average effect.

In practice, the random effect variance τ^2 will often be estimated with error. As shown in Appendix 12, overestimation will decrease the estimated calibration slope, while underestimation has the opposite effect. Fitting a standard model can be seen as an extreme case of the latter, setting the estimated between-center variance to zero.

7.4 Simulation study

This simulation study compares the performance of mixed effects and standard logistic regression models in multicenter validation data at the center and the population level. The design of the study is described first, followed by the results.

7.4.1 Design

We first created source populations, from which samples with different sizes were drawn. We fitted a random intercept model and a standard logistic regression model in each sample and tested them in the remaining part of the

source population, within the framework for performance evaluation presented in section 7.2.

We generated two source populations of approximately 20,000 patients: one population with heavily clustered data (intraclass correlation (ICC)=20%) and one with little clustering (ICC=5%).[216] We fixed the number of centers (J) at 20.[141] The number of patients per center (n_j) was drawn from a Poisson distribution with a separate, randomly generated lambda for each center, yielding center sizes ranging from approximately 600 to 2000.

We generated the data for the source populations according to a predefined true random intercept model. Each center was assigned a random center intercept a_j, generated from a normal distribution of which the variance was determined by the desired ICC. The true model included four normally distributed continuous predictors and four dichotomous predictors, each with a regression coefficient of 0.8. X_1 through X_4 were continuous with mean 0 and standard deviations 1, 0.6, 0.4 and 0.2, respectively; X_5 through X_8 were dummy variables with prevalence 0.2, 0.3, 0.3 and 0.4, respectively. We set the overall intercept α equal to -2.1 to obtain an event rate of the outcome Y_{ij} of 0.30. We computed the probability of an event (p_{ij}) for each patient from the generated predictors and random intercepts, using equation (7) and applying the inverse logit transformation. We generated Y_{ij} by comparing p_{ij} to a randomly drawn value from a uniform distribution:

(42)

$$U_{ij} \sim \text{unif}(0,1)$$
$$Y_{ij} = \begin{cases} 1 & \text{if } U_{ij} \leq p_{ij} \\ 0 & \text{if } U_{ij} > p_{ij} \end{cases}.$$

We drew samples from the source population with either 100 (for ICC=5% and ICC=20%) or 5 (only for ICC=20%) events per variable (EPV). The number of events to be sampled was calculated by multiplying the preset EPV value by nine (eight parameters for the regression coefficients plus one extra parameter for the random intercept variance). The required number of non-events to be sampled was computed in such a way that the event rate in the source population (0.3) was preserved. We sampled patients without replacement from all centers, without stratification for center. Each simulation was based on 1000 samples.

We built a random intercept logistic regression model and a standard logistic regression model containing all eight predictors in each sample. We used the following convergence criteria for the mixed effects model: a change of less

than 10^{-5} in deviances of the models fitted in the last two iterations, 10 to 100 iterations to fit the model, and no outlying estimated regression coefficients and standard errors (visual inspection). For the standard model, we used a positive convergence tolerance of 10^{-9}, a maximum of 50 iterations and a visual check of estimated regression coefficients and standard errors as convergence criteria. Samples with non-converging models were removed from the analysis.

We tested the models in the part of the source population that was not used for model development and as a result, the development set and the validation set are from the same population. We used two versions of the linear predictor for the random intercept model: the conditional linear predictor, including the center-specific intercept estimates (equation (10)) and the linear predictor assuming an average random intercept (equation (11)). The marginal linear predictor was obtained from the standard logistic regression model (equation (3)). Marginal predictions from the standard logistic regression model are very similar to marginalized predictions from the mixed effects logistic regression model (see Appendix 13). We computed the standard (equation (18)) and within-center calibration slopes (equation (20)) and intercepts (equations (31) and (36)), and the standard (equation (17)) and within-center c-statistic (equation (19)) for all predictions.

All simulations and calculations were performed in R version 2.14.0 (http://www.r-project.org/). The lmer function from the lme4 package[219] was used to fit mixed effects logistic regression models using Laplace approximation and the rms package was used for model evaluation.[5] The R code is provided in Appendix 14.

7.4.2 Results

This section first considers the results of the simulations with large samples (EPV=100) in situations with severe (ICC=20%) and moderate (ICC=5%) clustering. Next, it considers the results of the simulation with small samples (EPV=5) and severe clustering (ICC=20%).

Calibration and discrimination of models developed in large samples (EPV 100)

The conditional predictions from the random intercept model were well calibrated in samples with severe clustering (ICC=20%) at both the center level and the population level (Figure 32, squares; estimates are tabulated in

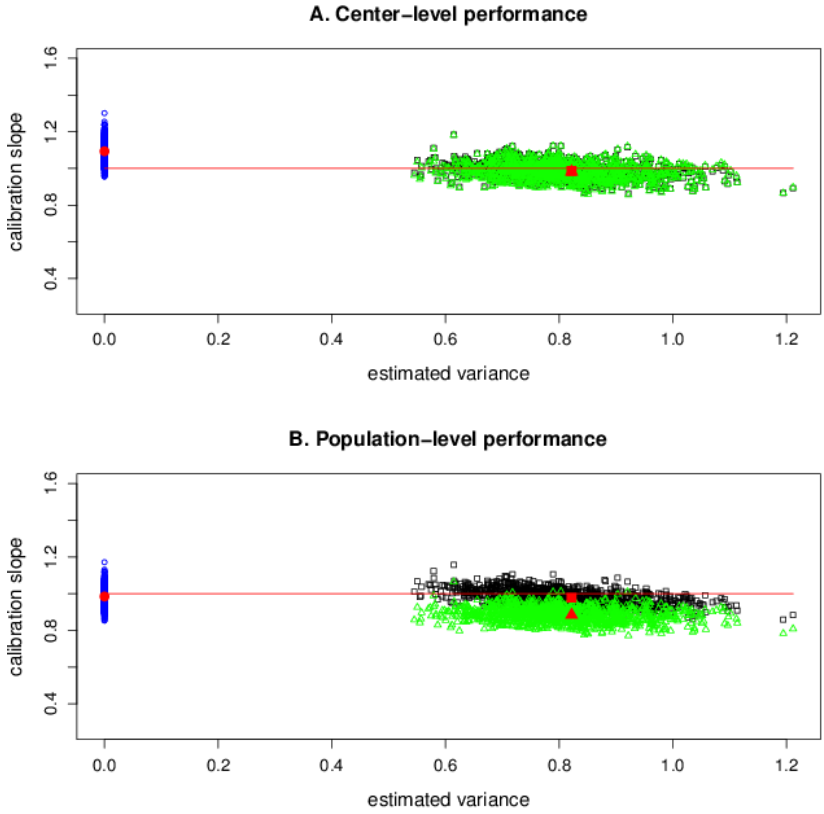


Figure 32. Center-level (panel A) and population-level (panel B) calibration slopes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor of the random intercept model assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.822 (ICC=20%). Small symbols indicate calibration slopes in the samples, large filled symbols indicate average calibration slopes at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.822) for the random intercept model. The horizontal line represents the ideal calibration slope.

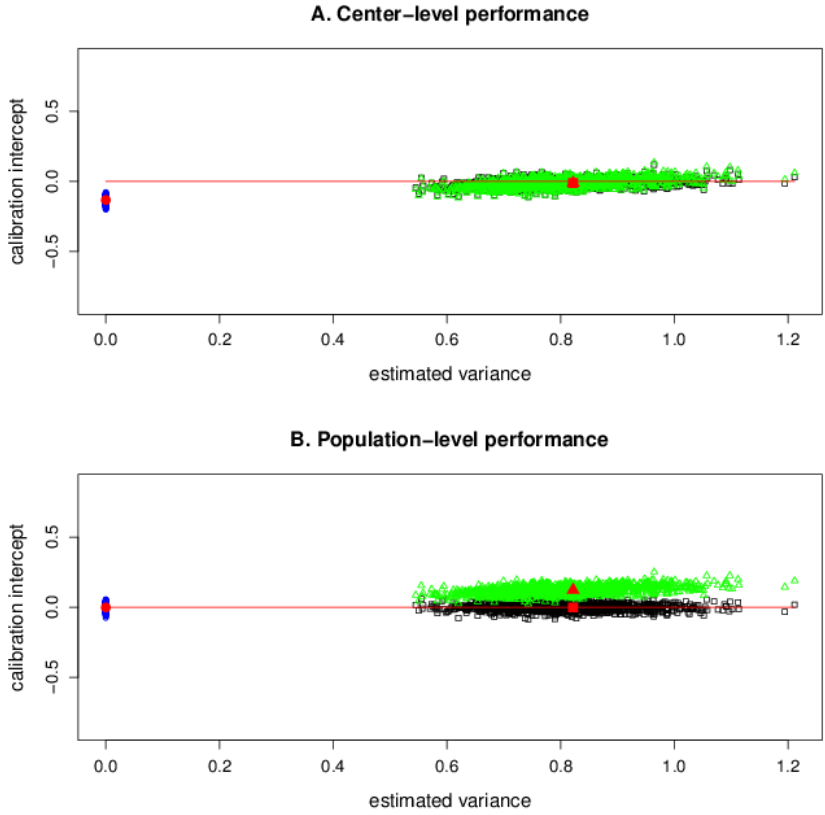


Figure 33. Center-level (panel A) and population-level (panel B) calibration intercepts of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor of the random intercept model assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.822 (ICC=20%). Small symbols indicate calibration intercepts in the samples, large filled symbols indicate average calibration intercepts at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.822) for the random intercept model. The horizontal line represents the ideal calibration intercept.

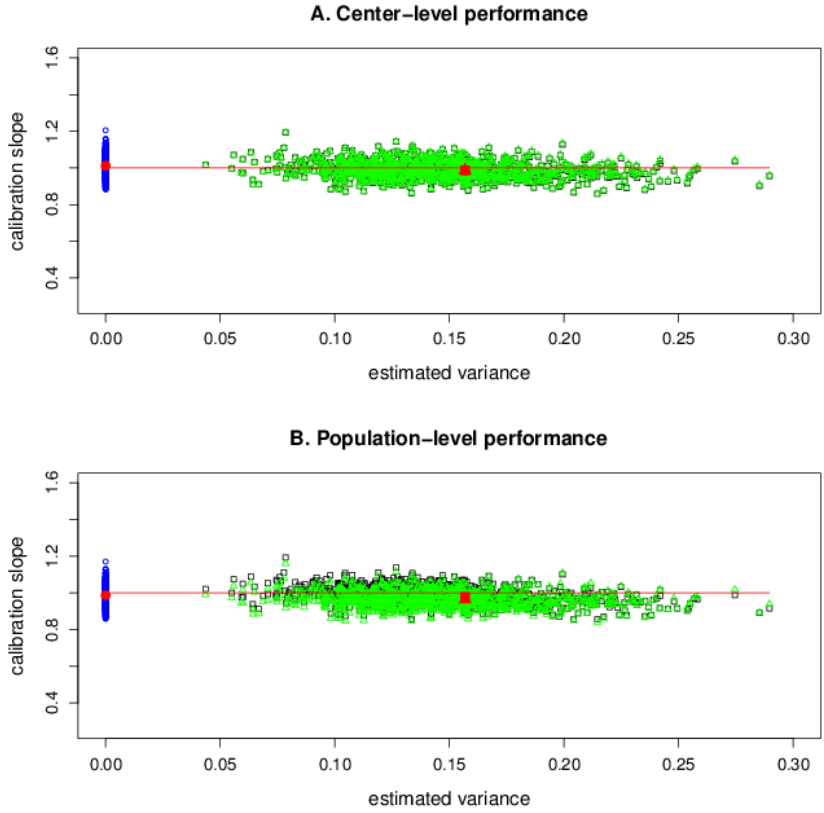


Figure 34. Center-level (panel A) and population-level (panel B) calibration slopes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.157 (ICC=5%). Small symbols indicate calibration slopes in the samples, large filled symbols indicate average calibration slopes at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.157) for the random intercept model. The horizontal line represents the ideal calibration slope.

Table A 7 in Appendix 15). The average calibration slopes close to one indicate that there was hardly any overfitting. The predictions from the random intercept model assuming average random intercepts were only calibrated at the center level (Figure 32A, triangles), while the predictions from the standard model were only calibrated at the population level (Figure 32B, circles). The center-level calibration slopes were larger than one for the predictions of the standard model (Figure 32A, circles) and the population-level calibration slopes were smaller than one for the predictions assuming an average random intercept (Figure 32B, triangles).

The association between the estimated random intercept variance and the within-center calibration slope is slightly negative and close to the theoretical approximation (see Figure A 7 in Appendix 15). Note that the within-center calibration slopes plotted in Figure 32 reflect the calibration slopes in the center with an average calibration slope, while the estimated $b_{cal\,j}$ (not shown) reflect center-specific differences from this slope. The average estimated variance of the $b_{cal\,j}$ was <0.0005 for the three types of predictions.

Calibration-in-the-large was also good for conditional predictions at the population and the center level, while the predictions assuming average random intercepts were only calibrated at the center level and the predictions from the standard model were only calibrated at the population level (Figure 33). The average estimated variance of the center-specific calibration intercepts was 0.83 for the predictions assuming an average random intercept and 0.89 for the predictions from the standard model. The conditional predictions yielded a much lower average estimated variance of center-specific calibration intercepts (0.05), indicating that most of the between-center differences in the event rates are accounted for by using random intercepts in the prediction.

The results from simulations with moderate clustering (ICC=5%) were similar to the results of the simulation with severe clustering, although differences in calibration between the three types of predictions were smaller due to the lower between-center variance (Figure 34 and Appendix 15, Figure A 8). The predictions from the standard model yielded within-center calibration slopes slightly above one (Figure 34A, circles), while the predictions assuming an average random intercept yielded population-level calibration slopes slightly below one (Figure 34B, triangles).

The empirical Bayes estimates are constant within each center and therefore do not influence the estimated within-center c-statistics. Hence, the obtained within-center c-statistics of the conditional predictions and the predictions for

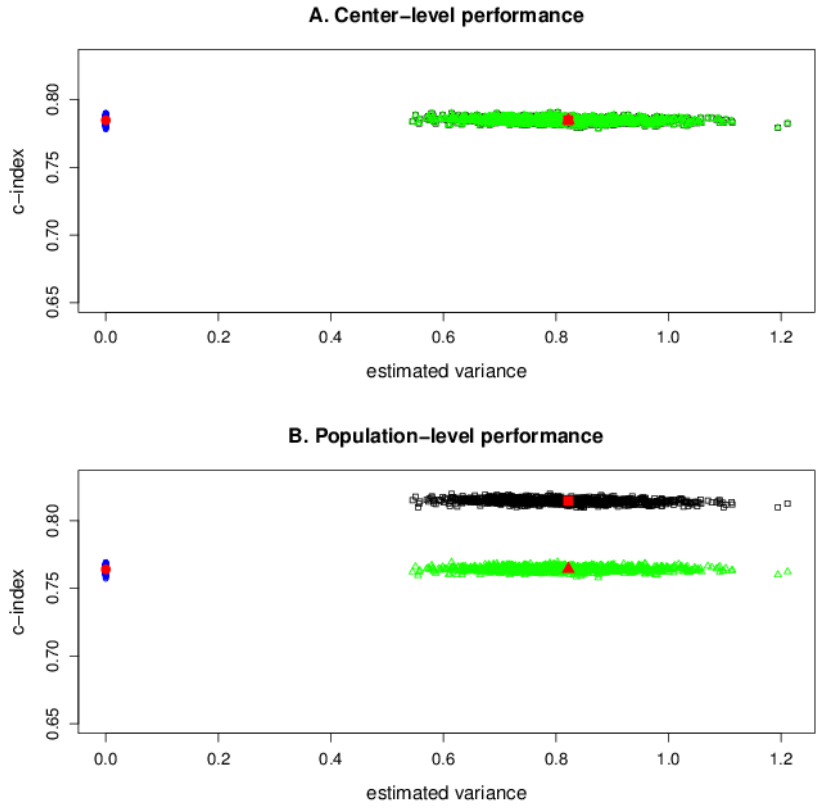


Figure 35. Center-level (panel A) and population-level (panel B) c-statistics of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.822 (ICC=20%). Small symbols indicate c-statistics in the samples, large filled symbols indicate average c-statistics at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.822) for the random intercept model.

an individual from an average center are by definition the same, and they are very similar to the within-center c-statistics of the standard model (Figure 35A). The population-level c-statistics for the predictions assuming an average random intercept (Figure 35B, triangles), were very similar to the population-level c-statistics for the predictions from the standard model. Higher population-level c-statistics were obtained with the conditional predictions, and this effect was even present in datasets with a low ICC (Appendix 15, Figure A 9, squares).

Calibration and discrimination of models developed in small samples (EPV 5)

The effects of the choice of the modeling technique on the calibration slope may in practice often be difficult to observe, as a result of the simultaneous effect of overfitting. 31 of 1000 (3%) simulations with EPV=5 were removed as they did not include patients from all centers, making it impossible to give conditional predictions in some centers. One sample (0.1%) was removed on account of a change of more than 10^{-5} in the deviance of the last two iterations and one sample (0.1%) was removed because it was outlying in a plot of estimated regression coefficients versus standard errors. It is reassuring that the performance of non-converged models was similar to the performance of converged models.

The random intercept variance was often underestimated (median observed variance 0.66, while the true variance is 0.82) and the estimates were very variable across samples, due to the small sample size (Figure 36). Calibration slopes were also highly variable and mostly smaller than one for all types of predictions, indicating overfitting (Figure 36). The predictions from the standard model gave within-center calibration slopes slightly closer to one than the conditional predictions and the predictions assuming an average center intercept (Figure 36A). This is the result of two opposite effects: small sample sizes lead to overfitting and decreased calibration slopes, whereas using marginal predictions leads to increased calibration slopes at the center level. The conditional predictions had a slightly lower population-level calibration slope than the predictions from the standard logistic regression model, whereas the predictions assuming an average random intercept gave the lowest population-level calibration slopes (Figure 36B).

The calibration-in-the-large was good for the conditional predictions, both at the center and the population level (Figure A 10 in Appendix 15). The within-center calibration intercepts were smaller than zero for the standard model, indicating that on average, predicted probabilities were too high. The population-level calibration intercepts were larger than zero for the

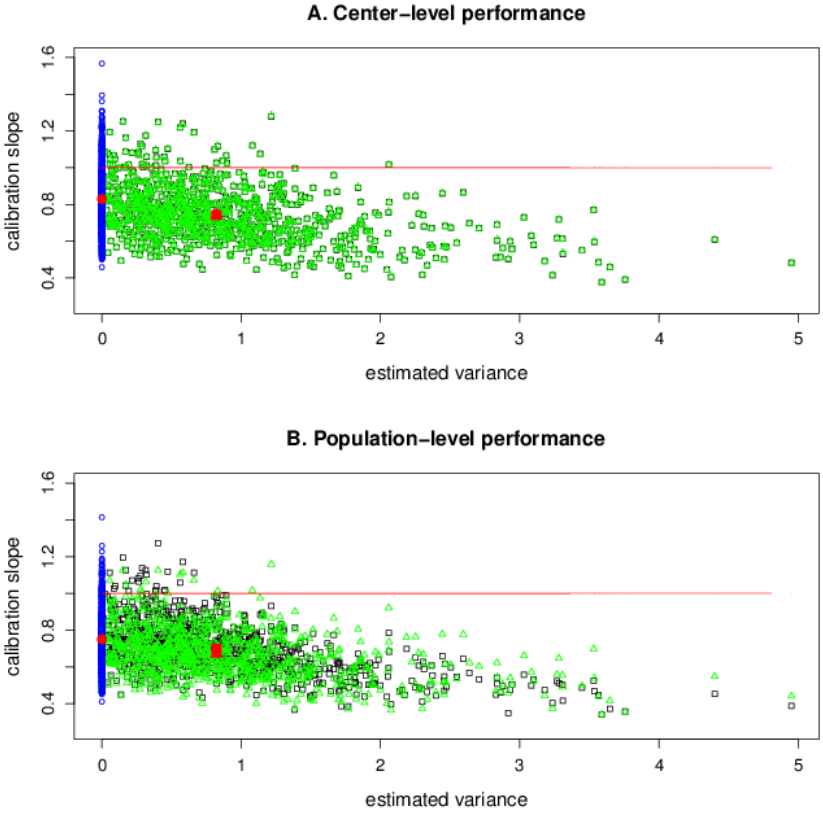


Figure 36. Center-level (panel A) and population-level (panel B) calibration slopes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor of the random intercept model assuming an average random intercept (triangles), by estimated random intercept variance in samples with 5 events per variable and true random effects variance=0.822 (ICC=20%). Small symbols indicate calibration slopes in the samples, large filled symbols indicate average calibration slopes at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.822) for the random intercept model. The horizontal line represents the ideal calibration slope.

predictions assuming an average random intercept, indicating that on average, predicted probabilities were too low (Figure A 10 in Appendix 15).

Discrimination was worse with small development samples than with large development samples, due to overfitting (Figure A 11 in Appendix 15). Higher population-level c-statistics were obtained with the conditional predictions than with the predictions from the standard model and the predictions assuming an average random intercept. The random intercept variance was sometimes estimated close to zero in the small development samples. In these cases, the c-statistic obtained using conditional predictions dropped to the level of the c-statistic of the standard model.

7.5 Empirical example

To illustrate the findings on real data, we developed and evaluated models to preoperatively diagnose ovarian cancer. The development dataset consisted of 3506 women with ovarian masses (949, 27% with malignancies), collected by the IOTA consortium during phases I, Ib and II (between 1997 and 2007) in 21 international centers. We used six clinical and ultrasound predictors: age, the proportion of solid tissue, the presence of more than ten locules, the number of papillary structures (0, 1, 2, 3, >3, linear effect), the presence of acoustic shadows, and the presence of ascites. This yielded an EPV of 136 for the random intercept model. The RICC was 15% ($\hat{\tau}^2$=0.59), accounting for

Table 24. Estimated regression coefficients from a mixed effects logistic regression model and a standard logistic regression model to predict the risk of ovarian mass malignancy.

	Mixed effects model $\hat{\beta}_{cond}$**(se)**	**Standard model** $\hat{\beta}_{marg}$**(se)**
Intercept	-4.729 (0.280)	-4.520 (0.203)
Age (in years)	0.036 (0.004)	0.035 (0.003)
Proportion of solid tissue	3.655 (0.189)	3.614 (0.181)
More than 10 locules (yes/no)	1.858 (0.191)	1.728 (0.183)
Number of papillations (0,1,2,3,>3)	0.633 (0.045)	0.592 (0.043)
Acoustic shadows (present/not present)	-2.151 (0.242)	-2.161 (0.232)
Ascites (present/not present)	2.975 (0.203)	2.848 (0.193)

the predictors. The regression coefficients of the standard model tended to be closer to zero than those of the mixed effects model, apart from the coefficient of acoustic shadows, whereas the standard errors were larger in the mixed effects model (Table 24).

All predictions (marginal, with average random intercept and conditional) were validated using conditional and standard performance measures (Figure 31), in a dataset of 2224 women (915 (41%) with malignancies), collected during IOTA phase III (between 2009 and 2012) in 15 of the 21 centers of the development set. The RICC was 14% ($\hat{\tau}^2$=0.53) after accounting for the linear predictor of the mixed effects model, assuming an average random intercept.

The calibration slope at the population level was close to one for the marginalized predictions from the random effects model (0.99) and slightly lower for the marginal predictions from the standard model (0.95). As expected, the calibration slope was lower for the predictions assuming an average random intercept (0.91). Surprisingly, the calibration slope of the conditional predictions was also lower (0.88) (Table 25). It is likely that this is due to differences in the true random center intercepts between the development and validation datasets.

The within-center calibration slopes for the predictions assuming an average random intercept and for the conditional predictions were slightly below 1 (0.94 and 0.93) (Table 25). The within-cluster calibration slope for the

Table 25. Population-level and center-level discrimination and calibration statistics for the marginal predicted risks from the standard logistic regression model and for the marginalized predicted risks from the mixed effects model, the predictions assuming an average random intercept, and the conditional predicted risks.

	Population level			**Center level**		
Prediction	**C**	**Cal slope**	**Cal int**	**C**	**Cal slope**	**Cal int**
Marginal	0.90	0.95	0.62	0.88	0.97 (0.014)	0.41 (0.554)
Marginalized from mixed effects model	0.90	0.99	0.60	0.88	1.02 (0.017)	0.39 (0.529)
Average cluster effect	0.90	0.91	0.70	0.88	0.94 (0.015)	0.48 (0.582)
Conditional	0.91	0.88	0.29	0.88	0.93 (0.009)	0.27 (0.248)

C: c-statistic; cal slope: calibration slope; cal int: calibration intercept. Numbers between brackets are the random variances of random center calibration intercepts and calibration slopes.

standard model was higher (0.97) than the within-center calibration slope for predictions assuming an average random center intercept, which is typical. The within-center calibration slope for the marginalized predictions from the mixed effects model was 1.02. The random variance of the center-specific calibration slopes was nearly half as large for the conditional predictions as for all other predictions, which indicates that the center-specific calibration was more stable when conditional predictions were used (Table 25).

The population-level calibration intercept was 0.29 for the conditional predictions, 0.62 for the marginal predictions from the standard model,0.60 for the marginalized predictions from the mixed effects model, and 0.70 for the predictions assuming an average random intercept. This is explained by the changed event rates within the centers: in 12 out of the 15 centers in the validation dataset, the event rate was higher than in the development set. The within-center calibration intercept was 0.27 for the conditional predictions, 0.41 for the marginal predictions from the standard model, 0.39 for the marginalized predictions from the mixed effects model, and 0.48 for the predictions assuming an average random intercept. The conditional predictions yielded the within-center calibration intercept closest to zero and the estimated variances of the center-specific calibration intercepts were half as large for conditional predictions as for all other types of predictions (Table 25).

At the center level, all predictions yielded a very similar c-statistic (0.88). At the population level, the discrimination of the conditional predictions was superior (0.91) to the other predictions (0.90). The discrepancy might have been higher, if the true random center intercepts in the validation data were more like the ones in the development data.

7.6 Discussion

This chapter investigated whether ignoring clustering in multicenter data influences the predictive performance of a risk model, comparing standard to mixed effects logistic regression. The results have shown that it does, but the consequences of ignoring clustering are dependent on the level at which the model is evaluated (the population or the center level) and on the aspect of predictive performance that is evaluated (calibration or discrimination) (Table 26).

Predictions from mixed effects models assuming an average random intercept are poorly calibrated at the population level, whereas marginal predictions are

Table 26. Schematic overview of the effect of the type of prediction on the conditional and standard performance measures, in the absence of overfitting and assuming a representative development dataset.

Performance measure		Marginal predictions	Predictions assuming an average random center intercept	Conditional predictions
Calibration	Conditional (center level)	Cal slope > 1 Cal intercept ≠ 0	Well calibrated	Well calibrated
	Standard (population level)	Well calibrated	Cal slope < 1 Cal intercept ≠ 0	Well calibrated
Discrimination	Conditional (center level)	Good discrimination	Good discrimination	Good discrimination
	Standard (population level)	Good discrimination	Good discrimination	Superior discrimination

Cal: calibration.

poorly calibrated at the center level. We showed that this is a consequence of the much-described finding that marginal regression coefficients are typically closer to zero than conditional regression coefficients.[138, 139, 230] The consequence, from a calibration perspective, is that predicted probabilities from a mixed effects model assuming an average random intercept model are too extreme to reflect the event rates in the entire population.[6, 65] For instance, less than 80% of patients with a predicted risk of 0.8 will experience the event, while of all patients in that center with a predicted risk of 0.2, more than 20% will experience the event. In contrast, conditional predictions from the mixed effects model (that include center-specific effects) were well calibrated at both the population level and the center level (Table 26). This is in line with earlier research showing that conditional predictions from mixed effects models yield better calibration-in-the-large at the center level.[152] Hence, we advise to use a mixed effects model to obtain better within-center calibration, but we must note that the degree of clustering in typical outcomes of risk models is generally small. The simulations in a source population with moderately weak clustering (ICC 5%) have shown that the calibration results of the standard logistic regression model and the mixed effects logistic regression model are very similar.

Additional simulations in small samples (EPV=5) showed that overfitting yields poorly calibrated results, both for standard and mixed effects logistic regression models. Calibration was poorer for the mixed effects model, because the problem of overfitting in small datasets was worsened by the fact that conditional regression coefficients are generally more extreme than marginal regression coefficients. Although the standard model was seemingly better calibrated, ignoring clustering is not an adequate solution for problems

caused by small sample sizes. The simulations also showed that the calibration of mixed effects models depends on the estimation of the between-center variance in heavily clustered data. Previous research has shown that the accuracy of the estimated between-center variance can be improved by increasing the number of centers.[146, 147, 215]

Discrimination at the population level was better for the conditional predictions obtained by mixed effects logistic regression than for the other predictions (Table 26). This was even observable when the degree of clustering was low. Center-specific intercept estimates contain additional information when comparing predicted probabilities for patients from different centers, enhancing discrimination.

The results of the simulation study suggest that conditional predictions perform the best in prediction. Center-specific random intercepts are required for conditional predictions, but these are not available for new centers. In the absence of data from the new center, numerical integration of the predictions over the estimated random effects distribution may be used. However, these marginal predicted probabilities will not be well calibrated at the center level. Another option is to substitute the center-specific random effect by zero. These predictions are easy to obtain and will be well calibrated in centers with an average effect. Alternative options are to estimate the center-specific intercept from the outcome prevalence of the new center or to use the intercept of a similar center from the model development set.[142, 143]

This study has the following limitations. It only considered mixed effects logistic regression to account for clustering. Other methods are available, such as fixed effects logistic regression with dummy variables for centers. Like the mixed effects logistic regression model, it offers center-specific predictions and the regression coefficients have a conditional interpretation (see section 3.2).[136] Hence, it may perform similarly to the mixed effects regression model in terms of discrimination and calibration. The optimal choice most likely depends on the number of centers, mixed effects models being more appropriate if the number of clusters is large.[134, 136, 140, 141, 146, 147, 215] Further, we assumed that the assumptions underlying the regression models hold. For example, we assumed that the random intercepts were normally distributed. This may not always be the case in practice but evidence to date suggests that random effects models are quite robust against violations of this assumption.[148, 149, 150] Random slopes were beyond the scope of this research. More research on how random slopes can be included in the development and external validation of risk models is needed.

7.7 Conclusion

Clustering in multicenter data should be accounted for when developing risk models. At the same time, the level at which risk models are used, determines how the performance should be assessed. It is important to understand that the choice of the model (standard or mixed effects logistic regression), the predictions used (marginal, assuming an average random intercept or conditional), and the level of performance evaluation (population or center level) will have an impact on the estimated predictive performance. We recommend the use of conditional predictions, when available, given their good performance at both the population and the center level.

8 General discussion

In this dissertation, I studied the consequences of wrongly assuming independence of observations in multicenter prediction research. In addition, I sought for alternative techniques that acknowledge clustering throughout the entire process of setting up a study to develop a risk model, analyzing the data, and externally validating the resulting model. A review of methodological research presented in chapter 2 showed that there is some consensus on good methodological practices for the development and validation of risk models, yet not on how to deal with multicenter data. In fact, the dependence of observations in multicenter prediction research is mostly ignored.

In chapter 3, I argued that mixed effects models may be most promising to yield good risk models. By the inclusion of a random center intercept in addition to several predictors, mixed effects models enable modeling differences in prevalence across centers. They do not assume independent observations and hence yield correct inferences. Moreover, unlike methods such as GEE and conditional logistic regression, they provide conditional predictions that are tailored to a specific center, for centers that were included in the development dataset. They can provide unbiased regression coefficients with as few as five centers and, in contrast to fixed effects models that include dummy variables to account for center differences, they can deal with small center sizes. An example of a mixed effects model is the Simple Rules risks scoring system, which includes random center intercepts and includes center type (oncology center versus other) as a predictor. This model discriminated well between benign and malignant ovarian tumors and the estimated risks of ovarian malignancy were well calibrated in oncology centers and other centers.

The simulation study in Chapter 7 indeed made clear that conditional predictions from a mixed effects model were well calibrated. This held true when calibration was assessed within centers and when it was assessed in the entire population of patients. The calibration was also more consistent across centers. The predictions from the standard logistic regression model, on the contrary, were not calibrated at the center level (calibration slopes typically >1 and calibration intercepts ≠0). In addition, the conditional predictions from the mixed effects model provided better discrimination at the population level. Predictions from the random intercept model assuming an average random center intercept were calibrated at the center level and discriminated between events and non-events equally well as the standard logistic regression model. However, the predictions were not calibrated at the population level (calibration slopes <1 and calibration intercepts ≠0).

The simulation study in Chapter 4 revealed that the severity of clustering did not increase the bias in estimated regression coefficients of mixed effects risk models. Instead, the driving force behind bias in the estimated regression coefficients and the predictive performance of a model at external validation was the sample size, and more specifically, the number of EPV in the multicenter sample (i.e., the ratio of the number of parameters to be estimated and the number of patients experiencing the outcome to be predicted). Samples with EPV 5 yielded overfitted risk models, which was confirmed in chapter 7 for all types of predictions (marginal, conditional and assuming an average center intercept). If backward variable selection was performed, an EPV of 50 was required to obtain unbiased estimates. In addition, the simulation studies showed that the random intercept variance was often underestimated. The bias decreased as the number of EPV increased (for the same number of centers) and as the number of centers increased (for the same number of EPV). Chapter 7 showed that the calibration of a mixed effects model depends on the estimated random effect variance, with larger calibration slopes as the variance was underestimated and vice versa.

Chapter 5 has demonstrated that measurements of predictors in multicenter studies may vary considerably between centers. This was the case for several predictors of ovarian malignancy, among others for the measured amount of fluid in the pouch of Douglas, presumably due to the position of the patient, and for the registered color score of intratumoral blood flow, a subjective measurement. The residual intraclass correlation (RICC) was proposed as a measure of the degree of clustering that can be applied in multicenter studies where each clinician assesses his own patients (in contrast to interrater agreement studies). The RICC takes the health status of the patients into account when it quantifies systematic differences in measurements across

centers. Predictors with heterogeneous measurements across centers may be excluded from a risk model to enhance the generalizability. In addition, statistical backward variable selection may be used to exclude predictors from the model. Chapter 5 showed that larger standard errors were obtained with the mixed effects model than with the standard logistic regression model. Consequently, fewer variables were retained during backward variable selection at the same level of significance. The predictive performance of the mixed effects and standard model was similar, despite the difference in the number of retained predictors.

Finally, chapter 6 showed that the predictive performance of risk and classification models for the pre-operative diagnosis of ovarian tumors may differ between centers at external validation. The quality of the discrimination between benign and malignant ovarian tumors was heterogeneous across centers, especially for RMI. Calibration-in-the-large varied as well, with RMI, LR2 and the Simple Rules overestimating the probability of malignancy in some centers and underestimating the risk in others. The results suggested that differences in the performance of models were related to the nature of the center (tertiary oncology referral center or other center type) and, to some extent, to the background training of the person applying the model. Random effects meta-analysis models proved to be very versatile tools to summarize the predictive performance across centers and to provide insight in the heterogeneity of performance.

8.1 Implications and recommendations

Based on the findings in this dissertation, I recommend adhering to existing methodological guidelines for the development and validation of risk models and using a modeling technique that reflects the structure of the data. In particular, mixed effects models should be used, as their conditional predictions are superior to the predictions of a standard logistic regression model, especially when the degree of clustering is strong. The need for center-specific models may in some circumstances be alleviated if we manage to include patient or center characteristics that explain the differences between centers in the risk model. When data from very few centers (less than five) is available, it would be preferable to use a fixed effects regression model, containing dummy variables for centers.[141]

In any case, large datasets are required to avoid overfitting. A reasonable guideline may be to collect a dataset with 10 EPV for predefined risk models, 20 EPV to alleviate the need for shrinkage techniques during estimation, and

50 EPV or more when data-driven predictor selection is used. The random intercept variance should be accounted for as an extra parameter in the EPV calculation. A reliably estimated random intercept variance improves calibration, but research has shown that a large number of clusters is needed for this.[146, 147, 215] One suggested guideline is to collect data from at least fifty clusters,[147] which may be hard to obtain in practice. Fortunately, the estimated effects of predictors in mixed effects models are generally unbiased if the EPV criteria are satisfied. Moreover, a sufficiently large number of EPV contributes to a good estimation of the random intercept variance.[53] The predictive performance of the model may therefore not be jeopardized in datasets with low numbers of centers provided that the EPV criteria are satisfied. Besides the number of centers, the choice of centers is crucial for the generalizability of the model: the chosen centers should reflect the different settings the model is intended for.

It is advisable to screen for systematic differences in measurements between centers in a multicenter study. These may arise for various reasons, among others differences in populations, differences in measurement equipment or procedures, or subjectivity of variables, and may threaten the generalizability of the risk model. The RICC is suited to monitor the differences in measurements as the study is ongoing, which allows solutions such as measurement protocol adjustments. The RICC reveals systematic differences in measurements (systematically higher for some centers and systematically lower for others) and indicates differences in underlying populations. An interrater agreement study, in contrast, also reveals non-systematic differences in measurements but does not unveil differences in centers' study populations.

In general, the best practice for predictor selection is to define in advance the predictors that are most promising based on clinical expertise. In addition, large systematic differences in measurements indicated by the RICC may be a reason to exclude a predictor from a risk model. Statistical variable selection may sometimes be unavoidable, for example when there are many clinically relevant candidate predictors and little is known about their association with the outcome. In this case, it is important to use methods that acknowledge clustering in every step of the statistical selection procedure, since standard errors are unreliable when the assumption of independent observations is violated.

Whether an investigator should evaluate a risk model at the population level or the center level depends on the situation. If the goal is to implement a risk model nationwide, for example, based on national guidelines, the discrimination at the national level can be considered. When risk models are

used to support decision-making within centers, they should perform well at the center level. Besides investigating the average predictive performance, it is also useful to investigate potential heterogeneity in model performance across centers. This can be done, for example, by adopting a meta-analytic framework, studying the variance of random effects and prediction intervals of performance statistics, or by investigating empirical Bayes estimates of center-specific calibration intercepts and slopes.[152, 218, 234] If centers are large, the center-specific performance can be studied. Differences in model performance can be caused by many things, including case-mix differences in the center populations and differences in definitions of variables.[98] It must therefore be kept in mind that obtaining good discrimination and calibration in every single center may be difficult. The local performance in specific centers may be improved by using conditional predictions from a mixed effects logistic regression model or by applying model updating techniques.[6, 100, 101, 142, 152, 245]

To make predictions for new individuals I recommend conditional predictions, because of their good performance at the center level and the population level. Center-specific random intercepts are required to obtain conditional predictions from a mixed effects model but are not available for new centers. In the absence of data from the new center, numerical integration of the predictions over the estimated random effects distribution may be used. However, these predicted probabilities have a population-averaged interpretation and may not be well calibrated at the center level. Another option is to substitute the center-specific random effect by zero. These predictions are easy to obtain and will be well calibrated in centers with an average effect. Alternative options are to estimate the center-specific intercept from the outcome prevalence of the new center or to use the intercept of a similar center from the model development set.[142, 143]

8.2 Future research

Chapters 4 through 7 have focused on mixed effects models. Chapter 3 discussed many other models and techniques that account for data clustering, such as fixed effects logistic regression, GEE and conditional logistic regression. In the future, I plan a direct comparison of the applicability and predictive performance of various techniques in several scenarios, with varying characteristics of the model development set, among which the number of centers, the number of patients per center, the distribution of center

effects, the presence of noise variables and the presence of center-predictor interactions ("random slopes").

An important research question is whether the effects of predictors may differ between centers in practice, how this should be handled when building a risk model, and whether accounting for such predictor-center interactions would improve the predictive performance. Two models that account for center-treatment interactions are random slope models and fixed effect models with interactions between center dummies and predictors. A sensible strategy to select center-treatment interactions is needed. This should rely as much as possible on expert knowledge. In addition, it should describe where in the model building process center-treatment interactions have to be investigated: before or after the selection of main effects, interactions between predictors, and transformations of linear variables.

Chapter 7 has illustrated that conditional predictions from a mixed effects logistic regression model are well calibrated at external validation and are able to discriminate reliably between events and non-events. However, the empirical Bayes estimates of center-specific intercepts required to make conditional predictions are only available for centers that were included in the model development dataset. Several updating techniques have been proposed to face the challenges specific to mixed effects models, among which frequentist, Bayesian and Empirical Bayes methods.[142, 143, 154] The applicability of each of these techniques depends on the available information, ranging from information available in the literature, to new patient data from new centers. I am currently planning a study to investigate the applicability and predictive performance of various updating techniques for mixed effects models in several scenarios, with varying characteristics of the model development set and the new centers in which the model will be applied.

Another interesting topic for future research is the composition of multicenter datasets. Researchers often strive for balanced datasets with equal contributions from all centers, but these are usually impractical or impossible to obtain because the inflow of patients differs strongly between hospitals.[214] Since the increased generalizability of results is one of the major reasons to perform multicenter research,[135] one has to ask whether imbalanced samples jeopardize the predictive performance of risk models in new data. Several scenarios for data collection need to be investigated. For example, a scenario in which the majority of data is collected in a small number of large, specialized centers, while the rest of the data is collected in various small centers. This situation occurs in practice when principal investigators work in

large university hospitals and invite smaller, regional centers to collaborate on a research project. Since the patient populations may differ, it is interesting to investigate how large the contributions of the regional centers ought to be in order to yield a generalizable risk model.

Finally, the evaluation of predictive performance in multicenter validation studies also requires some further research. For example, several methods to evaluate calibration need to be compared empirically. Center-specific flexible calibration curves are able to demonstrate moderate calibration but a large number of patients per center may be required to construct reliable curves for the individual centers.[64] Alternatively, a one-step meta-analysis can be performed using mixed effects logistic calibration. This only demonstrates weak calibration but has the advantage that empirical Bayes estimates of individual calibration intercepts and slopes are available to check calibration in individual centers. Since empirical Bayes estimates are shrunken towards the average if there is little center-specific information, this may be a more reliable way to assess center-specific calibration. In addition, a two-step meta-analysis of calibration intercepts or slopes may be performed, if individual patient data from the different centers is not available. The evaluation of clinical utility for decision-making in multicenter studies requires further research as well. NB estimates are correlated with their variance and consequently a random effects meta-analysis of NB may not be appropriate. Future research may look into variance-stabilizing functions of NB. A promising alternative is to derive a summary NB and probability statements concerning the expected benefit in new centers in a Bayesian meta-analysis of sensitivity, specificity and prevalence.

Although many interesting methodological research questions remain, the evidence to date suggests that applied prediction research would benefit from adopting mixed and random effects approaches for the development and validation of risk models, to fully exploit the information available in multicenter data.

Appendices

A1 Multiple imputation in the IOTA dataset

The serum CA125 tumor marker was not a mandatory variable, and measurements were missing in 31% of the patients included in phases I to III.[8, 183] This was dealt with using multiple imputation. Missing values are highly likely to have occurred for two reasons. First, following different local management practices some centers were more committed than others to measure CA125. The percentage of missing values per center varied between 0% and 94% with a median of 19%. Seventeen of 24 centers had at most 30% missing values, and four centers had clearly more than 50% missing values: Bologna (n_j=348, 69% missing values), Genk (n_j =428, 72% missing values), Barcelona (n_j=31, 84% missing values), Maurepas (n_j =64, 94% missing values).[8, 183] Second, investigators sometimes refrained from measuring CA125 based on the overall clinical picture and the appearance of the tumor on ultrasound. Hence, the value was more often missing in women with a benign tumor (36%) than in women with a borderline (18%), stage I invasive (20%), stage II-IV invasive (17%), or secondary metastatic (25%) tumor. [8, 183]

Predictive mean matching regression[179] was used for multiple imputation. This way, likely CA125 values were estimated for patients with a missing value, using variables that were related to either the level of CA125 itself, or to the unavailability of CA125 (i.e., a binary indicator indicating for each woman whether CA125 was missing or not). The center was included in the imputation model as a fixed effect. The log(log(CA125-1)) transformation was used to obtain a more symmetric distribution. The imputation process was repeated 100 times to generate multiple imputations of the missing values, resulting in 100 completed data sets. The results were combined using Rubin's rules.[8, 183]

The same approach was used in IOTA 4b data, where the CA125 level was missing for 19 patients (35%).[184] In all these cases, the RMI value was zero irrespective of the CA125 level, as the ultrasound score was zero. Hence, the missing data did not influence the results presented in this thesis.

A2 Technical overview of the design of the EPV simulation study

The generation of the source populations was done as follows:

1. The number of clusters (J) was fixed at 200. The number of observations per cluster was drawn from a Poisson distribution with a separate, randomly generated lambda for each cluster. This yielded cluster sizes ranging from approximately 200 to 1000. Hence, each population consisted of approximately 10,000 observations in total.
2. Each cluster was assigned a random cluster intercept a_j. They were generated from a normal distribution of which the variance was determined by the predefined ICC (0%, 5% or 20%) as follows: $\tau^2 = \frac{\frac{\pi^2}{3}\times ICC}{1-ICC}$.
3. Each of the Xs was generated from a normal (X_1~N(0,1), X_2~N(0,0.6), X_3~N(0,0.4), X_4~N(0,0.2)) or binomial (X_5 to X_8, with prevalence 0.2, 0.3, 0.3, and 0.4 respectively) distribution.
4. In one source populations (defined as set 6 in Table 12), a_j was correlated with X_1 and X_5. The correlation between X_1 and a_j was induced as follows: $X_{1\ correlated\ ij}=\sqrt{0.25}\times a_j+\sqrt{0.75}\times X_{1\ ij}$, after which $X_{1\ correlated}$ was rescaled to have the same mean and standard deviation as X_1. This yields a correlation of approximately 0.5 between $X_{1correlated\ ij}$ and a_j. To create $X_{5\ correlated\ ij}$, the success probability determining its distribution was related to a_j: $X_{5\ correlated\ ij}$~Bernoulli $\left(0.20+rescaled_a_j\right)$, where $rescaled_a_j=0.1\times\left(a_j-\bar{a}_j\right)/\max\left[abs\left(a_j-\bar{a}_j\right)\right]$, such that $-0.1\leq rescaled_a_j\leq 0.1$ and $rescaled_a_j \propto a_j$. The prevalence of $X_{5\ correlated\ ij}$ hence increased with the random cluster intercept: it is 12% at the cluster with the lowest a_j and 29% at the cluster with the highest a_j.
5. For each of the observations, the logit is computed from the generated Xs and a_j, using equation (10). Each β equals 0.8 and the intercept equals -2.1, to obtain a prevalence of Y_{ij} of 0.3.
6. The predicted probabilities of an event occurring are computed by exploiting the inverse logit transformation (see equation (3)).
7. Finally, to incorporate an element of randomness in the outcomes, the Y_{ij} are obtained by comparing the predicted probability of the

event to a randomly drawn value from a uniform distribution as follows:

$$U_{ij} \sim unif(0,1)$$
$$Y_{ij} = \begin{cases} 1 & \text{if } U_{ij} \leq p_{ij} \\ 0 & \text{if } U_{ij} > p_{ij} \end{cases}.$$

The sampling from the source population and the model building occurred as follows:

1. The number of clusters (J) to be drawn was prespecified according to the condition (see Table 12).
2. The number of events to be sampled was computed from the required EPV and the number of parameters (K+1) to be estimated (17 in sets 2 and 3 and 9 in all other conditions). The total sample size was computed from the number of events and the prevalence in the source population (0.3), or the predefined sample event rates p in set 5 (0.05, 0.1, 0.2 or 0.5).
3. Clusters were randomly drawn without replacement from the source population.
4. A list was generated with all observations from the sampled clusters. From this list, events and non-events were sampled separately, hence stratified for outcome Y_{ij}, and without replacement, to ensure the correct number of events was sampled and the prevalence of the outcome in the sample was as required.
5. In the obtained sample, a random intercept model was fitted. The model either contained the eight predefined predictors (X_1 through X_8), the eight predefined predictors and eight noise variables (X_1 through X_{16}, in set 2), or was built using backward selection (in set 3; α for deletion 0.10). A random intercept model was fitted in each subsequent step of the backward variable selection algorithm, to acknowledge clustering during model building.
6. The fitted model was evaluated in the source population, to obtain the c-statistic, the within-cluster c-statistic, the calibration slope and the within-cluster calibration slope. Predicted probabilities were obtained by omitting the random intercepts from the prediction equation (see equation (11)).[151]
7. Steps 1 through 6 were repeated 500 times.

A3 Examples of R code for the EPV simulation study

A3.1 Generation of source populations

```
rm(list=ls(all=TRUE))
library (rms)
library (lme4)

#Step 1. Fix the number of clusters (nb) and draw the number of observations per cluster (n2) from a Poisson distribution.
nb <- 200
mean <- 6.22
sd <- 0.3
L1<-rnorm(nb, mean, sd)
L<-round(exp(L1))
n2 <- rpois(nb, L)
an <-seq(1:nb)
anest <-rep(an, n2)
n <- length(anest) #number of observations in the source population
pat_nr <- seq(1:n)

#Step 2. Generate random cluster intercepts (randeff).
randeff <- rnorm(nb, 0, 0.416)  # this is to obtain an ICC of 5%
randeffect <-rep(randeff, n2)

## Step 3. Generate predictors X1 to X8
x1      <- rnorm(n, 0, 1)
x2      <- rnorm(n, 0, 0.6)
x3      <- rnorm(n, 0, 0.4)
x4      <- rnorm(n, 0, 0.2)
x5      <- rbinom(n, 1, 0.20)
x6      <- rbinom(n, 1, 0.30)
x7      <- rbinom(n, 1, 0.30)
x8      <- rbinom(n, 1, 0.40)

#Step 4: Introduce correlation between random intercept and predictors if required.
# Step 5: Set beta coefficients and obtain the logit.
b1      <- 0.8
```

```
b2      <- 0.8
b3      <- 0.8
b4      <- 0.8
b5      <- 0.8
b6      <- 0.8
b7      <- 0.8
b8      <- 0.8
logit   <-                                                          -
2.1+b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+randeffect

#Step 6: Calculate the predicted probabilities of experiencing an event
p       <- plogis(logit)

#Step 7: Obtain dichotomous outcome Y
y       <- ifelse(runif(n)<=p, 1, 0)
data <- as.data.frame(cbind(pat_nr, x1, x2, x3, x4, x5, x6, x7, x8, randeffect,
anest, y), dimnames = list(c(1:n), Cs(pat_nr, x1, x2, x3, x4, x5, x6, x7, x8,
randeffect, anest, y)))
```

A3.2 Sampling from the source population and model building within samples

This code corresponds to simulation condition 1.11 (Table 12). The code can be adjusted for other conditions with minor changes.

```
rm(list=ls(all=TRUE))
library (Hmisc)
library (rms)
library (lme4)
library(arm)
setwd("C:\\path for source population data")
load('name of source population') #load the source population
set.seed(choose a seed)
Nsample <- 500  #specify the number of samples to be drawn

#Step 1. Specify the number of clusters
Nsamplev2 <-30

#Step 2. Specify the number of events to be drawn
NEVENTS <- 450
NNONEVENTS <- round((NEVENTS/mean(data$y))-NEVENTS)
```

```
#Matrices to store results
Nanest_sample <- matrix(NA, nrow=Nsample, ncol=1)
anest_sample <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
Nevents_sample <- matrix(NA, nrow=Nsample, ncol=1)
Npats_sample <- matrix(NA, nrow=Nsample, ncol=1)
Pevents_sample <- matrix(NA, nrow=Nsample, ncol=1)
Npats_anest_sample <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)

mat_sd_ranef0 <- matrix(NA, nrow=Nsample, ncol=1)
mat_sd_ranef1 <- matrix(NA, nrow=Nsample, ncol=1)
mat_regcof <- matrix(NA, nrow=Nsample, ncol=9)
mat_SEregcof <- matrix(NA, nrow=Nsample, ncol=9)
mat_ranef <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
mat_SEranef <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
Niterm1 <- matrix(NA, nrow=Nsample, ncol=1)
Equaldeviance <- matrix(NA, nrow=Nsample, ncol=1)
mat_dims <- matrix(NA, nrow=Nsample, ncol=18)

realvalresultB <- matrix(NA,nrow=Nsample, ncol=5)
dimnames(realvalresultB) <- list(c(1:Nsample),
Cs(C,calib_slope,Cw,ML_calib_slope,var_ML_calib_slope))

loglik0 <- matrix(NA,nrow=Nsample,ncol=1)
loglik1 <- matrix(NA,nrow=Nsample,ncol=1)

#Start sampling from the domain
tmpA <- aggregate(data$pat_nr, list(anest = data$anest), FUN=length)
pt_anest <- as.vector(tmpA[,2])
names(pt_anest) <- tmpA[,1]

for (r in 1:Nsample){

#Step 3. Sample clusters (samp_anest) from the source population
samp_anest <- sample(x=names(pt_anest), size=Nsamplev2, replace = F)

#Get the data from sampled clusters (dsample_lev2)
loc.pt_anest <- pt_anest[samp_anest]
indices.cutoffs <- c(0,cumsum(loc.pt_anest))
ptnrs_samp2 <- matrix(NA,nrow=sum(loc.pt_anest),ncol=1)
for (i in 1:Nsamplev2){
```

```
ptnrs_samp2[(indices.cutoffs[i]+1):indices.cutoffs[i+1],] <-
data[data$anest==names(loc.pt_anest)[i],1]
}
dsample_lev2 <- data[ptnrs_samp2,]

#Step 4. Stratified sampling of events (sample_lev1_y1) and non-events
(sample_lev1_y0).
sdatay1 <- dsample_lev2[dsample_lev2$y==1,]
pt_nr_y1 <- as.vector(sdatay1$pat_nr)
sample_lev1_y1 <- sample(pt_nr_y1, NEVENTS , replace = FALSE)
sdatay0 <- dsample_lev2[dsample_lev2$y==0,]
pt_nr_y0 <- as.vector(sdatay0$pat_nr)
sample_lev1_y0 <- sample(pt_nr_y0, NNONEVENTS, replace = FALSE)
sample_lev1 <- c(sample_lev1_y0, sample_lev1_y1)
dsample <- data[sample_lev1,]

#Step 5. Fit a random intercept model in the sample: full model (m1) and null
model (m0)
iter_m1<-capture.output(m1<-
lmer(y~x1+x2+x3+x4+x5+x6+x7+x8+(1|anest), family = "binomial",
data=dsample, verbose=T ))
m0<-lmer(y~(1|anest), family = "binomial", data=dsample)
Last2it<-iter_m1[c(length(iter_m1)-1):c(length(iter_m1))]
Last2itsplit<-strsplit(Last2it,":")
Equal<- Last2itsplit [[1]][2]== Last2itsplit [[2]][2]#check convergence

# Set Npats_anest, anest and ranef to missing when we sampled <1
observation per cluster
anest_s <- length(unique(dsample$anest))
N_missing_anest <- Nsamplev2-Nanest_s
add_anest <- rep(x=-99, times=N_missing_anest)
if (Nanest_s < Nsamplev2){
anest_s <- c(unique(dsample$anest), add_anest)
Npats_anest_s <- c((aggregate(dsample$pat_nr, list(anest = dsample$anest),
FUN=length)[,2]),add_anest)
clustereff_s <- c(attr(m1, "ranef"),add_anest)
clustereff_s_SE <- c(as.numeric(se.ranef(m1)$anest), add_anest)
}
if (Nanest_s == Nsamplev2) {
Npats_anest_s <- aggregate(dsample$pat_nr, list(anest = dsample$anest),
FUN=length)[,2]
```

```
anest_s <- unique(dsample$anest)
clustereff_s <- attr(m1, "ranef")
clustereff_s_SE <- as.numeric(se.ranef(m1)$anest)
}

#save sample summaries
Nanest_sample[r,] <- Nanest_s
anest_sample[r,] <- anest_s
Npats_anest_sample[r,] <- Npats_anest_s
Nevents_sample[r,] <- sum(dsample$y)
Npats_sample[r,] <- length(dsample[,1])
Pevents_sample[r,] <- mean(dsample$y)

#save model parameters
mat_sd_ranef0[r,]<- sigma.hat(m0)$sigma$anest
mat_sd_ranef1[r,]<- sigma.hat(m1)$sigma$anest
mat_regcof[r,]<- as.numeric(fixef(m1))
mat_SEregcof[r,]<- se.coef(m1)$fixef
mat_ranef[r,] <- clustereff_s
mat_SEranef[r,] <- clustereff_s_SE
mat_dims[r,] <- as.numeric(attr(m1,"dims"))
Niterm1[r,] <- length(iter_m1)-1
Equaldeviance[r,] <- Equal

#save log likelihood of full and null model
loglik0[r,]<-logLik(m0)
loglik1[r,]<-logLik(m1)

#Step 6. Evaluate model performance in the source population
data$lpB <- cbind(1, as.matrix(data[,2:9])) %*% fixef(m1)
perfm_m1B<- lrm(data$y~data$lpB)
realvalresultB[r,1] <- perfm_m1B$stats[6] # C
realvalresultB[r,2] <- perfm_m1B$coefficients[2] #calibration slope
C.w.index <- rep(NA, length(unique(data$anest))) #within-cluster C.
Requires function concordance.prob.logistic from Van Oirbeek et al. 2012.[217]
n.comp <- rep(NA, length(unique(data$anest)))
for(p in 1:length(unique(data$anest))){
data_p <- data[data$anest==p,]
prediction <- plogis(data_p$lpB)
outcome <- data_p$y
result <- concordance.prob.logistic(outcome, prediction)
```

```
C.w.index[p] <- result$C
n.comp[p] <- result$n.comp
}
realvalresultB[r,3] <- sum(C.w.index*n.comp/sum(n.comp, na.rm = TRUE), na.rm = TRUE) #within-cluster C-index
calibslope_mlB <- glmer(y~lpB+(lpB|anest), family = "binomial", data=data)
realvalresultB[r,4] <- as.numeric(fixef(calibslope_mlB)[2]) #within-cluster calibration slope
realvalresultB[r,5] <- (sigma.hat(calibslope_mlB)$sigma$anest[2])^2 #random variance of the calibration slope

#save to check convergence
write.table(cbind(r, Niterm1[r,], iter_m1), file="C:\\path\\iter", col.names=F, row.names=F, append=T)
write.table(x=Equaldeviance, file=" C:\\path\\Equaldeviance", append=F, na="NA", col.names=F)
}

#Save all results
```

A4 Additional results from the EPV simulation study

A4.1 Bias in the estimated regression coefficients

Table A 1. The relative bias (%) in the estimated regression coefficients.

EPV	ICC	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
5	0	7.4 (-13.0 to 33.7)	11.7 (-48 to 63.4)	12.4 (-47.7 to 64.4)	15.0 (-71.2 to 101.5)	6.9 (-39.8 to 59.6)	8.6 (-35.0 to 48.5)	9.0 (-33.5 to 49.1)	9.1 (-28.8 to 51.8)
	5	11.6 (-13.8 to 33.8)	11.5 (-39.2 to 68.2)	14.6 (-37.5 to 72.8)	6.5 (-78.6 to 117.7)	4.1 (-47.6 to 51.0)	5.9 (-33.1 to 52.6)	7.0 (-32.7 to 46.0)	10.6 (-31.8 to 46.3)
	20	5.8 (-15.0 to 32.8)	6.1 (-45.1 to 60.9)	3.2 (-46.6 to 56.5)	-10.2 (-115.1 to 108.9)	5.4 (-48.5 to 54.6)	6.3 (-37.5 to 55.4)	9.5 (-35.9 to 56.6)	5.4 (-31.2 to 49.1)
10	0	5.3 (-7.6 to 19.5)	2.8 (-30.9 to 37.5)	3.5 (-30.4 to 38.4)	-3.1 (-57.3 to 55.5)	4.0 (-28.3 to 32.2)	0.7 (-21.2 to 33.0)	3.8 (-18.5 to 31.1)	4.1 (-22.5 to 29.2)
	5	4.9 (-10.7 to 19.0)	2.0 (-28.3 to 32.9)	4.7 (-26.4 to 36.5)	5.0 (-56.4 to 61.6)	1.0 (-29.8 to 32.5)	2.6 (-21.7 to 28.7)	2.7 (-23.9 to 32.6)	2.5 (-23.4 to 28.9)
	20	4.0 (-9.8 to 19.5)	5.0 (-29.9 to 40.5)	2.1 (-31.9 to 36.7)	8.9 (-60.5 to 74.2)	4.7 (-32.4 to 37.8)	-0.2 (-26.3 to 29.4)	-3.0 (-30.8 to 27.9)	5.1 (-23.3 to 32.3)
20	0	3.5 (-34.5 tot 13.5)	2.2 (-17.4 to 25.3)	2.9 (-16.9 to 26.1)	2.9 (-36.0 to 42.9)	3.1 (-21.9 to 24.3)	3.2 (-14.6 to 20.3)	-0.6 (-19.7 to 19.6)	0.2 (-19.0 to 16.5)
	5	2.7 (-7.7 to 11.7)	2.5 (-23.0 to 21.5)	5.3 (-20.9 to 24.8)	-0.06 (-41.7 to 39.7)	2.4 (-19.8 to 24.3)	-2.6 (-20.6 to 16.8)	2.2 (-16.1 to 19.3)	1.5 (-17.2 to 15.5)
	20	0.7 (-8.7 to 9.9)	5.5 (-15.4 to 29.1)	2.6 (-17.7 to 25.5)	-4.6 (-55.8 to 42.4)	-3.0 (-23.6 to 20.7)	2.7 (-19.2 to 21.3)	4.0 (-17.3 to 23.0)	-1.3 (-18.7 to 20.3)
50	0	1.1 (-4.4 to 7.0)	1.1 (-13.5 to 14.7)	1.7 (-13.0 to 15.5)	-0.4 (-27.1 to 25.1)	0.8 (-11.4 to 11.2)	0.4 (-11.3 to 12.4)	-0.4 (-11.1 to 11.7)	0.7 (-10.3 to 10.9)
	5	1.2 (-5.3 to 7.0)	-1.9 (-14.1 to 11.9)	0.7 (-11.7 to 14.9)	0.9 (-25.4 to 30.9)	1.3 (-10.6 to 14.3)	-0.8 (-10.5 to 11.4)	0.4 (-11.2 to 11.8)	1.4 (-9.9 to 12.3)
	20	0.9 (-5.6 to 7.5)	4.2 (-9.3 to 17.7)	1.3 (-11.8 to 14.5)	-2.1 (-26.1 to 24.7)	-0.6 (-13.6 to 15.0)	0.6 (-12.8 to 11.7)	0.8 (-11.0 to 13.1)	0.0 (-11.2 to 13.2)

Bias (interquartile range).

A4.2 Standard (population-level or "overall") c-statistics and calibration slopes

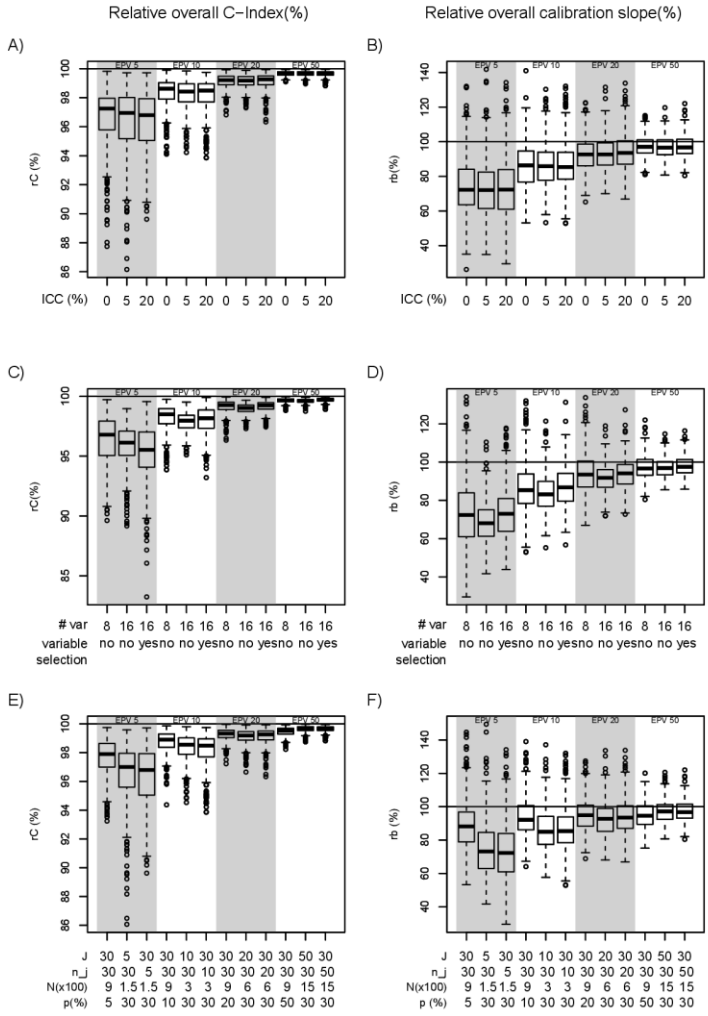


Figure A 1. The relative standard c-statistic and the relative standard calibration slope in relation to the number of events per variable (EPV), the amount of clustering (ICC), backward variable selection and sample size characteristics. rC: relative standard C-index; rb: relative standard calibration slope; #var: number of candidate predictors; J: the number of clusters, n_j: the average number of observations per cluster; N (x100): the total number of observations, to be multiplied by 100; p: prevalence. A) the effect of EPV and ICC on the relative discrimination; B) the effect of EPV and ICC on the relative calibration C) the effect of EPV and backward variable selection on the relative discrimination (ICC=20%); D) the effect of EPV and backward variable selection on the relative calibration (ICC=20%); E) the effect of EPV and sample characteristics on the relative discrimination (ICC=20%); F) the effect of EPV and sample characteristics on the relative calibration (ICC=20%). The box indicates the interquartile range (IQR), the bold horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

A4.3 Bias in the estimated random intercept variance

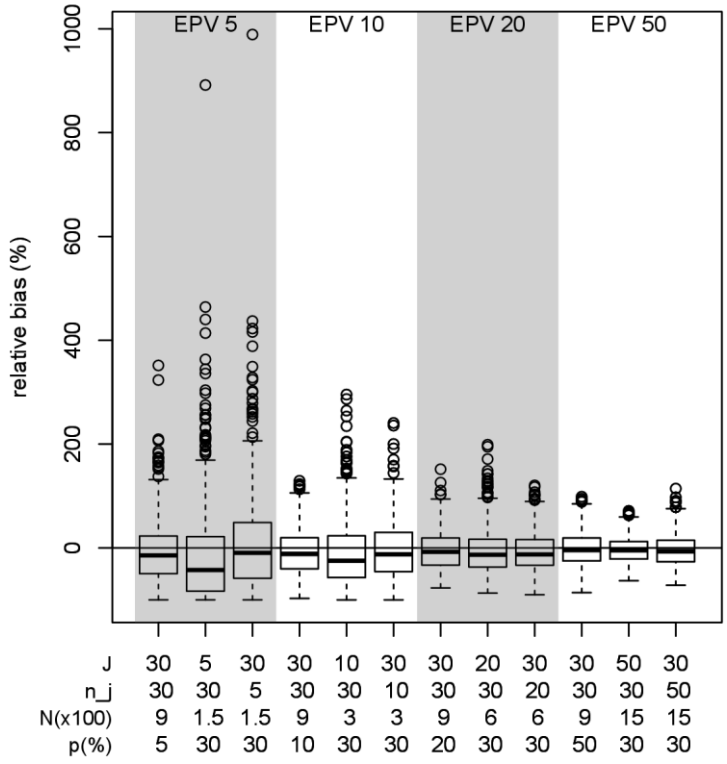


Figure A 2. The relative bias in the estimated random intercept variance (ICC=20%) in relation to the number of events per variable (EPV) and sample characteristics. ICC: intraclass correlation, J: the number of clusters, n_j: the average number of observations per cluster; N (x100): the total number of observations, to be multiplied by 100; p: prevalence. The box indicates the interquartile range (IQR), the bold horizontal line within the box indicates the median. Whiskers extend to the lowest and highest data points still within 1.5 IQR of the box. Outliers beyond these points are represented by dots.

A5 SAS macro to estimate the residual intraclass correlation

The RICC gives the proportion of variance of a variable due to differences between physicians, adjusted for several fixed effects that explain the differences between clusters.

It is a useful tool to screen for problematic variables in a multicenter study. It allows subjects to be nested in clusters and does not require that two or more physicians judge the same subject.

As such it can be used to detect problems with systematic differences in average measurements by physicians, due to e.g. cultural or regional differences, subjectivity of variables, miscalibration of measurement tools, deviations from the measurement protocol or unequal distributions of variables across clusters due to local referral patterns.

The outcome of this macro includes:

1) the residual intraclass correlation (RICC): gives the proportion of residual variance in the variable, after controlling for fixed effects. This will be 1 if within clusters, all subjects with a given level for the fixed predictors get the same score.

2) the proportion of variance at the cluster level (VPC): Gives the proportion of the total variance in the variable that is situated at the cluster level. This will be 1 only if within clusters, all subjects get the same score.

Notes:

- To use this script, indicate a dataset 'data', a variable 'score' (the variable for which you want to study clustering), variables 'predictor' (fixed effects to account for clustering), and a cluster indicator 'cluster';
- REML is used;
- Indicate whether the score variable's type is continuous (cont), ordinal (ord) or dichotomous (dich);
- Within-effects of the predictor need to be separated from between-effects. ‘Predictor’ is the list of (dummy) variables used to control for case-mix (usually health status). After 'between', indicate the mean/proportion of each of these variables per cluster;
- With macro_pred you can control for macro-variables, i.e., variables at the cluster level that can explain part of the residual ICC or VPC.

```
%MACRO variance_partition(data=, score=, cluster=,
predictor=, between=, macropred=, type=);

options spool serror symbolgen;

%let predictor= %UPCASE(&predictor);
%let between= %UPCASE(&between);
%let macropred= %UPCASE(&macropred);

*1. CONTINUOUS VARIABLE;

%IF &type=cont %THEN %DO;

      proc mixed data=&data;
            class &cluster ;
            model &score= &predictor &between &macropred/
s; /* we first include both between and within effects,
so we can separate them to prevent confounding*/
            random &cluster;
            ods output solutionf=fixed;
      run;

      data fixed;
            set fixed;
            effect=upcase(effect);
      run;

      %let n=%sysfunc(countw(&predictor)); /*now we will
isolate the within effect so we can continue to work with
only the within effect*/

      %do var=1 %to &n;
            %let withinvar= %qscan (&predictor, &var);

            data _null_;
                  set fixed;
                  if effect="&withinvar" then call symput
('estimate', estimate);
            run;

            data &data;
                  set &data;
                  coeff=&estimate;
                  part_var_&var = &withinvar * coeff;
```

```
            run;
      %end;

      data &data;
            set &data;
            within=sum(of part_var_1-part_var_&n);
            score_within=&score-within;/*consider     the
within effect as an offset*/
            drop coeff part_var_1-part_var_&n;
      run;

      proc mixed data=&data;/* run the model to determine
the residual variance at cluster and subject level*/
            class &cluster;
            model   score_within=&macropred;   /*we   have
included the within as an offset to take only this into
account,  and  thus  prevent  confounding  of  within  and
between cluster effects*/
            random intercept /sub=&cluster solution;  /*
the random intercept is computed, this captures also the
between-effect now*/
            ods       output       CovParms=cluster_resid
solutionr=RI; /* write variances to cluster_resid*/
      run;

      data cluster_resid;
            set cluster_resid;
            model='full'; /* indicate that this is the
residual variance by recognizing the full model, including
a fixed predictor, was fitted*/
      run;

      proc  mixed  data=&data;  /*  run  an  empty  model  to
determine  the  total  variance  at  cluster  and  subject
level*/
            class  &cluster ;
            model &score=;
            random intercept /sub=&cluster;
            ods  output  CovParms=total_cluster_resid;  /*
write variances to total_cluster_resid*/
      run;

      data total_cluster_resid;
            set total_cluster_resid;
            model='empty'; /* indicate that this is the
total variance by recognizing the empty model was fitted*/
```

```
      run;

      /* now we write the variance at the cluster and
subject to a macro variable. We do this for the empty
model and for the full model accounting for the
predictor*/

      data _null_;
            set total_cluster_resid;
            if CovParm="Intercept" then call symput
("var_tot_cluster", Estimate);
            if CovParm="Residual" then call symput
("var_tot_res", Estimate);
      run;

      data _null_;
            set cluster_resid;
            if CovParm="Intercept" then call symput
("var_res_cluster", Estimate);
            if CovParm="Residual" then call symput
("var_res_res", Estimate);
      run;

      /* we compute the variance coefficient to quantify
the total variability*/

      proc means data=&data;
            var &score;
            output out=VC mean=mean std=std;
      run;

      data vc;
            set vc;
            Vc=mean/std;
            if _type_=0 then call symput ("coeff_var",
vc);
      run;

      /* we compute the variance partitioning measures of
interest*/

      data outcome;
            variation_coefficient=&coeff_var;
            Residual_variance_cluster=&var_res_cluster;
            Residual_variance_subject=&var_res_res;
            Total_variance_cluster=&var_tot_cluster;
```

```
            Total_variance_subject=&var_tot_res;

      Total_variance=Total_variance_subject+Total_varian
ce_cluster;
            Variance_explained=Total_variance-
Residual_variance_cluster-Residual_variance_subject;

      Proportion_cluster=Residual_variance_cluster/Total
_variance;

      Proportion_explained=Variance_explained/Total_vari
ance;

      Proportion_subject=Residual_variance_subject/Total
_variance;

      Residual_ICC=Residual_variance_cluster/(residual_v
ariance_cluster+residual_variance_subject);
      run;

      title 'Total variance';
      proc print data=outcome label noobs;
            label  variation_coefficient="coefficient  of
variation"        total_variance="total        variance"
proportion_explained="explained"
proportion_subject="subject"
proportion_cluster="cluster";
            format                        proportion_cluster
proportion_explained proportion_subject percent10.1;
            var    variation_coefficient    total_variance
proportion_cluster                    proportion_explained
proportion_subject;
      run;

      title 'Residual variance';
      proc print data=outcome label noobs;
            label residual_icc="residual ICC";
            format residual_icc percent10.1;
            var Residual_ICC;
      run;

      data _null_;
      call symput ("var_tot_cluster", ".");
      call symput ("var_tot_res", ".");
      call symput ("var_res_cluster", ".");
      call symput ("var_res_res", ".");
```

```
      call symput ("coeff_var", ".");
      run;

%END;

%IF &type=dich %THEN %DO;


      proc glimmix data=&data method=laplace;  /*first we
compute the model including both within and between
effects*/
            class &cluster ;
            model     &score     =&predictor     &between
&macropred/s dist=binomial;
            random &cluster;
            ods output ParameterEstimates=fixed;
      run;

      data fixed;
            set fixed;
            effect=upcase(effect);
      run;

      %let  n=%sysfunc(countw(&predictor));  /*  now  we
separate the between from the wihtin effect*/

      %do var=1 %to &n;
            %let withinvar= %qscan (&predictor, &var);

            data _null_;
                  set fixed;
                  if effect="&withinvar" then call symput
('estimate', estimate);
            run;

            %put &estimate;

            data &data;
                  set &data;
                  coeff=&estimate;
                  part_var_&var = &withinvar * coeff;
            run;
      %end;

      data &data;
            set &data;
```

```
            within=sum(of part_var_1-part_var_&n);
            drop coeff part_var_1-part_var_&n;
      run;

      proc glimmix data=&data method=laplace;
            class &cluster ;
            model     &score=     &macropred/offset=within
dist=binomial  s;  /*  we  continue  with  only  the  within
effect, as an offset*/
            random  intercept  /sub=&cluster  solution;  /*
the random intercept captures also the between effect of
the predictor*/
            output out=lp  pred(noblup)=lp;/* we save the
linear  predictor  as  we  will  need  this  to  computed  the
explained variance*/
            ods       output       CovParms=cluster_resid
solutionr=RI; /* write variances to total_cluster_resid*/
      run;

      proc means data=lp var; /*we compute the variance
of the linear predictor. This is the variance explained
by the model. Note that we do not need to fit a full and
empty  model  here,  as  was  the  case  with  the  continuous
model*/
            where lp ne 0; /* sas puts the lp to 0 if the
cluster has a missing value. Since these are not used for
the computation of the variances in the multilevel model,
we  don't  use  them  for  the  variance  of  the  fixed  part
either*/
            var lp;
            output out=var_lp var=var_lp;
      run;

      data _null_;
            set var_lp;
            if _type_=0 then call symput ("var_explained",
var_lp);
      run;

      data _null_;
            set cluster_resid;
            if   CovParm="Intercept"   then   call   symput
("var_cluster", Estimate);
      run;
```

```
      proc means data=&data; /*we compute the prevalence
to quantify the total variance*/
            var &score;
            output out=prev mean=prev;
      run;

      data _null_;
            set prev;
            if _type_=0 then call symput ("prevalence",
prev);
      run;

      data outcome; /* we compute the variance
partitioning measures of interest*/
            prevalence=&prevalence;
            Residual_variance_cluster=&var_cluster;
            pi=constant("PI");
            Residual_variance_subject=pi*pi/3;
            Variance_explained=&var_explained;

      Total_variance=Residual_variance_cluster+Residual_
variance_subject+variance_explained;

      Proportion_cluster=Residual_variance_cluster/Total
_variance;

      Proportion_explained=Variance_explained/Total_vari
ance;

      Proportion_subject=Residual_variance_subject/Total
_variance;

      Residual_ICC=Residual_variance_cluster/(residual_v
ariance_cluster+residual_variance_subject);
      run;

      title 'Total variance';
      proc print data=outcome label noobs;
            label     total_variance="total     variance"
proportion_explained="explained"
proportion_subject="subject"
proportion_cluster="cluster";
            format     prevalence     proportion_cluster
proportion_explained proportion_subject percent10.1;
```

```
            var         prevalence         total_variance
proportion_cluster                    proportion_explained
proportion_subject;
      run;

      title 'Residual variance';
      proc print data=outcome label noobs;
            label residual_icc="residual ICC";
            format residual_icc percent10.1;
            var Residual_ICC;
      run;

      data _null_;
      call symput ("var_cluster", ".");
      call symput ("var_explained", ".");
      run;

%END;

%IF &type=ord %THEN %DO;

      proc glimmix data=&data method=laplace; /*first we
compute the model including both within and between
effects*/
            class &cluster ;
            model &score =&predictor &between &macropred/
link=cumlogit dist=multinomial  s;
            random &cluster;
            ods output ParameterEstimates=fixed;
      run;

      data fixed;
            set fixed;
            effect=upcase(effect);
      run;

      %let  n=%sysfunc(countw(&predictor));  /*we  compute
the within-effect*/

      %do var=1 %to &n;
            %let withinvar= %qscan (&predictor, &var);

            data _null_;
                  set fixed;
                  if effect="&withinvar" then call symput
('estimate', estimate);
```

```
            run;

            %put &estimate;

            data &data;
                  set &data;
                  coeff=&estimate;
                  part_var_&var = &withinvar * coeff;
            run;
      %end;

      data &data;
            set &data;
            within=sum(of part_var_1-part_var_&n);
            drop coeff part_var_1-part_var_&n;
      run;

      proc glimmix data=&data method=laplace; /* we
continue with only the within effect, as an offset*/
            class &cluster ;
            model     &score=     &macropred/offset=within
link=cumlogit dist=multinomial  s;
            random intercept /sub=&cluster solution;
            output out=lp  pred(noblup)=lp;
            ods       output       CovParms=cluster_resid
solutionr=RI; /* write variances to total_cluster_resid*/
      run;

      data lp_sort; set lp; run;
      proc sort data=lp_sort; by _level_;run;

      proc means data=lp_sort var;
            where lp ne 0 and _level_=1;  /* sas puts the
lp to 0 if the cluster has a missing value. Since these
are not used for the computation of the variances in the
multilevel model, we don't use them for the variance of
the fixed part either. We use level one. This is arbitrary
but the variance is identical for all levels.*/
            var lp;
            output out=var_lp var=var_lp;
      run;

      data _null_;
            set var_lp;
            if _type_=0 then call symput ("var_explained",
var_lp);
```

```
      run;

      data _null_;
            set cluster_resid;
            if  CovParm="Intercept"  then  call  symput
("var_cluster", Estimate);
      run;

      data outcome;
            Residual_variance_cluster=&var_cluster;
            pi=constant("PI");
            Residual_variance_subject=pi*pi/3;
            Variance_explained=&var_explained;

      Total_variance=Residual_variance_cluster+Residual_
variance_subject+variance_explained;

      Proportion_cluster=Residual_variance_cluster/Total
_variance;

      Proportion_explained=Variance_explained/Total_vari
ance;

      Proportion_subject=Residual_variance_subject/Total
_variance;

      Residual_ICC=Residual_variance_cluster/(residual_v
ariance_cluster+residual_variance_subject);
      run;

      title 'Total variance';
      proc print data=outcome label noobs;
            label     total_variance="total     variance"
proportion_explained="explained"
proportion_subject="subject"
proportion_cluster="cluster";
            format                      proportion_cluster
proportion_explained proportion_subject percent10.1;
            var     total_variance     proportion_cluster
proportion_explained proportion_subject;
      run;

      title 'Residual variance';
      proc print data=outcome label noobs;
            label residual_icc="residual ICC";
            format residual_icc percent10.1;
```

```
            var Residual_ICC;
      run;

      data _null_;
      call symput ("var_cluster", ".");
      call symput ("var_explained", ".");
      run;


%END;

%MEND variance_partition;
```

A6 Difference in net sensitivity and net amount of avoided false positives per 100 patients

Table A 2. Increase in net sensitivity (%) of the IOTA Simple Rules risk scoring system (SRs risks) compared to the Risk of Malignancy Index (RMI), the LR2 model by IOTA, and the ADNEX model by IOTA for all patients and for oncology and other centers.

All patients	**Increase in net sensitivity [95% CI]**		
Threshold probability	SRs risks vs LR2	SRs risks vs RMI	SRs risks vs ADNEX
5%	2% [1% to 4%]	RMI harmful	0% [-1% to 1%]
10%	4% [3% to 6%]	RMI harmful	0% [-1% to 2%]
15%	4% [2% to 6%]	RMI harmful	0% [-2% to 1%]
20%	5% [3% to 7%]	20% [17% to 23%]	0% [-1% to 2%]
30%	5% [3% to 8%]	16% [13% to 19%]	1% [-1% to 3%]
50%	1% [-2% to 5%]	6% [2% to 10%]	-6% [-9% to -2%]
Oncology centers	**Increase in net sensitivity [95% CI]**		
Threshold probability	SRs risks vs LR2	SRs risks vs RMI	SRs risks vs ADNEX
5%	LR2 harmful	RMI harmful	-1% [-2% to 0%]
10%	4% [2% to 6%]	RMI harmful	0% [-1% to 1%]
15%	3% [1% to 5%]	RMI harmful	0% [-2% to 1%]
20%	5% [3% to 7%]	RMI harmful	0% [-1% to 2%]
30%	6% [4% to 9%]	17% [13% to 20%]	1% [-1% to 4%]
50%	1% [-2% to 5%]	3% [-1% to 7%]	-7% [-10% to -3%]
Non-oncology centers	**Increase in net sensitivity [95% CI]**		
Threshold probability	SRs risks vs LR2	SRs risks vs RMI	SRs risks vs ADNEX
5%	5% [1% to 9%]	RMI harmful	2% [0% to 5%]
10%	9% [3% to 15%]	29% [20% to 37%]	3% [-1% to 8%]
15%	7% [0% to 14%]	25% [16% to 33%]	2% [-4% to 8%]
20%	5% [-2% to 14%]	20% [10% to 29%]	1% [-6% to 9%]
30%	1% [-7% to 11%]	13% [2% to 23%]	-4% [-12% to 3%]
50%	2% [-10% to 13%]	21% [6% to 37%]	2% [-9% to 13%]

Table A 3. Net amount of additional false positives avoided per 100 patients by the IOTA Simple Rules risk scoring system (SRs risks) compared to the Risk of Malignancy Index (RMI), the LR2 model by IOTA, and the ADNEX model by IOTA for all patients, oncology centers and other centers.

All women	**Avoided false positives per 100 women [95% CI]**		
Threshold probability	SRs risks vs LR2	SRs risks vs RMI	SRs risks vs ADNEX
5%	19 [10 to 30]	RMI harmful	-2 [-8 to 5]
10%	16 [10 to 23]	RMI harmful	1 [-3 to 6]
15%	9 [4 to 13]	RMI harmful	0 [-4 to 3]
20%	9 [5 to 12]	32 [27 to 38]	1 [-2 to 4]
30%	5 [3 to 8]	15 [12 to 19]	1 [-1 to 3]
50%	0 [-1 to 2]	2 [1 to 4]	-2 [-4 to -1]
Oncology centers	**Avoided false positives per 100 women [95% CI]**		
Threshold probability	SRs risks vs LR2	SRs risks vs RMI	SRs risks vs ADNEX
5%	LR2 harmful	RMI harmful	-6 [-15 to 3]
10%	17 [10 to 25]	RMI harmful	0 [-4 to 5]
15%	9 [4 to 15]	RMI harmful	-1 [-5 to 3]
20%	10 [6 to 15]	RMI harmful	1 [-3 to 4]
30%	7 [4 to 10]	19 [15 to 24]	2 [-1 to 4]
50%	1 [-1 to 3]	2 [0 to 4]	-3 [-5 to -2]
Non-oncology centers	**Avoided false positives per 100 women [95% CI]**		
Threshold probability	SRs risks vs LR2	SRs risks vs RMI	SRs risks vs ADNEX
5%	16 [4 to 31]	RMI harmful	8 [-1 to 19]
10%	15 [5 to 27]	48 [32 to 66]	5 [-2 to 13]
15%	8 [0 to 16]	26 [15 to 38]	3 [-4 to 8]
20%	4 [-2 to 11]	15 [7 to 24]	1 [-5 to 7]
30%	0 [-3 to 5]	6 [1 to 10]	-2 [-5 to 1]
50%	0 [-2 to 2]	4 [1 to 7]	0 [-2 to 2]

A7 Center-specific case-mix and population models to generate a heterogeneous multicenter dataset

Table A 4. The center-specific distribution of the predictor X and the true association between the predictor X and the outcome Y, the number of patients sampled for the validation dataset from center j (n_j), the outcome prevalence in center j (p_j), the observed center-specific c-statistic (C), the observed calibration intercept and the observed calibration slope.

Center	X	Model	n_j	p_j	C	Calibration intercept	calibration slope
1	X~N(0.2,1)	logit*[b]=-2+2*X	101	0.27	0.91	-0.14	1.36
2	X~N(0.2,1)	logit=-2+2*X	94	0.24	0.80	-0.28	0.90
3	X~N(0.2,1)	logit=-2+2*X	94	0.29	0.94	0.04	1.72
4	X~N(0.2,1)	logit=-2+2*X	108	0.29	0.82	0.26	1.01
5	X~N(0.2,1)	logit=-2+2*X	114	0.23	0.89	-0.33	1.22
6	X~N(0.2,1)	logit=-2+2*X	95	0.32	0.88	0.04	1.33
7	X~N(0.2,1)	logit=-2+2*X	91	0.25	0.92	0.05	1.52
8	X~N(0.2,1)	logit=-2+2*X	104	0.29	0.89	0.13	1.22
9	X~N(0.2,1)	logit=-2+2*X	98	0.23	0.90	-0.40	1.74
10	X~N(0.2,1)	logit=-2+2*X	89	0.36	0.90	0.29	1.36
11	X~N(0.2,1)	logit=-2+3.09*X	96	0.31	0.89	0.23	1.45
12	X~N(0.2,1)	logit=-2+1.40*X	104	0.25	0.83	0.04	1.02
13	X~N(0.2,1)	logit=-2+3.10*X	100	0.42	0.95	0.86	2.28
14	X~N(0.2,1)	logit=-2+2.44*X	90	0.36	0.94	0.61	1.77

15	X~N(0.2,1)	logit=-2+0.60*X	99	0.11	0.67	-1.43	0.40
16	X~N(0.2,1)	logit=-2+2.07*X	103	0.26	0.87	0.10	1.25
17	X~N(0.2,1)	logit=-2+1.77*X	86	0.23	0.91	-0.31	1.37
18	X~N(0.2,1)	logit=-2+1.29*X	104	0.17	0.77	-0.66	0.71
19	X~N(0.2,1)	logit=-2+1.31*X	89	0.19	0.79	-0.40	0.68
20	X~N(0.2,1)	logit=-2+0.80*X	101	0.19	0.77	-0.55	0.77
21	X~N(0.2,1)	logit=-1+2*X	90	0.42	0.93	0.93	1.65
22	X~N(0.2,1)	logit=-3+2*X	99	0.15	0.88	-0.89	1.21
23	X~N(0.2,1)	logit=-3+2*X	105	0.15	0.84	-0.91	0.98
24	X~N(0.2,0.5)	logit=-2+2*X	94	0.23	0.73	-0.02	1.26
25	X~N(0.2,1.32)	logit=-2+2*X	115	0.25	0.91	-0.57	1.14
26	X~N(0.2,0.5)	logit=-1+2*X	93	0.38	0.81	0.79	1.85
27	X~N(0.2,1.32)	logit=-1+2*X	82	0.48	0.93	1.07	1.37
28	X~N(0.2,0.5)	logit=-3+2*X	89	0.09	0.95	-1.12	3.20
29	X~N(0.2,1.32)	logit=-3+2*X	93	0.17	0.90	-1.18	1.08
30	X~N(0.2,1); select *[a] X<x~N(0,0.8)	logit=-2+2*X	104	0.13	0.78	0.10	1.09
31	X~N(0.2,1); select X>x~N(0,0.8)	logit=-2+2*X	93	0.54	0.74	0.41	0.87
32	X~N(0.2,1); select X<x~N(0,0.8)	logit=--1+2*X	125	0.26	0.69	1.04	0.64
33	X~N(0.2,1); select X>x~N(0,0.8)	logit=-1+2*X	94	0.53	0.87	0.89	1.52

34	X~N(0.2,1); select X<x~N(0,0.8)	logit=-3+2*X	66	0.05	0.68	-1.46	0.58
35	X~N(0.2,1); select X>x~N(0,0.8)	logit=-3+2*X	98	0.21	0.84	-1.20	1.35
36	X~N(0.2,1)	logit=-1+1.02*X	110	0.41	0.66	0.82	0.45
37	X~N(0.2,1)	logit=-3+2.93*X	104	0.26	0.92	0.18	1.77
38	X~N(0.2,1)	logit=-3+2.28*X	81	0.12	0.87	-0.99	1.23
39	X~N(0.2,0.5)	logit=-2+1.13*X	100	0.12	0.58	-0.57	0.35
40	X~N(0.2,1.32)	logit=-2+1.17*X	113	0.19	0.79	-0.99	0.57
41	X~N(0.2,0.5)	logit=-1+0.62*X	106	0.34	0.65	0.75	0.63
42	X~N(0.2,1.32)	logit=-1+2.13*X	109	0.47	0.93	1.36	1.49
43	X~N(0.2,0.5)	logit=-3+0.93*X	107	0.10	0.58	-1.02	0.26
44	X~N(0.2,1.32)	logit=-3+1.42*X	98	0.12	0.82	-1.78	0.69
45	X~N(0.2,1); select X<x~N(0,0.8)	logit=-2+1.58*X	110	0.13	0.76	0.06	0.66
46	X~N(0.2,1); select X>x~N(0,0.8)	logit=-2+3.38*X	116	0.55	0.88	0.81	1.80
47	X~N(0.2,1); select X<x~N(0,0.8)	logit=-1+1.56*X	113	0.19	0.81	0.60	0.86
48	X~N(0.2,1); select X>x~N(0,0.8)	logit=-1+2.21*X	97	0.71	0.91	1.48	2.02
49	X~N(0.2,1); select X<x~N(0,0.8)	logit=-3+1.245*X	101	0.04	0.77	-1.30	0.80
50	X~N(0.2,1); select X>x~N(0,0.8)	logit=-3+2.08*X	102	0.25	0.80	-0.83	1.06

*[a] Comparing X to x and selecting X only if X<x is done to generate homogeneous centers with a less severe case-mix in terms of predictor X. Homogeneous centers with a more severe case-mix in terms of predictor X are chosen selecting X>x. *[b] logit=log($p(Y_{ij}=1)/(1-p(Y_{ij}=1))$), where Y is the dichotomous outcome we want to predict. In this column, the true prediction equation is given. Y outcomes are generated based on this prediction equation and the simulated X values. An element of randomness in the Y is obtained by comparing the $p(Y_{ij}=1)$ to a randomly drawn U_{ij} from a uniform distribution supported on [0,1], and assigning a value of 1 to Y_{ij} if it is greater than U_{ij}.

A8 Observed differences in center-specific decision curves in simulated data

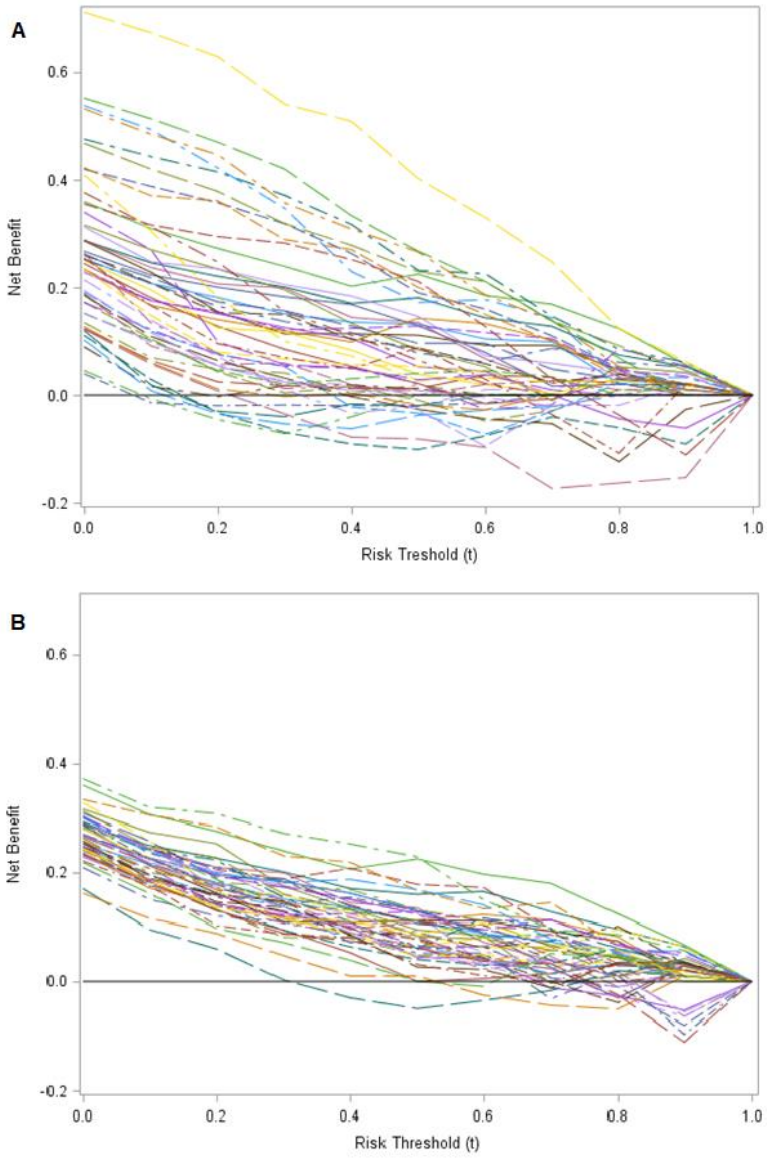


Figure A 3. Observed differences in center-specific decision curves in a simulated heterogeneous (panel A) and homogeneous (panel B) population.

A9 Results for meta-analyses of center-specific net benefit using various weights

Table A 5. Results for meta-analyses of center-specific net benefit using various weights.

t	n_j	1/J	$n_{events\ j}$	inverse variance	random effects	random effects log(1-NB)
0	0.27 (0.26-0.28)	0.27 (0.25-0.28)	0.34 (0.32-0.35)	0.21 (0.20-0.22)	0.26 (0.22-0.30) [-0.01-0.53]	0.27 (0.23-0.31) [-0.05-0.49]
0.1	0.20 (0.19-0.22)	0.20 (0.19-0.22)	0.28 (0.26-0.30)	0.14 (0.13-0.16)	0.20 (0.16-0.24) [-0.09-0.49]	0.21 (0.16-0.25) [-0.13-0.44]
0.2	0.16 (0.15-0.17)	0.16 (0.15-0.17)	0.23 (0.22-0.25)	0.10 (0.08-0.11)	0.15 (0.11-0.20) [-0.13-0.44]	0.16 (0.12-0.20) [-0.18-0.40]
0.3	0.12 (0.11-0.14)	0.13 (0.11-0.14)	0.19 (0.17-0.20)	0.06 (0.05-0.07)	0.12 (0.08-0.16) [-0.14-0.38]	0.12 (0.08-0.16) [-0.17-0.34]
0.4	0.10 (0.09-0.11)	0.10 (0.09-0.11)	0.16 (0.14-0.17)	0.04 (0.03-0.05)	0.10 (0.06-0.13) [-0.13-0.32]	0.10 (0.06-0.13) [-0.15-0.29]
0.5	0.08 (0.07-0.09)	0.08 (0.07-0.09)	0.12 (0.11-0.14)	0.02 (0.01-0.03)	0.07 (0.05-0.10) [-0.11-0.26]	0.07 (0.05-0.10) [-0.13-0.24]
0.6	0.06 (0.04-0.07)	0.06 (0.04-0.07)	0.09 (0.08-0.11)	0.02 (0.01-0.02)	0.05 (0.03-0.08) [-0.10-0.21]	0.05 (0.03-0.08) [-0.11-0.19]
0.7	0.04 (0.02-0.05)	0.04 (0.02-0.05)	0.06 (0.05-0.08)	0.01 (0.00-0.01)	0.04 (0.02-0.06) [-0.07-0.15]	0.04 (0.02-0.05) [-0.08-0.14]
0.8	0.02 (0.00-0.04)	0.02 (0.00-0.04)	0.03 (0.02-0.05)	0.01 (0.00-0.01)	0.02 (0.01-0.03) [-0.02-0.06]	0.02 (0.01-0.03) [-0.02-0.05]
0.9	0.01 (-0.02-0.04)	0.01 (-0.02-0.04)	0.02 (-0.01-0.05)	0.01 (0.00-0.01)	0.01 (0.01-0.01) [0.01-0.01]	0.01 (0.00-0.01) [0.01-0.01]
1	0.00 (0.00-0.00)	0.00 (0.00-0.00)	0.00 (0.00-0.00)	0.00 (0.00-0.00)	0.00 (0.00-0.00) [0.00-0.00]	0.00 (0.00-0.00) [0.00-0.00]

Summary estimate (95% confidence interval) [95% prediction interval]. t: risk threshold; n_j: number of observations per center; 1/J: equal weights, with J the number of centers; $n_{events\ j}$: number of events per center.

A10 Additional results of the random effects meta-analysis of NB in IOTA data

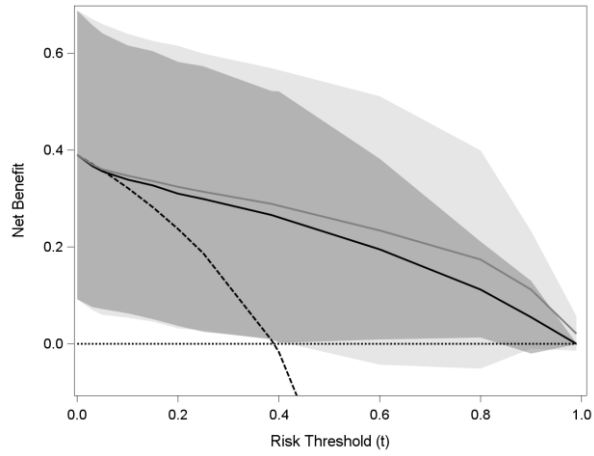


Figure A 4. Decision curves for LR2 and a refitted LR2 model obtained by random effects analysis of net benefit in oncology centers. Black full curve and dark prediction interval: original LR2; grey curve and light prediction interval: refitted LR2; black dashed curve: treat all in oncology centers; black dotted horizontal line: treat none.

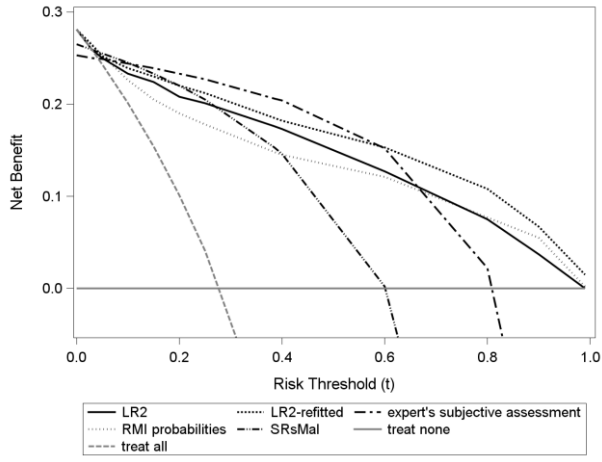


Figure A 5. Decision curves for LR2, a refitted LR2 model, a probability model of RMI, The Simple Rules classifying inconclusive cases as malignant, expert's subjective assessment, treat all and treat none, obtained by random effects analysis of net benefit.

A11 Formulas for the c-statistic and logistic calibration in a comprehensive framework

Table A 6. Formulas for the c-statistic and logistic calibration in a comprehensive framework of the standard and within-center validation of marginal predictions, conditional predictions and predictions assuming an average random intercept.

	Marginal predictions	Conditional predictions	Predictions assuming an average random intercept
Standard validation			
c-statistic	$\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(\hat{p}_{LR\,ij} > \hat{p}_{LR\,i'j'} \text{ and } y_{ij}=1 \text{ and } y_{i'j'}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(y_{ij}=1 \text{ and } y_{i'j'}=0\right)}$	$\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(\hat{p}_{MLR\,c\,ij} > \hat{p}_{MLR\,c\,i'j'} \text{ and } y_{ij}=1 \text{ and } y_{i'j'}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(y_{ij}=1 \text{ and } y_{i'j'}=0\right)}$	$\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(\hat{p}_{MLR\,a\,ij} > \hat{p}_{MLR\,a\,i'j'} \text{ and } y_{ij}=1 \text{ and } y_{i'j'}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(y_{ij}=1 \text{ and } y_{i'j'}=0\right)}$
Calibration slope β_{cal}	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+\beta_{cal}LP_{LR\,ij}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+\beta_{cal}LP_{MLR\,c\,ij}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+\beta_{cal}LP_{MLR\,a\,ij}$
Calibration intercept α_{cal}	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+LP_{LR\,ij}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+LP_{MLR\,c\,ij}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+LP_{MLR\,a\,ij}$
Conditional validation			
c-statistic	$\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(\hat{p}_{LR\,ij} > \hat{p}_{LR\,i'j} \text{ and } y_{ij}=1 \text{ and } y_{i'j}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(y_{ij}=1 \text{ and } y_{i'j}=0\right)}$	$\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(\hat{p}_{MLR\,c\,ij} > \hat{p}_{MLR\,c\,i'j} \text{ and } y_{ij}=1 \text{ and } y_{i'j}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(y_{ij}=1 \text{ and } y_{i'j}=0\right)}$	$\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(\hat{p}_{MLR\,a\,ij} > \hat{p}_{MLR\,a\,i'j} \text{ and } y_{ij}=1 \text{ and } y_{i'j}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(y_{ij}=1 \text{ and } y_{i'j}=0\right)}$
Calibration slope $\beta_{cal\,w}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+\beta_{cal\,w}LP_{LR\,ij}+b_jLP_{LR\,ij}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+\beta_{cal\,w}LP_{MLR\,c\,ij}+b_jLP_{MLR\,c\,ij}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+\beta_{cal\,w}LP_{MLR\,a\,ij}+b_jLP_{MLR\,a\,ij}$
	$\binom{a_j}{b_j}\sim N\left(\begin{matrix}0\\0\end{matrix},\begin{matrix}\tau_a^2 & \tau_{ab}\\ \tau_{ab} & \tau_b^2\end{matrix}\right)$	$\binom{a_j}{b_j}\sim N\left(\begin{matrix}0\\0\end{matrix},\begin{matrix}\tau_a^2 & \tau_{ab}\\ \tau_{ab} & \tau_b^2\end{matrix}\right)$	$\binom{a_j}{b_j}\sim N\left(\begin{matrix}0\\0\end{matrix},\begin{matrix}\tau_a^2 & \tau_{ab}\\ \tau_{ab} & \tau_b^2\end{matrix}\right)$
Calibration intercept $\alpha_{cal\,w}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+LP_{LR\,ij}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+LP_{MLR\,c\,ij}$	$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+LP_{MLR\,a\,ij}$
	$(a_j)\sim N(0,\tau_a^2)$	$(a_j)\sim N(0,\tau_a^2)$	$(a_j)\sim N(0,\tau_a^2)$

i: patient indicator; j: center indicator; n_j: number of observations in center j; J: number of centers; y: event indicator; p: probability; $\hat{p}$: predicted probability; LP: linear predictor; LR: standard logistic regression; MLR c: mixed effects logistic regression (conditional); MLR a: mixed effects logistic regression (assuming an average random intercept); β_{cal} : calibration slope; $\beta_{cal\,w}$: within-cluster calibration slope; b_j: within-center calibration slope; α_{cal} : calibration slope; $\alpha_{cal\,w}$: random calibration slope; a_j: random calibration slope; τ_a^2: variance of the random calibration intercepts; τ_a^2: variance of the random calibration slopes; τ_{ab}: covariance of the random calibration intercept and slope.

A12 Calibration with a biased estimate of the between-center variance

If two mixed effects models are fitted on the same dataset, but one has a biased estimate of the between-center variance $\hat{\tau}'^2$ (e.g., underestimated in a sample with very few centers), the following equations hold:

$$\frac{\hat{\beta}_{\text{conditional}}}{\sqrt{1+\tau^2 c^2}} \approx \frac{\hat{\beta}_{\text{conditional'}}}{\sqrt{1+\hat{\tau}'^2 c^2}}$$

$$\hat{\beta}_{\text{conditional'}} \approx \hat{\beta}_{\text{conditional}} \times \sqrt{\frac{1+\hat{\tau}'^2 c^2}{1+\tau^2 c^2}},$$

with c defined as above. Hence, the within-center calibration slopes of these two models will differ by factor $\sqrt{\frac{1+\tau^2 c^2}{1+\hat{\tau}'^2 c^2}}$. If the model with the correct estimate of the between-center variance is perfectly calibrated, the model with the biased estimate will yield calibration slopes smaller than one if τ^2 is overestimated and calibration slopes larger than one if τ^2 is underestimated.

A13 Correspondence between predictions from the standard logistic regression model and marginalized predictions from the mixed effects logistic regression model

The logistic regression model and the mixed effects logistic regression model were developed in a sample with 100 EPV, drawn from a population with clustered data (ICC 20%). The predictions depicted are the marginalized predicted probabilities of the mixed effects logistic regression model and marginal predicted probabilities obtained by the standard logistic regression model.

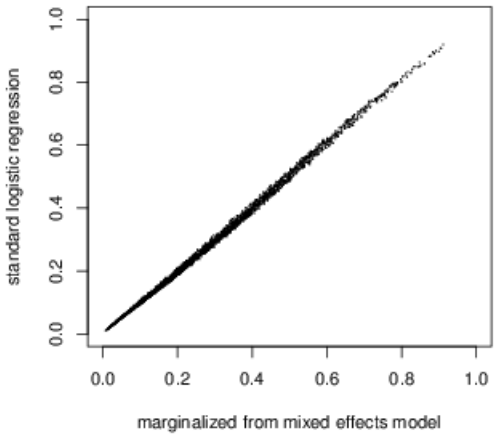


Figure A 6. Marginal predictions from a standard logistic regression model versus marginalized predictions from a mixed effects logistic regression model.

A14 R code for the simulation study of the impact of ignoring clustering on predictive performance

```
#####1. Generation of source populations
rm(list=ls(all=TRUE))
library (rms)
library (lme4)

set.seed(1986)

#Step 1. Fix the number of clusters (nb) and draw the number of observations
per #cluster (n2) from a poisson distribution.
nb <- 20
mean <- 6.82
sd <- 0.3
L1<-rnorm(nb, mean, sd)
L<-round(exp(L1))
n2 <- rpois(nb, L)
cent <-seq(1:nb)
center <-rep(cent, n2)
n <- length(center) #number of observations in the source population
pat_nr <- seq(1:n)

#Step 2. Generate random cluster intercepts (randeff).
set.seed(2022)
randeff <- rnorm(nb, 0, 0.9068997)  # this is to obtain an ICC of 20%
randeffect <-rep(randeff, n2)

## Step 3. Generate predictors x1 to x8
x1      <- rnorm(n, 0, 1)
x2      <- rnorm(n, 0, 0.6)
x3      <- rnorm(n, 0, 0.4)
x4      <- rnorm(n, 0, 0.2)
x5      <- rbinom(n, 1, 0.20)
x6      <- rbinom(n, 1, 0.30)
x7      <- rbinom(n, 1, 0.30)
x8      <- rbinom(n, 1, 0.40)

# Step 4: Set beta coefficients and obtain the logit.
b1      <- 0.8
b2      <- 0.8
b3      <- 0.8
```

```
b4      <- 0.8
b5      <- 0.8
b6      <- 0.8
b7      <- 0.8
b8      <- 0.8
logit   <-                                                          -
2.1+b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+randeffect

#Step 5: Calculate the predicted probabilities of experiencing an event
p       <- plogis(logit)

#Step 6: Obtain dichotomous outcome y
y       <- ifelse(runif(n)<=p, 1, 0)
data <- as.data.frame(cbind(pat_nr, x1, x2, x3, x4, x5, x6, x7, x8, randeffect,
center, y), dimnames = list(c(1:n), Cs(pat_nr, x1, x2, x3, x4, x5, x6, x7, x8,
randeffect, center, y)))

####2.Sampling from the source population and model building within
samples
#This code corresponds to sampling with 100 events per variable. The code
can be adjusted for other conditions with minor changes.
rm(list=ls(all=TRUE))
library (Hmisc)
library (rms)
library (lme4)
library(arm)
Nsample <- 1000  #specify the number of samples to be drawn

#Step 1. Specify the number of clusters
Nsamplev2 <-20

#Step 2. Specify the number of events to be drawn
NEVENTS <- 900
NNONEVENTS <- round((NEVENTS/mean(data$y))-NEVENTS)

#Matrices to store results
Ncenter_sample <- matrix(NA, nrow=Nsample, ncol=1)
center_sample <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
Nevents_sample <- matrix(NA, nrow=Nsample, ncol=1)
Npats_sample <- matrix(NA, nrow=Nsample, ncol=1)
Pevents_sample <- matrix(NA, nrow=Nsample, ncol=1)
Npats_center_sample <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)

mat_sd_ranef0 <- matrix(NA, nrow=Nsample, ncol=1)
```

```
mat_sd_ranef1 <- matrix(NA, nrow=Nsample, ncol=1)
mat_regcof <- matrix(NA, nrow=Nsample, ncol=9)
mat_SEregcof <- matrix(NA, nrow=Nsample, ncol=9)
mat_ranef <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
mat_SEranef <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
Niterm1 <- matrix(NA, nrow=Nsample, ncol=1)
Equaldeviance <- matrix(NA, nrow=Nsample, ncol=1)

mat_regcofu <- matrix(NA, nrow=Nsample, ncol=9)
mat_SEregcofu <- matrix(NA, nrow=Nsample, ncol=9)
Niterm1u <- matrix(NA, nrow=Nsample, ncol=1)
Equaldevianceu <- matrix(NA, nrow=Nsample, ncol=1)

realvalresultB <- matrix(NA,nrow=Nsample, ncol=8) #Validation results for
predictions average random intercept
dimnames(realvalresultB) <- list(c(1:Nsample), Cs(C,calib_slope,
calib_interc, Cw,ML_calib_slope,var_ML_calib_slope,
ML_calib_interc,var_ML_calib_interc))
realvalresultBcond <- matrix(NA,nrow=Nsample, ncol=8) ) #Validation
results for conditional predictions
dimnames(realvalresultBcond) <- list(c(1:Nsample), Cs(C,calib_slope,
calib_interc, Cw,ML_calib_slope,var_ML_calib_slope,
ML_calib_interc,var_ML_calib_interc))
realvalresultBu <- matrix(NA,nrow=Nsample, ncol=8) ) #Validation results
for marginal predictions
dimnames(realvalresultBu) <- list(c(1:Nsample), Cs(C,calib_slope,
calib_interc, Cw,ML_calib_slope,var_ML_calib_slope,
ML_calib_interc,var_ML_calib_interc))loglik0 <-
matrix(NA,nrow=Nsample,ncol=1)

loglik0 <- matrix(NA,nrow=Nsample,ncol=1)
loglik1 <- matrix(NA,nrow=Nsample,ncol=1)
loglik0u <- matrix(NA,nrow=Nsample,ncol=1)
loglik1u <- matrix(NA,nrow=Nsample,ncol=1

#Start sampling from the domain
tmpA <- aggregate(data$pat_nr, list(center = data$center), FUN=length)
pt_center <- as.vector(tmpA[,2])
names(pt_center) <- tmpA[,1]

for (r in 1:Nsample){
set.seed(r)

#Step 3. Sample clusters (samp_center) from the source population
```

```
samp_center <- names(pt_center)

#Get the data from sampled clusters (dsample_lev2)
loc.pt_center <- pt_center[samp_center]
indices.cutoffs <- c(0,cumsum(loc.pt_center))
ptnrs_samp2 <- matrix(NA,nrow=sum(loc.pt_center),ncol=1)
for (i in 1:Nsamplev2){
ptnrs_samp2[(indices.cutoffs[i]+1):indices.cutoffs[i+1],] <-
data[data$center==names(loc.pt_center)[i],1]
}
dsample_lev2 <- data[ptnrs_samp2,]
dsample_lev2<-dsample_lev2[order(dsample_lev2$center), ]

#Step 4. Stratified sampling of events (sample_lev1_y1) and non-events
(sample_lev1_y0).
sdatay1 <- dsample_lev2[dsample_lev2$y==1,]
pt_nr_y1 <- as.vector(sdatay1$pat_nr)
sample_lev1_y1 <- sample(pt_nr_y1, NEVENTS , replace = FALSE)
sdatay0 <- dsample_lev2[dsample_lev2$y==0,]
pt_nr_y0 <- as.vector(sdatay0$pat_nr)
sample_lev1_y0 <- sample(pt_nr_y0, NNONEVENTS, replace = FALSE)
sample_lev1 <- c(sample_lev1_y0, sample_lev1_y1)
dsample <- data[sample_lev1,]

#Step 5. Fit a random intercept model in the sample: full model (m1) and null
#model (m0)
        #mixed effects model
iter_m1<-capture.output(m1<-
lmer(y~x1+x2+x3+x4+x5+x6+x7+x8+(1|center), family = "binomial",
data=dsample, verbose=T ))
m0<-lmer(y~(1|center), family = "binomial", data=dsample)
Last2it<-iter_m1[c(length(iter_m1)-1):c(length(iter_m1))]
Last2itsplit<-strsplit(Last2it,":")
Equal<- Last2itsplit [[1]][2]== Last2itsplit [[2]][2]#check convergence
        #standard model
m1u<-glm(y~x1+x2+x3+x4+x5+x6+x7+x8,family=binomial ,data=dsample,
maxit=50) #deviance en fixed parameter schattingen voor alle iteraties tijdens
het fitten
m0u<-glm(y~ 1,family=binomial, data=dsample, maxit=50)#nulmodel
Equalu<- m1u$converge

# Set Npats_center, center and ranef to missing when we sampled <1
observation per cluster
center_s <- length(unique(dsample$center))
```

```
N_missing_center <- Nsamplev2-Ncenter_s
add_center <- rep(x=-99, times=N_missing_center)
if (Ncenter_s < Nsamplev2){
center_s <- c(unique(dsample$center), add_center)
Npats_center_s <- c((aggregate(dsample$pat_nr, list(center = dsample$center), FUN=length)[,2]),add_center)
clustereff_s <- c(attr(m1, "ranef"),add_center)
clustereff_s_SE <- c(as.numeric(se.ranef(m1)$center), add_center)
}
if (Ncenter_s == Nsamplev2) {
Npats_center_s <- aggregate(dsample$pat_nr, list(center = dsample$center), FUN=length)[,2]
center_s <- unique(dsample$center)
clustereff_s <- attr(m1, "ranef")
clustereff_s_SE <- as.numeric(se.ranef(m1)$center)
}

#save sample summaries
Ncenter_sample[r,] <- Ncenter_s
center_sample[r,] <- center_s
Npats_center_sample[r,] <- Npats_center_s
Nevents_sample[r,] <- sum(dsample$y)
Npats_sample[r,] <- length(dsample[,1])
Pevents_sample[r,] <- mean(dsample$y)

#save model parameters
mat_sd_ranef0[r,]<- sigma.hat(m0)$sigma$center
mat_sd_ranef1[r,]<- sigma.hat(m1)$sigma$center
mat_regcof[r,]<- as.numeric(fixef(m1))
mat_SEregcof[r,]<- se.coef(m1)$fixef
mat_ranef[r,] <- clustereff_s
mat_SEranef[r,] <- clustereff_s_SE
mat_dims[r,] <- as.numeric(attr(m1,"dims"))
Niterm1[r,] <- length(iter_m1)-1
Equaldeviance[r,] <- Equal
mat_regcofu[r,]<- m1u$coef
mat_SEregcofu[r,]<- summary(m1u)$coefficients[, 2]
Niterm1u[r,] <- m1u$iter
Equaldevianceu[r,] <- Equalu

#save log likelihood of full and null model
loglik0[r,]<-logLik(m0)
loglik1[r,]<-logLik(m1)
loglik0u[r,]<-logLik(m0u)
```

```
loglik1u[r,]<-logLik(m1u)

#Step 6. Evaluate model performance in the test set
        #Predictions assuming average random intercept
datatest<-data.frame(cbind(data[-sample_lev1,2:9],     data$y[-sample_lev1],
data$center[-sample_lev1]))
datatest$lpB <- cbind(1, as.matrix(data[-sample_lev1,2:9])) %*% fixef(m1)
names(datatest)<-c("x1","x2",  "x3",  "x4","x5","x6","x7","x8",  "y","center",
"lpB")
perfm_m1B<- lrm(datatest$y~datatest$lpB)
perfm_m1Bi<- lrm(datatest$y~offset(datatest$lpB))
realvalresultB[r,1] <- perfm_m1B$stats[6] # C
realvalresultB[r,2] <- perfm_m1B$coefficients[2] #calib slope
realvalresultB[r,3] <- perfm_m1Bi$coefficients[1] #calib intercept
C.w.index <- rep(NA, length(unique(data$center)))
n.comp <- rep(NA, length(unique(data$center)))
for(p in 1:length(unique(data$center))){
data_p <- datatest[datatest$center==p,]
prediction <- plogis(data_p$lpB)
outcome <- data_p$y
result <- concordance.prob.logistic(outcome, prediction)
C.w.index[p] <- result$C
n.comp[p] <- result$n.comp
}
realvalresultB[r,4] <- sum(C.w.index*n.comp/sum(n.comp, na.rm = TRUE),
na.rm = TRUE) # Cwithin
calibslope_mlB  <-  lmer(y~lpB+(lpB|center),  family  =  "binomial",
data=datatest)
realvalresultB[r,5] <- as.numeric(fixef(calibslope_mlB)[2]) #within calib
slope
realvalresultB[r,6] <- (sigma.hat(calibslope_mlB)$sigma$center[2])^2 #var
within calib slope
calibint_mlB  <-  lmer(y~offset(lpB)+(1|center),  family  =  "binomial",
data=datatest)
realvalresultB[r,7] <- as.numeric(fixef(calibint_mlB)[1]) # within calib slope
realvalresultB[r,8] <- (sigma.hat(calibint_mlB)$sigma$center[1])^2 #var
within calib slope

        #Conditional predictions
ri<-ifelse(datatest$center==1,clustereff_s[1],
        ifelse(datatest$center==2,clustereff_s[2],
        ifelse (datatest$center==3, clustereff_s[3],
        ifelse (datatest$center==4, clustereff_s[4],
        ifelse (datatest$center==5, clustereff_s[5],
```

```
        ifelse (datatest$center==6, clustereff_s[6],
        ifelse (datatest$center==7, clustereff_s[7],
        ifelse (datatest$center==8, clustereff_s[8],
        ifelse (datatest$center==9, clustereff_s[9],
        ifelse (datatest$center==10, clustereff_s[10],
        ifelse (datatest$center==11, clustereff_s[11],
        ifelse (datatest$center==12, clustereff_s[12],
        ifelse (datatest$center==13, clustereff_s[13],
        ifelse (datatest$center==14, clustereff_s[14],
        ifelse (datatest$center==15, clustereff_s[15],
        ifelse (datatest$center==16, clustereff_s[16],
        ifelse (datatest$center==17, clustereff_s[17],
        ifelse (datatest$center==18, clustereff_s[18],
        ifelse (datatest$center==19, clustereff_s[19],
        ifelse (datatest$center==20, clustereff_s[20],0
        ))))))))))))))))))))
lpc<-datatest$lpB+ri
cond<-as.data.frame(cbind(datatest$center,datatest$y,lpc))
colnames(cond)<-c("center","y","lp")
perfm_m1B<- lrm(cond$y~cond$lp)
realvalresultBcond[r,1] <- perfm_m1B$stats[6] # C
realvalresultBcond[r,2] <- perfm_m1B$coefficients[2] #calib slope
perfm_m1Bi<- lrm(cond$y~offset(cond$lp))
realvalresultBcond[r,3] <- perfm_m1Bi$coefficients[1] #calib intercept
C.w.index <- rep(NA, length(unique(cond$center)))
n.comp <- rep(NA, length(unique(cond$center)))
for(p in 1:length(unique(cond$center))){
data_p <- cond[cond$center==p,]
prediction <- plogis(data_p$lp)
outcome <- data_p$y
result <- concordance.prob.logistic(outcome, prediction)
C.w.index[p] <- result$C
n.comp[p] <- result$n.comp
}
realvalresultBcond[r,4] <- sum(C.w.index*n.comp/sum(n.comp, na.rm =
TRUE), na.rm = TRUE) # Cwithin
calibslope_mlB <- lmer(y~lp+(lp|center), family = "binomial", data=cond)
realvalresultBcond[r,5] <- as.numeric(fixef(calibslope_mlB)[2]) #within
calib slope
realvalresultBcond[r,6] <- (sigma.hat(calibslope_mlB)$sigma$center[2])^2
#var within calib slope
calibinter_mlB <- lmer(y~offset(lp)+(1|center), family = "binomial",
data=cond)
```

```
realvalresultBcond[r,7] <- as.numeric(fixef(calibinter_mlB)[1]) #within calib
intercept
realvalresultBcond[r,8] <- (sigma.hat(calibinter_mlB)$sigma$center[1])^2
#var within calib intercept

#Marginal predictions
datatest$lpBu <- cbind(1, as.matrix(data[-sample_lev1,2:9])) %*% m1u$coef
names(datatest)<-c("x1","x2", "x3", "x4","x5","x6","x7","x8", "y","center",
"lpB", "lpBu")
perfm_m1Bu<- lrm(datatest$y~datatest$lpBu)
realvalresultBu[r,1] <- perfm_m1Bu$stats[6] # C
realvalresultBu[r,2] <- perfm_m1Bu$coefficients[2] #calib slope
perfm_m1Bui<- lrm(datatest$y~offset(datatest$lpBu))
realvalresultBu[r,3] <- perfm_m1Bui$coefficients[1] #calib slope
C.w.index <- rep(NA, length(unique(data$center)))
n.comp <- rep(NA, length(unique(data$center)))
for(p in 1:length(unique(data$center))){
data_p <- datatest[datatest$center==p,]
prediction <- plogis(data_p$lpBu)
outcome <- data_p$y
result <- concordance.prob.logistic(outcome, prediction)
C.w.index[p] <- result$C
n.comp[p] <- result$n.comp
}
realvalresultBu[r,4] <- sum(C.w.index*n.comp/sum(n.comp, na.rm = TRUE),
na.rm = TRUE) # Cwithin
calibslope_mlBu <- lmer(y~lpBu+(lpBu|center), family = "binomial",
data=datatest)
realvalresultBu[r,5] <- as.numeric(fixef(calibslope_mlBu)[2]) #within calib
slope
realvalresultBu[r,6] <- (sigma.hat(calibslope_mlBu)$sigma$center[2])^2
#var within calib slope
calibinter_mlBu <- lmer(y~offset(lpBu)+(1|center), family = "binomial",
data=datatest)
realvalresultBu[r,7] <- as.numeric(fixef(calibinter_mlBu)[1]) # within calib
intercept
realvalresultBu[r,8] <- (sigma.hat(calibinter_mlBu)$sigma$center[1])^2 #var
within calib intercept

#write convergence results
write.table(cbind(r, Niterm1[r,], iter_m1), file="path\\iter", col.names=F,
row.names=F, append=T)
write.table(x=Equaldeviance, file=" path \\EPV 100 ICC
20%\\Equaldeviance", append=F, na="NA", col.names=F)
```

```
write.table(cbind(r, Niterm1u[r,], iter_m1), file=" path \\iteru", col.names=F, row.names=F, append=T)
write.table(x=Equaldevianceu, file=" path \\Equaldevianceu", append=F, na="NA", col.names=F)

} #end of the simulation loop

#Save all results
```

A15 Additional results for the simulation study of the impact of ignoring clustering on predictive performance

A15.1 Detailed calibration results of simulations with EPV=100 and ICC=20%

Table A 7. Calibration slopes and intercepts at the center and population level for marginal predictions from the standard logistic regression model and predictions assuming an average random center intercept and conditional predictions from the random intercept logistic regression model.

Level of performance evaluation	Calibration measure	Type of prediction		
		Marginal predictions	**Predictions assuming an average random center intercept**	**Conditional predictions**
Conditional (center level)	Calibration slope (1 is ideal)	1.09 (0.05) [<0.0005]	0.98 (0.05) [<0.0005]	0.99 (0.05) [<0.0005]
	Calibration intercept (0 is ideal)	-0.13 (0.02) [0.83]	-0.01 (0.04) [0.89]	-0.01 (0.03) [0.05]
Standard (population level)	Calibration slope (1 is ideal)	0.99 (0.05)	0.88 (0.04)	0.98 (0.04)
	Calibration intercept (0 is ideal)	-0.00 (0.02)	0.12 (0.03)	-0.00 (0.03)

Results are presented as mean (sd) [mean between-center variance of the random calibration intercept or slope].

A15.2 Relation between the estimated random intercept variance and the calibration slope

If the random intercept variance was estimated correctly (0.822), the predictions from the random intercept model assuming average random intercepts were well calibrated at the center level. As predicted, the calibration slope was lower when the random intercept variance was overestimated and higher when the random intercept variance was underestimated. In the extreme case where the estimated random intercept variance was set to zero by using a standard logistic regression model for prediction (circles), the within-center calibration slope was higher on average. The association between the estimated random intercept variance and the within-center calibration slope was negative and close to the theoretical approximation.

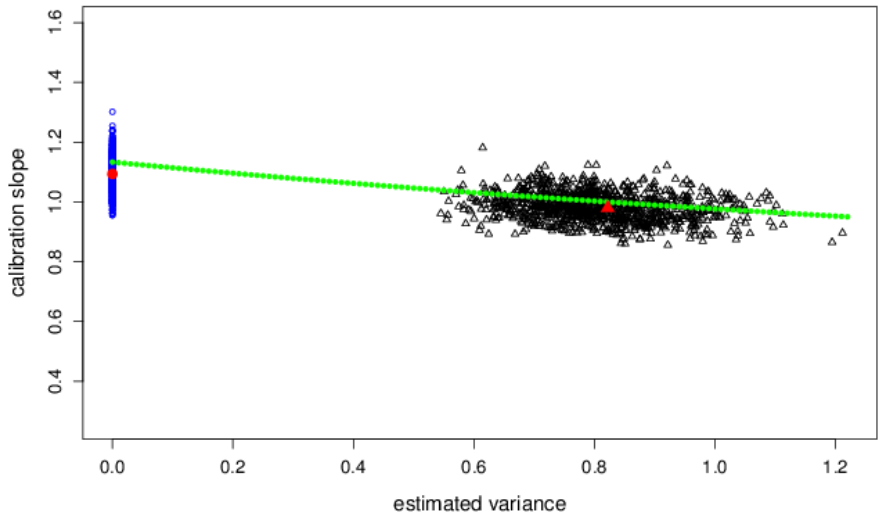


Figure A 7. Center-level calibration slopes of the standard logistic regression model (circles) and the random intercept model assuming average random intercepts (triangles), by estimated random intercept variance. Small symbols indicate calibration slopes observed in the samples, large filled symbols indicate observed average calibration slopes at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.822, ICC=20%) for the random intercept model. The dotted line represents the expected relationship.

A15.3 Calibration intercepts and c-statistics obtained with development samples with EPV 100 in a population with ICC=5%

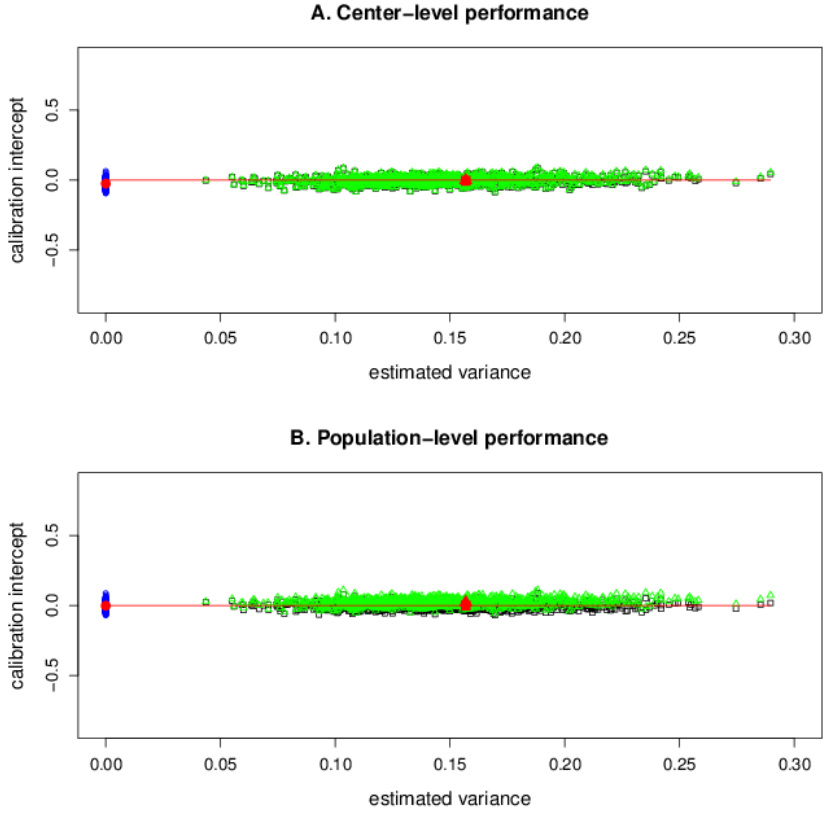


Figure A 8. Center-level (panel A) and population-level (panel B) calibration intercepts of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.157 (ICC=5%). Small symbols indicate calibration intercepts in the samples, large filled symbols indicate average calibration intercepts at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.157) for the random intercept model. The horizontal line represents the ideal calibration intercepts.

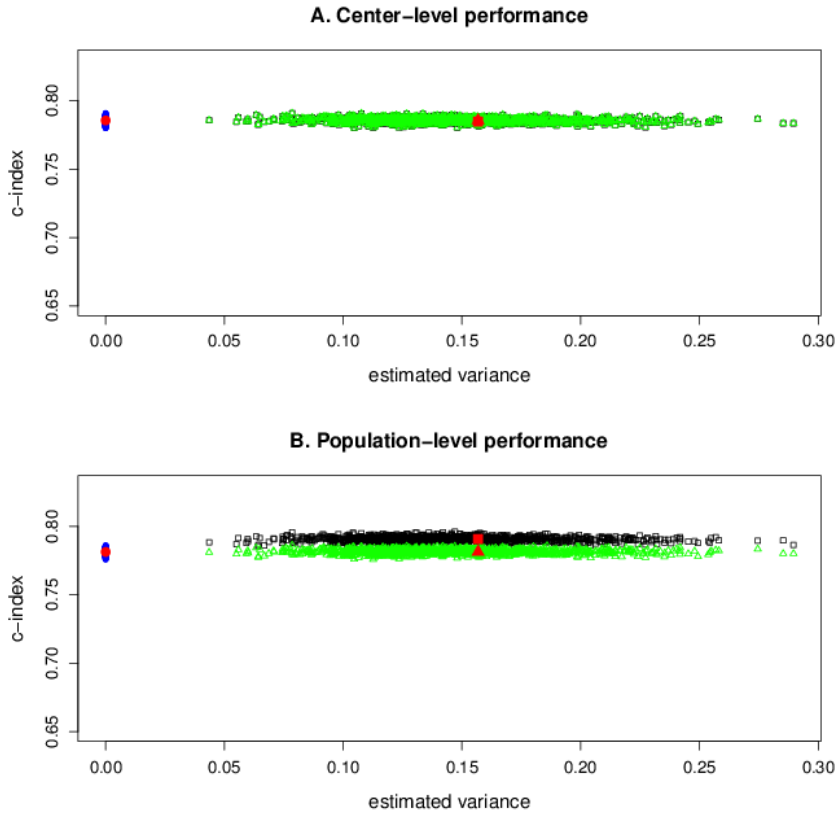


Figure A 9. Center-level (panel A) and population-level (panel B) c-statistics of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.157 (ICC=5%). Small symbols indicate c-statistics in the samples, large filled symbols indicate average c-statistics at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.157) for the random intercept model.

A15.4 Calibration intercepts and c-statistics obtained with development samples with EPV 5 in a population with ICC=20%

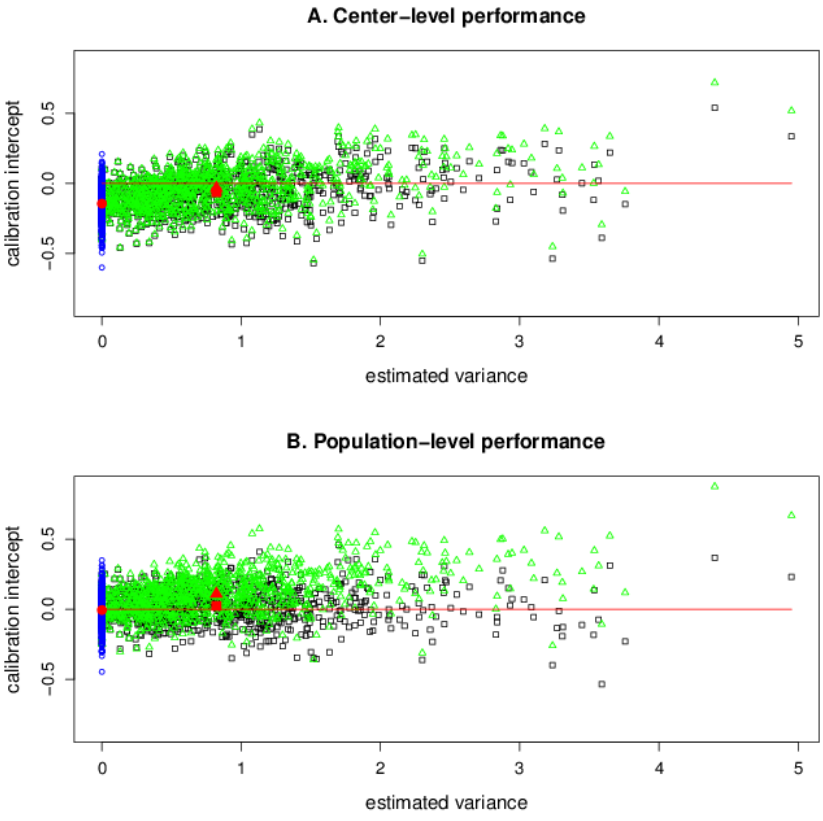


Figure A 10. Center-level (panel A) and population-level (panel B) calibration intercepts of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model from the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 5 events per variable and true random effects variance=0.822 (ICC=20%). Small symbols indicate calibration slopes in the samples, large filled symbols indicate average calibration slopes at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.822) for the random intercept model. The horizontal line represents the ideal calibration intercept.

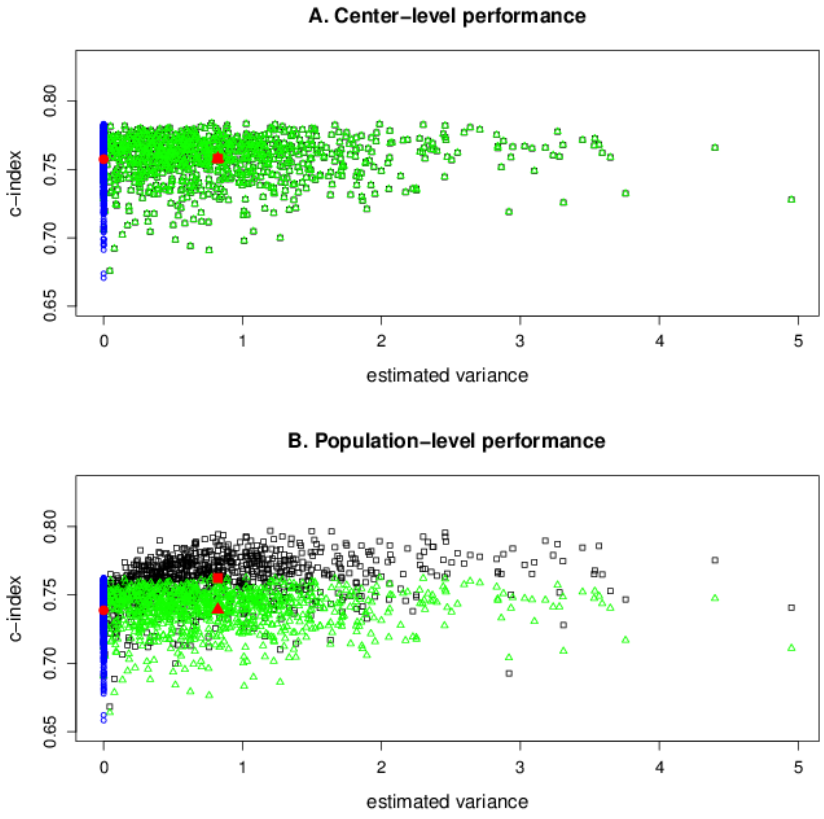


Figure A 11. Population-level (panel B) and center-level (panel A) c-statistics of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 5 events per variable and true random effects variance=0.822 (ICC=20%). Small symbols indicate c-statistics in the samples, large filled symbols indicate average c-statistics at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.822) for the random intercept model.

References

1. Hand DJ. Wonderful Examples, but Let's not Close Our Eyes. 2014: 98-100.

2. Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol*. 2016; 214: 79-90.e36.

3. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: Explanation and Elaboration. *Ann Intern Med*. 2015; 162: W1-W73.

4. Schoorel EN, Vankan E, Scheepers HC, et al. Involving women in personalised decision-making on mode of delivery after caesarean section: the development and pilot testing of a patient decision aid. *BJOG*. 2014; 121: 202-9.

5. Harrell FE. *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer, 2001.

6. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer US, 2009.

7. Timmerman D, Testa AC, Bourne T, et al. Logistic Regression Model to Distinguish Between the Benign and Malignant Adnexal Mass Before Surgery: A Multicenter Study by the International Ovarian Tumor Analysis Group. *J Clin Oncol*. 2005; 23: 8794-801.

8. Van Calster B, Van Hoorde K, Valentin L, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary

metastatic tumours: prospective multicentre diagnostic study. *BMJ*. 2014; 349: g5920.

9. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989; 81: 1879-86.

10. Hand DJ. Deconstructing Statistical Questions. *J R Stat Soc Ser A*. 1994; 157: 317-56.

11. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M and Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One*. 2012; 7: e32844.

12. Wyatt JC and Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ*. 1995; 311: 1539-41.

13. Vickers AJ and Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). *Urology*. 2010; 76: 1298-301.

14. Macleod MR, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. *Lancet*. 2014; 383: 101-4.

15. Kaijser J, Sayasneh A, Van Hoorde K, et al. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. *Hum Reprod Update*. 2014; 20: 449-62.

16. Peat G, Riley RD, Croft P, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. *PLoS Med*. 2014; 11: e1001671.

17. Altman DG. The time has come to register diagnostic and prognostic research. *Clin Chem*. 2014; 60: 580-2.

18. Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P and Ingelsson E. Risk Prediction Measures for Case-Cohort and Nested Case-Control Designs: An Application to Cardiovascular Disease. *Am J Epidemiol*. 2012; 175: 715-24.

19. Collins GS, Reitsma JB, Altman DG and Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015; 350: g7594.

20. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. *Ann Intern Med*. 2003; 138: W1-12.

21. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *Epidemiology*. 2007; 18: 805-35.

22. Liew SM, Doust J and Glasziou P. Cardiovascular risk scores do not account for the effect of treatment: a review. *Heart*. 2011; 97: 689-97.

23. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013; 10: e1001381.

24. Khudyakov P, Gorfine M, Zucker D and Spiegelman D. The impact of covariate measurement error on risk prediction. *Stat Med*. 2015; 34: 2353-67.

25. Carroll RJ, Delaigle A and Hall P. Nonparametric Prediction in Measurement Error Models. *Journal of the American Statistical Association*. 2009; 104: 993-1014.

26. Vach W. *Regression Models as a Tool in Medical Researc*. Boca Raton, Fl: Chapman & Hall/CRC, 2013.

27. Stiell IG and Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med*. 1999; 33: 437-47.

28. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009; 338: b2393.

29. Ambler G, Omar RZ and Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Stat Methods Med Res*. 2007; 16: 277-98.

30. Vergouwe Y, Royston P, Moons KG and Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. *J Clin Epidemiol*. 2010; 63: 205-14.

31. White IR, Royston P and Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011; 30: 377-99.

32. Moons KG, Donders RA, Stijnen T and Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006; 59: 1092-101.

33. Janssen KJ, Vergouwe Y, Donders AR, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem*. 2009; 55: 994-1001.

34. Little RJ and Rubin DB. Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*. 2002.

35. Bourne T, Wynants L, Peters M, et al. The impact of complaints procedures on the welfare, health and clinical practise of 7926 doctors in the UK: a cross-sectional survey. *BMJ Open*. 2015; 5: e006687.

36. Steyerberg EW, van der Ploeg T and Van Calster B. Risk prediction with machine learning and regression methods. *Biom J*. 2014; 56: 601-6.

37. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning*. Springer, 2009.

38. Kruppa J, Liu Y, Biau G, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. *Biometrical journal Biometrische Zeitschrift*. 2014; 56: 534-63.

39. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T and Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *Journal of biomedical informatics*. 2015; 54: 283-93.

40. 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. *Artif Intell Med*. 2009; 46: 139-54.

41. Van Holsbeke C, Van Calster B, Bourne T, et al. External Validation of Diagnostic Models to Estimate the Risk of Malignancy in Adnexal Masses. *Clin Cancer Res*. 2012; 18: 815-25.

42. Austin PC, Tu JV, Ho JE, Levy D and Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol*. 2013; 66: 398-407.

43. Tollenaar N and Van der Heijden P. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *J R Stat Soc Ser A*. 2013; 176: 565-84.

44. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med*. 2004; 66: 411-21.

45. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr. and Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making*. 2001; 21: 45-56.

46. Ambler G, Seaman S and Omar RZ. An evaluation of penalised survival methods for developing prognostic models with rare events. *Stat Med*. 2012; 31: 1150-61.

47. Nemes S, Jonasson JM, Genell A and Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med Res Methodol*. 2009; 9: 56.

48. Samworth RJ. Stein's Paradox. *Eureka*. 2012; 62: 38-41.

49. Efron B and Morris CN. *Stein's paradox in statistics*. WH Freeman, 1977.

50. Peduzzi P, Concato J, Kemper E, Holford TR and Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996; 49: 1373-9.

51. Harrell FE, Lee KL and Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996; 15: 361-87.

52. Steyerberg EW, Eijkemans MJ and Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol*. 1999; 52: 935-42.

53. Wynants L, Bouwmeester W, Moons KG, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *J Clin Epidemiol*. 2015; 68: 1406-14.

54. Ogundimu EO, Altman DG and Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol*. 2016; 76: 175-82.

55. Chatfield C. Model Uncertainty, Data Mining and Statistical Inference. *J R Stat Soc Ser A*. 1995; 158: 419-66.

56. 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 Med Res Methodol*. 2013; 13.

57. Royston P and Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*. Chichester: John Wiley & Sons, 2008.

58. Royston P, Altman DG and Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006; 25: 127-41.

59. Greenland S. Tests for interaction in epidemiologic studies: A review and a study of power. *Stat Med*. 1983; 2: 243-51.

60. Bottomley C, Van Belle V, Kirk E, Van Huffel S, Timmerman D and Bourne T. Accurate prediction of pregnancy viability by means of a simple scoring system. *Hum Reprod*. 2013; 28: 68-76.

61. Morrow DA and Cook NR. Determining decision limits for new biomarkers: clinical and statistical considerations. *Clin Chem*. 2011; 57: 1-3.

62. Steyerberg EW, Van Calster B and Pencina MJ. [Performance measures for prediction models and markers: evaluation of predictions and classifications]. *Rev Esp Cardiol*. 2011; 64: 788-94.

63. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L and Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ*. 2013; 185: E537-44.

64. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ and Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016; 74: 167-76.

65. Cox DR. Two Further Applications of a Model for Binary Regression. *Biometrika*. 1958; 45: 562-5.

66. Austin PC and Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014; 33: 517-35.

67. Hosmer DW and Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. *Stat Med*. 2002; 21: 2723-38.

68. Peek N, Arts DG, Bosman RJ, van der Voort PH and de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epidemiol*. 2007; 60: 491-501.

69. Pepe MS and Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008; 100: 978-9.

70. Vickers AJ and Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. *Semin Oncol*. 2010; 37: 31-8.

71. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ and Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest*. 2012; 42: 216-28.

72. Vickers AJ and Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making*. 2006; 26: 565-74.

73. Vickers AJ, Van Calster B and Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016; 352: i6.

74. Pauker SG and Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980; 302: 1109-17.

75. Van Calster B and Vickers AJ. Calibration of Risk Prediction Models: Impact on Decision-Analytic Performance. *Med Decis Making*. 2014.

76. Vickers AJ. Prediction models: revolutionary in principle, but do they do more good than harm? *J Clin Oncol*. 2011; 29: 2951-2.

77. Mallett S, Halligan S, Thompson M, Collins GS and Altman DG. Interpreting diagnostic accuracy studies for patient care. *BMJ*. 2012; 345.

78. Balachandran VP, Gonen M, Smith JJ and DeMatteo RP. Nomograms in oncology: more than meets the eye. *Lancet Oncol*. 2015; 16: e173-e80.

79. Baker SG. Putting Risk Prediction in Perspective: Relative Utility Curves. *J Natl Cancer Inst*. 2009; 101: 1538-42.

80. Localio AR and Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. *Ann Intern Med*. 2012; 157: 294-5.

81. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D and Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. *Med Decis Making*. 2013; 33: 490-501.

82. Collins GS and Moons KG. Comparing risk prediction models. *BMJ*. 2012; 344: e3186.

83. Pepe MS, Kerr KF, Longton G and Wang Z. Testing for improvement in prediction model performance. *Stat Med*. 2013; 32: 1467-82.

84. Cook NR and Paynter NP. Performance of reclassification statistics in comparing risk prediction models. *Biometrical journal Biometrische Zeitschrift*. 2011; 53: 237-58.

85. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr. and Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008; 27: 157-72; discussion 207-12.

86. Leening MJ, Vedder MM, Witteman JC, Pencina MJ and Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med*. 2014; 160: 122-31.

87. Pepe MS, Fan J, Feng Z, Gerds T and Hilden J. The Net Reclassification Index (NRI): A Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Statistics in Biosciences*. 2015; 7: 282-95.

88. Leening MJ, Steyerberg EW, Van Calster B, D'Agostino RB, Sr. and Pencina MJ. Net reclassification improvement and integrated discrimination

improvement require calibrated models: relevance from a marker and model perspective. *Stat Med*. 2014; 33: 3415-8.

89. Hilden J and Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med*. 2014; 33: 3405-14.

90. Justice AC, Covinsky KE and Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999; 130: 515-24.

91. Steyerberg EW, Harrell FE, Borsboom G, Eijkemans MJC, Vergouwe Y and Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001; 54: 774-81.

92. Konig IR, Malley JD, Weimar C, Diener HC and Ziegler A. Practical experiences on the necessity of external validation. *Stat Med*. 2007; 26: 5499-511.

93. Siontis GC, Tzoulaki I, Castaldi PJ and Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015; 68: 25-34.

94. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014; 14: 40.

95. Vergouwe Y, Steyerberg EW, Eijkemans MJ and Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005; 58: 475-83.

96. Collins GS, Ogundimu EO and Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016; 35: 214-26.

97. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW and Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015; 68: 279-89.

98. Vergouwe Y, Moons KGM and Steyerberg EW. External Validity of Risk Models: Use of Benchmark Values to Disentangle a Case-Mix Effect From Incorrect Coefficients. *Am J Epidemiol*. 2010; 172: 971-80.

99. Strobl AN, Vickers AJ, Van Calster B, et al. Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific cancer risk calculators. *Journal of biomedical informatics*. 2015; 56: 87-93.

100. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ and Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004; 23: 2567-86.

101. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE and Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008; 61: 76-86.

102. Ankerst DP, Koniarski T, Liang Y, et al. Updating risk prediction tools: a case study in prostate cancer. *Biometrical journal Biometrische Zeitschrift*. 2012; 54: 127-42.

103. Mallett S, Royston P, Dutton S, Waters R and Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med*. 2010; 8: 20.

104. Collins GS, Mallett S, Omar O and Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*. 2011; 9: 103.

105. Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*. 2012; 9: 1-12.

106. Collins GS. How can I validate a nomogram? Show me the model. *Ann Oncol*. 2015; 26: 1034-5.

107. Van Belle V and Van Calster B. Visualizing Risk Prediction Models. *PLoS One*. 2015; 10: e0132614.

108. Reilly BM and Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med*. 2006; 144: 201-9.

109. Moons KGM, Altman DG, Vergouwe Y and Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Br Med J*. 2009; 338: b606.

110. Stiell IG, Clement CM, Grimshaw J, et al. Implementation of the Canadian C-Spine Rule: prospective 12 centre cluster randomised trial. *BMJ*. 2009; 339: b4146.

111. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM and Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ*. 2012; 344: e686.

112. Van den Bruel A, Aertgeerts B, Bruyninckx R, Aerts M, Buntinx F and Hall H. Signs and symptoms for diagnosis of serious infections in children: a prospective study in primary care. *Br J Gen Pract*. 2007; 57: 538-46.

113. Siontis KC, Siontis GC, Contopoulos-Ioannidis DG and Ioannidis JP. Diagnostic tests often fail to lead to changes in patient outcomes. *J Clin Epidemiol*. 2014; 67: 612-21.

114. Kawamoto K, Houlihan CA, Balas EA and Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*. 2005; 330: 765.

115. McMahon MJ, Luther ER, Bowes WAJ and Olshan AF. Comparison of a Trial of Labor with an Elective Second Cesarean Section. *N Engl J Med*. 1996; 335: 689-95.

116. Flamm BL and Geiger AM. Vaginal Birth After Cesarean Delivery: An Admission Scoring System. *Obstet Gynecol*. 1997; 90: 907-10.

117. Dinsmoor MJ and Brock EL. Predicting failed trial of labor after primary cesarean delivery. *Obstet Gynecol*. 2004; 103: 282-6.

118. Ozdemir I, Yucel N and Yucel O. Rupture of the pregnant uterus: a 9-year review. *Arch Gynecol Obstet*. 2005; 272: 229-31.

119. Ofili-Yebovi D, Ben-Nagi J, Sawyer E, et al. Deficient lower-segment Cesarean section scars: prevalence and risk factors. *Ultrasound Obstet Gynecol*. 2008; 31: 72-7.

120. Bij de Vaate AJM, Brölmann HAM, van der Voet LF, van der Slikke JW, Veersema S and Huirne JAF. Ultrasound evaluation of the Cesarean scar: relation between a niche and postmenstrual spotting. *Ultrasound Obstet Gynecol*. 2011; 37: 93-9.

121. Vikhareva Osser O, Jokubkiene L and Valentin L. High prevalence of defects in Cesarean section scars at transvaginal ultrasound examination. *Ultrasound Obstet Gynecol*. 2009; 34: 90-7.

122. Wang CB, Chiu WWC, Lee CY, Sun YL, Lin YH and Tseng CJ. Cesarean scar defect: correlation between Cesarean section number, defect size, clinical symptoms and uterine position. *Ultrasound Obstet Gynecol*. 2009; 34: 85-9.

123. Menada Valenzano M, Lijoi D, Mistrangelo E, Costantini S and Ragni N. Vaginal Ultrasonographic and Hysterosonographic Evaluation of the Low Transverse Incision after Caesarean Section: Correlation with Gynaecological Symptoms. *Gynecol Obstet Invest*. 2006; 61: 216-22.

124. Regnard C, Nosbusch M, Fellemans C, et al. Cesarean section scar evaluation by saline contrast sonohysterography. *Ultrasound Obstet Gynecol*. 2004; 23: 289-92.

125. Armstrong V, Hansen WF, Van Voorhis BJ and Syrop CH. Detection of cesarean scars by transvaginal ultrasound. *Obstet Gynecol*. 2003; 101: 61-5.

126. Naji O, Daemen A, Smith A, et al. Visibility and measurement of Cesarean section scars in pregnancy: a reproducibility study. *Ultrasound Obstet Gynecol*. 2012; 40: 549-56.

127. Jastrow N, Antonelli E, Robyr R, Irion O and Boulvain M. Inter- and intraobserver variability in sonographic measurement of the lower uterine segment after a previous Cesarean section. *Ultrasound Obstet Gynecol*. 2006; 27: 420-4.

128. Jastrow N, Chaillet N, Roberge S, Morency AM, Lacasse Y and Bujold E. Sonographic lower uterine segment thickness and risk of uterine scar defect: a systematic review. *J Obstet Gynaecol Can*. 2010; 32: 321-7.

129. Martins WP, Barra DA, Gallarreta FMP, Nastri CO and Filho FM. Lower uterine segment thickness measurement in pregnant women with previous Cesarean section: reliability analysis using two- and three-dimensional transabdominal and transvaginal ultrasound. *Ultrasound Obstet Gynecol*. 2009; 33: 301-6.

130. Eden KB, McDonagh M, Denman MA, et al. New Insights on Vaginal Birth After Cesarean Can It Be Predicted? *Obstet Gynecol*. 2010; 116: 967-81.

131. Naji O, Daemen A, Smith A, et al. Changes in Cesarean section scar dimensions during pregnancy: a prospective longitudinal study. *Ultrasound Obstet Gynecol*. 2013; 41: 556-62.

132. Naji O, Abdallah Y, Bij De Vaate AJ, et al. Standardized approach for imaging and measuring Cesarean section scars using ultrasonography. *Ultrasound Obstet Gynecol*. 2012; 39: 252-9.

133. R Development Core Team. R: A language and environment for statistical computing. 2011: http://www.R-project.org/ (accessed).

134. Snijders TAB and Bosker RJ. *Multilevel analysis : an introduction to basic and advanced multilevel modeling*. 2nd ed. London: Sage, 2012.

135. Sprague S, Matta JM, Bhandari M, et al. Multicenter collaboration in observational research: improving generalizability and efficiency. *J Bone Joint Surg Am*. 2009; 91 Suppl 3: 80-6.

136. Molenberghs G and Verbeke G. *Models for discrete longitudinal data*. New York: Springer, 2005.

137. Localio AR, Berlin JA, Ten Have TR and Kimmel SE. Adjustments for center in multicenter studies: an overview. *Ann Intern Med*. 2001; 135: 112-23.

138. Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Stat Methods Med Res*. 1992; 1: 249-73.

139. Zeger SL, Liang K-Y and Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*. 1988; 44: 1049-60.

140. Kahan BC. Accounting for centre-effects in multicentre trials with a binary outcome – when, why, and how? *BMC Med Res Methodol*. 2014; 14: 1-11.

141. Kahan BC and Harhay MO. Many multicenter trials had few events per center, requiring analysis via random-effects models or GEEs. *J Clin Epidemiol*. 2015; 68: 1504-11.

142. Debray TPA, Moons KGM, Ahmed I, Koffijberg H and Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med*. 2013; 32: 3158-80.

143. Snell KI, Hua H, Debray TP, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol*. 2016; 69: 40-50.

144. Upton G and Cook I. *A Dictionary of Statistics*. 2nd ed. Oxford: Oxford University Press, 2008.

145. Browne WJ and Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*. 2006; 1: 473-514.

146. Paccagnella O. Sample Size and Accuracy of Estimates in Multilevel Models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2011; 7: 111-20.

147. Moineddin R, Matheson FI and Glazier RH. A simulation study of sample size for multilevel logistic regression models. *BMC Med Res Methodol*. 2007; 7.

148. Neuhaus JM, McCulloch CE and Boylan R. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Stat Med*. 2013; 32: 2419-29.

149. Kahan BC and Morris TP. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects? *Stat Med*. 2013; 32: 1136-49.

150. Maas CJ and Hox JJ. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*. 2004; 46: 427-40.

151. Skrondal A and Rabe-Hesketh S. Prediction in multilevel generalized linear models. *J R Stat Soc Ser A*. 2009; 172: 659-87.

152. Bouwmeester W, Twisk J, Kappen T, Klei W, Moons K and Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Med Res Methodol*. 2013; 13: 19.

153. Pavlou M, Ambler G, Seaman S and Omar RZ. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Med Res Methodol*. 2015; 15: 59.

154. Bouwmeester W. Prediction models: systematic reviews and clustered study data. Utrecht: Utrecht University, 2012.

155. Carpenter J and Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med*. 2000; 19: 1141-64.

156. Normand S-LT. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med*. 1999; 18: 321-59.

157. Higgins JP, Thompson SG and Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A, (Statistics in Society)*. 2009; 172: 137-59.

158. Higgins JP and Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002; 21: 1539-58.

159. Steyerberg EW, Eijkemans MJ, Van Houwelingen JC, Lee KL and Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. *Stat Med*. 2000; 19: 141-60.

160. Debray TP, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW and Moons KG. Incorporating published univariable associations in diagnostic and prognostic modeling. *BMC Med Res Methodol*. 2012; 12: 121.

161. Debray TP, Koffijberg H, Vergouwe Y, Moons KG and Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med*. 2012; 31: 2697-712.

162. Debray TP, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW and Moons KG. Meta-analysis and aggregation of multiple published prediction models. *Stat Med*. 2014; 33: 2341-62.

163. Timmerman D, Van Calster B, Testa A, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the

International Ovarian Tumor Analysis (IOTA) group. *Am J Obstet Gynecol*. 2016; 214: 424-37.

164. Jayson GC, Kohn EC, Kitchener HC and Ledermann JA. Ovarian cancer. *The Lancet*. 2014; 384: 1376-88.

165. Prat J. Staging classification for cancer of the ovary, fallopian tube, and peritoneum. *Int J Gynaecol Obstet*. 2014; 124: 1-5.

166. Cancer Research UK. Ovarian cancer statistics. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer#heading-Zero (accessed 31/05/2016).

167. Belgian Cancer Registry. Incidence Fact Sheet: Ovarian cancer. http://www.kankerregister.org/Incidence%20Fact%20Sheets (accessed 31/05/2016).

168. American Cancer Society. Suvrival rates for ovarian cancer, by stage. http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-survival-rates (accessed 31/05/2016).

169. Jacobs IJ, Menon U, Ryan A, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *The Lancet*. 2016; 387: 945-56.

170. National Insititute for Health and Care Excellence. Ovarian cancer: recognition and initial management. 2011.

171. Dodge JE, Covens AL, Lacchetti C, et al. Preoperative identification of a suspicious adnexal mass: A systematic review and meta-analysis. *Gynecol Oncol*. 2012; 126: 157-66.

172. Vergote I, De Brabanter J, Fyles A, et al. Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. *Lancet*. 2001; 357: 176-82.

173. Earle CC, Schrag D, Neville BA, et al. Effect of surgeon specialty on processes of care and outcomes for ovarian cancer patients. *J Natl Cancer Inst*. 2006; 98: 172-80.

174. Verleye L, Vergote I and van der Zee AG. Patterns of care in surgery for ovarian cancer in Europe. *Eur J Surg Oncol*. 2010; 36 Suppl 1: S108-14.

175. Timmerman D, Schwarzler P, Collins WP, et al. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. *Ultrasound Obstet Gynecol*. 1999; 13: 11-6.

176. Kaijser J, Bourne T, Valentin L, et al. Improving strategies for diagnosing ovarian cancer: a summary of the International Ovarian Tumor Analysis (IOTA) studies. *Ultrasound Obstet Gynecol*. 2013; 41: 9-20.

177. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C and Grudzinskas JG. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *BJOG*. 1990; 97: 922-9.

178. Timmerman D, Valentin L, Bourne TH, Collins WP, 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 Obstet Gynecol*. 2000; 16: 500-5.

179. Schenker N and Taylor JMG. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*. 1996; 22: 425-46.

180. Van Holsbeke C, Van Calster B, Testa AC, et al. Prospective Internal Validation of Mathematical Models to Predict Malignancy in Adnexal Masses: Results from the International Ovarian Tumor Analysis Study. *Clin Cancer Res*. 2009; 15: 684-91.

181. Timmerman D, Van Calster B, Testa AC, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. *Ultrasound Obstet Gynecol*. 2010; 36: 226-34.

182. Daemen A, Valentin L, Fruscio R, et al. Improving the preoperative classification of adnexal masses as benign or malignant by second-stage tests. *Ultrasound Obstet Gynecol*. 2011; 37: 100-6.

183. Testa A, Kaijser J, Wynants L, et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. *Br J Cancer*. 2014; 111: 680-8.

184. Sayasneh A, Wynants L, Preisler J, et al. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. *Br J Cancer*. 2013; 108: 2448-54.

185. Nunes N, Ambler G, Hoo WL, et al. A prospective validation of the IOTA logistic regression models (LR1 and LR2) in comparison to subjective pattern recognition for the diagnosis of ovarian cancer. *Int J Gynecol Cancer*. 2013; 23: 1583-9.

186. Nunes N, Yazbek J, Ambler G, Hoo W, Naftalin J and Jurkovic D. Prospective evaluation of the IOTA logistic regression model LR2 for the diagnosis of ovarian cancer. *Ultrasound Obstet Gynecol*. 2012; 40: 355-9.

187. Meys EM, Kaijser J, Kruitwagen RF, et al. Subjective assessment versus ultrasound models to diagnose ovarian cancer: A systematic review and meta-analysis. *Eur J Cancer*. 2016; 58: 17-29.

188. Van Calster B, Timmerman D, Valentin L, et al. Triaging women with ovarian masses for surgery: observational diagnostic study to compare RCOG guidelines with an International Ovarian Tumour Analysis (IOTA) group protocol. *BJOG*. 2012; 119: 662-71.

189. Kaijser J, Van Gorp T, Van Hoorde K, et al. 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. *Gynecol Oncol*. 2013; 129: 377-83.

190. Moore RG, McMeekin DS, Brown AK, et al. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecol Oncol*. 2009; 112: 40-6.

191. Timmerman D, Testa AC, Bourne T, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound Obstet Gynecol*. 2008; 31: 681-90.

192. Timmerman D, Ameye L, Fischerova D, et al. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. *BMJ*. 2010; 341: c6839.

193. Sayasneh A, Kaijser J, Preisler J, et al. A multicenter prospective external validation of the diagnostic performance of IOTA simple descriptors and rules to characterize ovarian masses. *Gynecol Oncol*. 2013; 130: 140-6.

194. Nunes N, Ambler G, Foo X, Naftalin J, Widschwendter M and Jurkovic D. Use of IOTA simple rules for diagnosis of ovarian cancer: meta-analysis. *Ultrasound Obstet Gynecol*. 2014; 44: 503-14.

195. Ruiz de Gauna B, Rodriguez D, Olartecoechea B, et al. Diagnostic performance of IOTA simple rules for adnexal masses classification: a comparison between two centers with different ovarian cancer prevalence. *Eur J Obstet Gynecol Reprod Biol*. 2015; 191: 10-4.

196. Peces Rama A, Llanos Llanos MC, Sanchez Ferrer ML, Alcazar Zambrano JL, Martinez Mendoza A and Nieto Diaz A. Simple descriptors and simple rules of the International Ovarian Tumor Analysis (IOTA) Group: a prospective study of combined use for the description of adnexal masses. *Eur J Obstet Gynecol Reprod Biol*. 2015; 195: 7-11.

197. Knafel A, Banas T, Nocun A, et al. The Prospective External Validation of International Ovarian Tumor Analysis (IOTA) Simple Rules in the Hands of Level I and II Examiners. *Ultraschall Med*. 2016; 37: 516-23.

198. Alcazar JL, Pascual MA, Graupera B, et al. External validation of IOTA simple descriptors and simple rules for classifying adnexal masses. *Ultrasound Obstet Gynecol*. 2016; 48: 397-402.

199. Alcazar JL, Pascual MA, Olartecoechea B, et al. IOTA simple rules for discriminating between benign and malignant adnexal masses: prospective external validation. *Ultrasound Obstet Gynecol*. 2013; 42: 467-71.

200. Hartman CA, Juliato CRT, Sarian LO, et al. Ultrasound criteria and CA 125 as predictive variables of ovarian cancer in women with adnexal tumors. *Ultrasound Obstet Gynecol*. 2012; 40: 360-6.

201. Fathallah K, Huchon C, Bats AS, et al. Validation externe des critères de Timmerman sur une série de 122 tumeurs ovariennes. *Gynécologie Obstétrique & Fertilité*. 2011; 39: 477-81.

202. Tantipalakorn CW, Chanane; Khunamornpong, Surapan; Sukpan, Kornkanok; Tongsong, Theera;. IOTA Simple Rules in Differentiating between Benign and Malignant Ovarian Tumors. *Asian Pac J Cancer Prev*. 2014; 15: 5123-6.

203. Silvestre L, Martins WP and Candido-dos-Reis FJ. Limitations of three-dimensional power Doppler angiography in preoperative evaluation of ovarian tumors. *Journal of Ovarian Research*. 2015; 8: 1-6.

204. Tinnangwattana D, Vichak-Ururote L, Tontivuthikul P, Charoenratana C, Lerthiranwong T and Tongsong T. IOTA Simple Rules in Differentiating between Benign and Malignant Adnexal Masses by Non-expert Examiners. *Asian Pac J Cancer Prev*. 2015; 16: 3835-8.

205. Ameye L, Timmerman D, Valentin L, et al. Clinically oriented three-step strategy for assessment of adnexal pathology. *Ultrasound Obstet Gynecol*. 2012; 40: 582-91.

206. Szubert S, Wojtowicz A, Moszynski R, et al. External validation of the IOTA ADNEX model performed by two independent gynecologic centers. *Gynecol Oncol*. 2016; 142: 490-5.

207. Araujo KG, Jales RM, Pereira PN, et al. Performance of the IOTA ADNEX model in the preoperative discrimination of adnexal masses in a gynecologic oncology center. *Ultrasound Obstet Gynecol*. 2016.

208. Sayasneh A, Ferrara L, De Cock B, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model: a multicentre external validation study. *Br J Cancer*. 2016; 115: 542-8.

209. Meys EM, Jeelof LS, Achten NM, et al. Estimating the risk of malignancy in adnexal masses: an external validation of the ADNEX model and comparison with other frequently used ultrasound methods. *Ultrasound Obstet Gynecol*. 2016.

210. Joyeux E, Miras T, Masquin I, Duglet PE, Astruc K and Douvier S. [Before surgery predictability of malignant ovarian tumors based on ADNEX model and its use in clinical practice]. *Gynecol Obstet Fertil*. 2016; 44: 557-64.

211. Epstein E, Van Calster B, Timmerman D and Nikman S. Subjective ultrasound assessment, the ADNEX model and ultrasound-guided tru-cut biopsy to differentiate disseminated primary ovarian cancer from metastatic non-ovarian cancer. *Ultrasound Obstet Gynecol*. 2016; 47: 110-6.

212. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A and Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol*. 2011; 64: 993-1000.

213. Vittinghoff E and McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*. 2007; 165: 710-8.

214. Senn S. Some controversies in planning and analysing multi-centre trials. *Stat Med*. 1998; 17: 1753-65; discussion 99-800.

215. Maas CJM and Hox JJ. Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2005; 1: 86-92.

216. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S and Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol*. 2004; 57: 785-94.

217. Van Oirbeek R and Lesaffre E. Assessing the predictive ability of a multilevel binary regression model. *Computational Statistics & Data Analysis*. 2012; 56: 1966-80.

218. van Klaveren D, Steyerberg E, Perel P and Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol*. 2014; 14: 5.

219. Bates D, Maechler M and Bolker B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. 2011: http://CRAN.R-project.org/package=lme4 (accessed October 7, 2014).

220. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr. and Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med*. 2000; 19: 1059-79.

221. Shrout PE and Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979; 86: 420-8.

222. McGraw KO and Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996; 1: 30-46.

223. Rahim-Williams B, Riley JL, 3rd, Williams AK and Fillingim RB. A quantitative review of ethnic group differences in experimental pain response: do biology, psychology, and culture matter? *Pain Med*. 2012; 13: 522-40.

224. Guthrie LB, Oken E, Sterne JA, et al. Ongoing monitoring of data clustering in multicenter studies. *BMC Med Res Methodol*. 2012; 12: 29.

225. Ren S, Yang S and Lai S. Intraclass correlation coefficients and bootstrap methods of hierarchical binary outcomes. *Stat Med*. 2006; 25: 3576-88.

226. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995; 57: 289-300.

227. Kahan B and Morris T. Assessing potential sources of clustering in individually randomised trials. *BMC Med Res Methodol*. 2013; 13: 58.

228. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S and Mark DB. Comparing risk-adjustment methods for provider profiling. *Stat Med*. 1997; 16: 2645-64.

229. White IR and Royston P. Imputing missing covariate values for the Cox model. *Stat Med*. 2009; 28: 1982-98.

230. Neuhaus JM, Kalbfleisch JD and Hauck WW. A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. *International Statistical Review / Revue Internationale de Statistique*. 1991; 59: 25-35.

231. Biard L, Labopin M, Chevret S and Resche-Rigon M. Investigating covariate-by-centre interaction in survival data. *Stat Methods Med Res*. 2016.

232. Van Calster B and Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making*. 2015; 35: 162-9.

233. Hilden J. Prevalence-free utility-respecting summary indices of diagnostic power do not exist. *Stat Med*. 2000; 19: 431-40.

234. Riley RD, Ahmed I, Debray TP, et al. Summarising and validating test accuracy results across multiple studies for use in clinical practice. *Stat Med*. 2015; 34: 2081-103.

235. Chu H and Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol*. 2006; 59: 1331-2; author reply 2-3.

236. Chu H, Nie L, Cole SR and Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Stat Med*. 2009; 28: 2384-99.

237. Pennells L, Kaptoge S, White IR, et al. Assessing Risk Prediction Models Using Individual Participant Data From Multiple Studies. *Am J Epidemiol*. 2014; 179: 621-32.

238. Snell K. Examining prediction model performance across multiple studies: which scale is best for meta-analysis of the C-statistic and other measures? *36th Annual Conference of the International Society for Clinical Biostatics*. Utrecht, the Netherlands2015.

239. Kerr KF, Brown MD, Zhu K and Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *J Clin Oncol*. 2016; 34: 2534-40.

240. Royston P, Parmar MK and Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Stat Med*. 2004; 23: 907-26.

241. Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. *The American statistician*. 2008; 62: 314-20.

242. Vickers A, Cronin A, Elkin E and Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak*. 2008; 8: 53.

243. Barendregt JJ, Doi SA, Lee YY, Norman RE and Vos T. Meta-analysis of prevalence. *J Epidemiol Community Health*. 2013; 67: 974-8.

244. van Klaveren D, Götz HM, Op de Coul EL, Steyerberg EW and Vergouwe Y. Prediction of Chlamydia trachomatis infection to facilitate selective screening on population and individual level: a cross-sectional study of a population-based screening programme. *Sex Transm Infect*. 2016; 92: 433-40.

245. Van Houwelingen HC and Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med*. 1995; 14: 1999-2008.

Curriculum vitae

Laure Wynants was born in Hasselt, Belgium on December 2, 1986.

She obtained a Master's degree in Sociology, magna cum laude, at the KU Leuven in 2008 and was a researcher at the KU Leuven Department of Sociology in 2008 and 2009. She obtained an advanced Master's degree in Quantitative Analysis in the Social Sciences, magna cum laude, at the HUB in Brussels in 2010.

In 2009 she enrolled in the Master of Science in Statistics program at the KU Leuven. She majored in biostatistics, wrote a thesis on the calibration of a sensor for continuous blood glucose monitoring, and graduated summa cum laude in 2011. In the summer of 2011 she did an internship at SAS, participating in the 'Summer of SAS' program. As of October 2011 she has been working in the STADIUS center for dynamical systems, signal processing and data analytics at the KU Leuven, preparing the current PhD thesis under the supervision of prof. dr. Ben Van Calster, prof. dr. ir. Sabine Van Huffel and prof. dr. Dirk Timmerman. She has been a PhD fellow from the Flemish Agency for Innovation by Science and Technology (IWT) since January 2013. Her main research interests are clinical prediction modeling, methods for diagnostic research, multicenter studies, and methods to handle clustered data.

Her work on the meta-analysis of clinical utility of models for the pre-operative diagnosis of ovarian cancer was awarded the Lee Lusted Student Prize in Quantitative Methods and Theoretical Developments at the annual meeting of the Society for Medical Decision Making in 2014. From November 2013 to January 2014 she was a visiting researcher at the Department of Public Health, Erasmus MC, Rotterdam where she investigated sample size

requirements for multicenter prediction research with prof. dr. Yvonne Vergouwe. She received a travel grant from the Flanders Research Foundation (FWO) to support this research. In 2015 she was awarded a travel grant from the FWO to present the results of an interim analysis of measurement differences between centers in the International Endometrial Tumor Analysis (IETA) study at the 25th World Congress on Ultrasound in Obstetrics and Gynecology in Montréal, Canada. Later in 2015 she was awarded a grant from the FWO to travel to the Egyptian IVF-ET center in Cairo to analyze a dataset on ICSI (intracytoplasmic sperm injection) in collaboration with prof. dr. Mona Aboulghar. In 2017 she will receive a grant from the KU Leuven Research Fund to continue her research.

List of publications

Papers in international journals

Wynants L, Vergouwe Y, Van Huffel S, Timmerman D and Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. *Stat Methods Med Res*. 2016: e-pub ahead of print.

Bourne T, Vanderhaegen J, Vranken R, **Wynants L**, et al. Doctors' experiences and their perception of the most stressful aspects of complaints processes in the UK: an analysis of qualitative survey data. *BMJ Open*. 2016; 6: e011711.

Wynants L, Collins GS and Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG*. 2016: e-pub ahead of print.

Timmerman D, Van Calster B, Testa A, Savelli L, Fischerova D, Froyman W, **Wynants L**, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis (IOTA) group. *Am J Obstet Gynecol*. 2016; 214: 424-37.

Van Calster B, Van Hoorde K, Froyman W, Kaijser J, **Wynants L**, et al. Practical guidance for applying the ADNEX model from the IOTA group to discriminate between different subtypes of adnexal tumors. *Facts Views Vis Obgyn*. 2015; 7: 32-41.

Wynants L, Bouwmeester W, Moons KG, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *J Clin Epidemiol*. 2015; 68: 8.

Bourne T, **Wynants L**, Peters M, et al. The impact of complaints procedures on the welfare, health and clinical practise of 7926 doctors in the UK: a cross-sectional survey. *BMJ Open*. 2015; 5: e006687.

Testa A, Kaijser J, **Wynants L**, et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. *Br J Cancer*. 2014; 111: 680-8.

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 Med Res Methodol*. 2013; 13.

Sharma S, Goodarzi M, **Wynants L**, Ramon H and Saeys W. Efficient use of pure component and interferent spectra in multivariate calibration. *Anal Chim Acta*. 2013; 778: 15-23.

Naji O, **Wynants L**, Smith A, et al. Does the presence of a Caesarean section scar affect implantation site and early pregnancy outcome in women attending an early pregnancy assessment unit? *Hum Reprod*. 2013; 28: 1489-96.

Sayasneh A, **Wynants L**, Preisler J, et al. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. *Br J Cancer*. 2013; 108: 2448-54.

Naji O, **Wynants L**, Smith A, et al. Predicting successful vaginal birth after Cesarean section using a model based on Cesarean scar features examined by transvaginal sonography. *Ultrasound Obstet Gynecol*. 2013; 41: 672-8.

Letters and replies in international journals

Naji O, **Wynants L**, Smith A, et al. Reply. *Ultrasound Obstet Gynecol*. 2013; 42: 123-4.

Van Calster B, **Wynants L**, Kaijser J, Van Huffel S and Timmerman D. The independent effect of tumor size in predicting ovarian malignancy. *Eur J Obstet Gynecol Reprod Biol*. 2012; 162: 237-8.

Conference abstracts in international journals

Wynants L, Verbakel J, Van Huffel S, Timmerman D and Van Calster B. Clinical utility of prediction models for ovarian tumor diagnosis: a decision curve analysis. *Diagnostic and Prognostic Research*. 2016; forthcoming.

Wynants L, Vergouwe Y, Van Huffel S, Timmerman D and Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. *Diagnostic and Prognostic Research*. 2016; forthcoming.

Wynants L, Installé A, Pascual M, et al. OC07.01: The International Endometrial Tumor Analysis (IETA) study: interim analysis of measurement differences between centres. *Ultrasound Obstet Gynecol*. 2015; 46: 14.

Wynants L, Van Huffel S and Van Calster B. Prediction models in multicenter studies: methodological aspects and current state of the art. *Archives of Public Health*. 2015; 73: O3.

Sayasneh A, Preisler J, **Wynants L**, et al. Diagnostic performance of IOTA Logistic Regression (LR2) model compared to the Risk of Malignancy Index to characterize adnexal masses: a multicentre prospective study. *BJOG*. 2013; 120: 372-3.

Unpublished conference contributions

Wynants L, Vergouwe Y, Van Huffel S, Timmerman D and Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. Oral presentation at the 37th Annual Conference of the International Society for Clinical Biostatistics, Birmingham, UK, 24/8/2016.

Wynants L, Verbakel J, Van Huffel S, Timmerman D and Van Calster B. Clinical utility of prediction models for ovarian tumor diagnosis: a decision curve analysis. Poster presentation at the 16th Biennial European Conference of the Society for Medical Decision Making, London, UK, 13/7/2016.

Wynants L, Van Huffel S, and Van Calster B. Evaluating the Clinical Utility of Prediction Models in a Multicenter Population using Decision-Analytic Measures. Oral presentation at the 36th Annual Conference of the International Society for Clinical Biostatistics, Utrecht, The Netherlands, 24/8/2015.

Wynants L, Bouwmeester W, Van Huffel S, Van Calster B, and Vergouwe Y. The number of events per variable needed to build logistic prediction models in clustered data: a simulation study. Oral presentation at the 35th

Annual Conference of the International Society for Clinical Biostatistics, Vienna, Austria, 28/8/ 2014.

Wynants L, Bouwmeester W, Moons KG, M M, Timmerman D, Van Huffel S, et al. The number of events per variable needed to develop risk prediction models from clustered data using multilevel modeling: a simulation study. Poster presented at Leuven Statistics Days 2014 04/12/2014; Leuven, Belgium.

Wynants L, Timmerman D, Van Huffel S, Van Calster B. Evaluating the Clinical Utility of Prediction Models in a Heterogeneous Multicenter Population Using Decision-Analytic Measures: The Random Effect Weighted Net Benefit. 36th Annual North American Meeting of the Society of Medical Decision Making; 20/10/2014; Miami, Florida 2014. Moderated poster session.

Wynants L, Bouwmeester W, Van Huffel S, Van Calster B, and Vergouwe Y. The number of events per variable needed to build logistic prediction models in clustered data: a simulation study. Oral presentation at the 35th Annual Conference of the International Society for Clinical Biostatistics, Vienna, Austria, 28/8/ 2014.

Wynants L, Timmerman D, Bourne T, Van Huffel S, and Van Calster B. Screening for Data Clustering in Multicenter Studies: The Residual Intraclass Correlation. Poster presented at the 34th Annual Conference of the International Society for Clinical Biostatistics, Munich, Germany, 26/08/2013.

Wynants L, Timmerman D, Van Huffel S, and Van Calster B. The Use of Variance Partition Coefficients to Asses Interoperator Variability in Ultrasound Features of Adnexal Masses and Clinical Characteristics of Patients. Poster presented at the Annual Meeting of the Belgian Statistical Society, Liège, Belgium 24/10/2012.

Wynants L, Timmerman D, Van Huffel S, and Van Calster B. Variable Selection for Prediction Models Based on Multicenter Data. Poster presented at the Annual Conference of the International Society of Clinical Biostatistics Bergen, Norway, 19/08/2012.

Wynants L, Timmerman D, Van Huffel S, and Van Calster B. Variable Selection for Prediction Models Based on Multicenter Data. Poster presented at the Leuven Statistics Days Leuven, Belgium, 07/06/2012.

Invited talks

Wynants L and Van Hoorde K. Biostatistics Made Very Easy for Clinicians. IOTA congress 2015, Leuven, Belgium, 17/04/2015.

Wynants L and Van Hoorde K. Biostatistics Made Easy. IOTA congress 2013, Leuven, Belgium, 27/04/2013.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING (ESAT)
STADIUS CENTER FOR DYNAMICAL SYSTEMS,
SIGNAL PROCESSING AND DATA ANALYTICS
Kasteelpark Arenberg 10 box 2446
B-3001 HEVERLEE, BELGIUM
tel. + 32 16 32 76 70
laure.wynants@kuleuven.be
www.esat.kuleuven.be/stadius/