KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Data-driven Diagnostic Decision Support Systems

A Secondary Use for Early Pregnancy and Kidney Transplant Research Databases

Thibaut Vaulet

Supervisors:
Prof. dr. ir. B. De Moor
Prof. dr. D. Timmerman

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

February 2022

Data-driven Diagnostic Decision Support Systems

A Secondary Use for Early Pregnancy and Kidney Transplant Research Databases

Thibaut VAULET

Examination committee:
Prof. dr. ir. P. Sas, chair
Prof. dr. ir. B. De Moor, supervisor
Prof. dr. D. Timmerman, supervisor
Prof. dr. ir. J. Suykens
Prof. dr. ir. J. Davis
Prof. dr. M. Naesens
Prof. dr. T. Bourne
(Imperial College, London)
Dr. M. Al-Memar
(Imperial College, London)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

February 2022

© 2022 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Thibaut Vaulet, Celestijnenlaan 200A box 2402, B-3001 Leuven (Belgium)

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

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

Acknowledgements

I would like to acknowledge the people who contributed directly or indirectly to this thesis.

First, I would like to thank my promotor Professor Bart De Moor for having welcomed me in his team of engineers and mathematicians. I am grateful for the freedom, the independence and the trust you granted me during these years at ESAT. Thanks to you and your impressive network of collaborations, I was able to work on many stimulating joint projects.

I would also like to thank my co-promotor Professor Dirk Timmerman and his team from the Department of Gynecology of UZ Leuven. Thank you for the enriching meetings and discussions over gynecology and obstetrics. You have been a crucial link in the collaboration between ESAT and the Imperial College, London which has resulted in various successful research projects. I would also like to extend a special thanks to Professor Van Calster for his valuable inputs and tips on statistical issues.

I am deeply grateful to Professor Tom Bourne, Dr. Maya Al-Memar, Dr. Hanine Fourie and their colleagues from the London Queen Charlotte's and Chelsea Hospital, and from Imperial College for the stimulating projects proposed to me. Thank you for your time and your involvement in the EPOS-related projects. Thank you for producing such a nice and extensive dataset on early pregnancy. A very special thanks to Maya and her colleague Chris for their medical assistance during a family trip in London.

I also owe a debt of gratitude to Professor Marteen Naesens and his team of collaborators from the Department of Nephrology of the UZ Leuven. Thank you for your endless support and guidance, your critical in-depth reviews of my manuscripts, your efficiency and your original research ideas. Thank you for the amazing dataset you provided me with.

I also would like to extend my sincere gratitude to Professor Suykens and

Professor Davis for their great didactical skills and for their feedback and constructive remarks during each milestone of this PhD. Their valuable suggestions have contributed to improve this manuscript. I would also like to thank Professor Sas for accepting to preside my examination committee.

I am naturally grateful to the Fonds Wetenschappelijk Onderzoek (FWO) for granting me the SB PhD scholarship.

I would like to acknowledge the support of my current and former colleagues from the STADIUS research group – Alexandre, Arnaud, Axel, Bob, Bram, Christof, Dusan, Giulia, Gorana, Hannes, Katrien, Lola, Marc, Mauricio, Melanie, Naomi, Nico, Nikhil, Oliver, Tina, Wanqiu, Willem, Xi - for providing such a pleasant work environment, for the fun chitchats and the constructive discussions. I am especially thankful to Mauricio who introduced me and guided me through the dark processes of computer aided control system design ! Thanks for your patience and your guidance through this field that was initially completely unknown to me. Thanks also to Giulia and Christof, my fellow teaching assistants of CACSD, for sharing the workload and the fun chats during the examinations periods. I would also like to give a special thanks to Dusan Popovic who introduced me to the world of machine learning.

I am also thankful to the administrative and technical staff at ESAT- Aldona, Elsy, Ida, John, Maarten, Marc and Wim. I would like to give a special shout-out for Ida, our cherished secretary, for her efficiency in solving administrative problems and her availability. Thanks a lot for all your help throughout this PhD Ida !

I am also grateful to my family and friends for their support throughout this process. A special thanks to my mum who never hesitated to come from the Ardennes to take care of my kids when we ran out of daycare solutions due to unexpected sickness or Covid-related closure of the schools. I am also thankful to my two lovely children, Shakti and Siddhartha, who, besides always cheering me up, were kind enough to understand that their dad had to work some weekends and could not go to the zoo or to the playgroundwith them.

But above all, I would like to thank my beloved wife, Vartika, who supported me from the very first day of my transition from medicine to machine learning. None of this would have been possible without your relentless support and your unconditional love. Despite your heavy schedule at the hospital, you created time out of nowhere for me and my work. You have always kept your faith in me, even when I doubted myself. I am deeply grateful to you.

Thibaut
Herent, January 2022

Abstract

With the advent of information technology in healthcare, data occupy a central place in clinical practice. This digital dependency calls for efficient tools to assist physicians in managing and processing the explosion of data resources. These tools are commonly known as *clinical decision support systems* (CDSS). They cover a broad range of applications from alerting physicians of drug allergy to the facilitation of administrative data encoding. In this thesis, we focused on *diagnostic decision support systems* (DDSS), a subtype of CDSS devoted to enhancing the diagnostic decision process with relevant information provided at the patient-level. DDSS often target a very specific medical condition. Therefore, their development largely depends on applied research centered around a predefined medical problem. In this thesis, we expose the research on DDSS initiated by three different medical challenges.

The first challenge focuses on patient similarity. Physicians often solve current cases by referring to prior encountered patients. To exploit the valuable clinical databases for similar patient retrieval, proper metrics of inter-patient similarity must first be defined. In this project, we focused on tree-based ensemble algorithms, which consist of combinations of individual decision trees, to learn task-specific patient similarity metrics. While requiring little optimization, the flexibility of these algorithms makes them suitable to learn metrics for various medical outcomes (e.g. diagnosis, prognosis or scoring systems). We demonstrated that these approaches, by natively handling the presence of missing values, provide powerful task-specific patient similarity metrics for heterogeneous clinical databases.

Our second use case targets the problem of viability prediction at the end of the first trimester of pregnancy using early gestation data. Our main contributions consist in the development and the internal validation of two calibrated models to predict the first trimester outcome based on demographics, clinical, sonographic and biomarkers predictors. We also extensively assessed the added value of plasmatic biomarkers variables in the predictive models. In addition, we

studied the imputation of incomplete data at prediction time using a modified version of fixed chained equation imputation. This online imputation framework allows to perform accurate model predictions on moderately incomplete queries. Finally, we leveraged a post-hoc interpretability framework, along with relevant visualization plots, to facilitate the understanding of the models' predictions from the clinical end-users.

In the third project, we focus on kidney transplant biopsies and explore another aspect of DDSS. While most of the existing DDSS are based on pre-established diagnostic classifications, we first aimed to improve the existing classification of acute kidney transplant rejection. The current categorization, based on the histological assessment of kidney transplant biopsies, produces categories that are not mutually exclusive and makes use of lesions that are not specific of the disease processes. While this complex classification accurately depicts the observed histological reality, its clinical interpretation remains difficult, leading to potential unstable clinical decisions. In this project, we derived and validated six novel phenotypes of acute rejection by semi-supervised consensus clustering, a special form of clustering where external guidance is added to enforce the creation of clinically meaningful clusters.

Finally, the last part of this thesis, as a synthesis of the use cases research, outlines a generic framework for data-driven DDSS. Additional theoretical considerations with regard to DDSS are exposed. In particular, we discuss healthcare-related data and their preprocessing, the choice of appropriate clinical predictive models for DDSS and important notions around clinical model interpretability.

Beknopte samenvatting

Met de komst van informatietechnologie in de gezondheidszorg nemen gegevens een centrale plaats in de klinische praktijk in. Deze digitale afhankelijkheid vraagt om efficiënte hulpmiddelen om artsen te helpen bij het beheer en de verwerking van de explosie aan gegevensbronnen. Deze hulpmiddelen zijn algemeen bekend als *klinische beslissingsondersteunende systemen* (CDSS). Ze bestrijken een brede waaier van toepassingen, gaande van het waarschuwen van artsen voor geneesmiddelenallergie tot het vergemakkelijken van het coderen van administratieve gegevens. In dit proefschrift hebben we ons gericht op *diagnostische beslissingsondersteunende systemen* (DDSS), een subtype van CDSS dat zich toelegt op het verbeteren van het diagnostische beslissingsproces met relevante informatie op patiëntniveau. DDSS zijn vaak gericht op een specifieke medische aandoening. Daarom hangt hun ontwikkeling grotendeels af van toegepast onderzoek rond een vooraf gedefinieerd medisch probleem. In deze dissertatie belichten we het onderzoek naar DDSS geïnitieerd door drie verschillende medische uitdagingen.

De eerste uitdaging richt zich op de gelijkenis van patiënten. Artsen lossen vaak actuele gevallen op door te verwijzen naar patiënten die ze eerder hebben ontmoet. Om de waardevolle klinische databanken te exploiteren voor het terugvinden van gelijkaardige patiënten, moeten eerst de juiste metrieken rond gelijkenissen tussen patiënten worden gedefinieerd. In dit project hebben we ons gericht op beslisboom-gebaseerde “ensemble” algoritmen, die bestaan uit combinaties van individuele beslisbomen, om taakspecifieke patiëntmetrieken te leren. Terwijl ze weinig optimalisatie vereisen, maakt de flexibiliteit van deze algoritmes hen geschikt om metrieken te leren voor verschillende medische uitkomsten (bv. diagnose, prognose of scoringssystemen). We toonden aan dat deze aanpak, door op een natuurlijke manier om te gaan met de aanwezigheid van ontbrekende waarden, krachtige taakspecifieke patiënt similariteitsmetrieken oplevert voor heterogene klinische databases.

Onze tweede uitdaging richt zich op het probleem van de levensvatbaarheid

voorspelling op het einde van het eerste trimester van de zwangerschap met behulp van gegevens over de vroege zwangerschap. Onze belangrijkste bijdragen bestaan uit de ontwikkeling en de interne validatie van twee gekalibreerde modellen om de uitkomst van het eerste trimester te voorspellen op basis van demografische, klinische, sonografische en biomarkervoorspellers. We hebben ook de toegevoegde waarde van plasmatische biomarkervariabelen in de voorspellende modellen uitvoerig geëvalueerd. Bovendien bestudeerden we de imputatie van onvolledige gegevens op het moment van de voorspelling, met behulp van een aangepaste versie van "fixed chained equation" imputatie. Dit online imputatiesysteem maakt het mogelijk om accurate modelvoorspellingen te doen op matig incomplete data. Tenslotte hebben we gebruik gemaakt van een post-hoc interpretatiekader, samen met relevante visualisatieplots, om het begrip van de voorspellingen van de modellen door de klinische eindgebruikers te vergemakkelijken.

In het derde project richten we ons op niertransplantatiebiopsieën en onderzoeken we een ander aspect van DDSS. Terwijl de meeste van de bestaande DDSS gebaseerd zijn op vooraf vastgestelde diagnostische classificaties, hebben we ons eerst gericht op het verbeteren van de bestaande classificatie van acute niertransplantaatafstoting. De huidige classificatie, gebaseerd op de histologische beoordeling van niertransplantatiebiopsieën, levert categorieën op die elkaar niet uitsluiten en maakt gebruik van laesies die niet specifiek zijn voor de ziekteprocessen. Hoewel deze complexe classificatie de waargenomen histologische realiteit accuraat weergeeft, blijft de klinische interpretatie moeilijk, wat kan leiden tot onstabiele klinische beslissingen. In dit project hebben we zes nieuwe fenotypes van acute afstoting afgeleid en gevalideerd door semi-gesuperviseerde consensus clustering, een speciale vorm van clustering waarbij externe sturing wordt toegevoegd om de creatie van klinisch zinvolle clusters af te dwingen.

Ten slotte schetst het laatste deel van dit proefschrift, als synthese van het use-case onderzoek, een generiek kader voor data-gedreven DDSS. Aanvullende theoretische overwegingen met betrekking tot DDSS worden belicht. In het bijzonder bespreken we gezondheidszorg-gerelateerde gegevens en hun voorbewerking, de keuze van geschikte klinische voorspellende modellen voor DDSS en belangrijke noties rond de interpreteerbaarheid van klinische modellen.

Abbreviations

α-FP	α-foeto-protein
β-hCG	β-human chorionic gonadotrophin
ABMR	antibody-mediated rejection
ARI	adjusted rand index
AUC	area under the curve
BI-SS	stability selection within bootstrap imputation
BMI	body-mass index
CA-125	cancer antigen 125
CBR	case-based reasoning
CDSS	clinical decision support system
CI	confidence interval
CRL	crown-rump length
CV	cross-validation
DCA	decision curve analysis
DDSS	diagnostic decision support system
DRMST	difference in restricted mean survival times
E2	estradiol
ECI	estimated calibration index
EHR	electronic health record
FH	fetal heartbeat
FI	feature importance
FNI	fraction of novel information
GA	gestational age

GBM	gradient boosting machine
GDPR	general data protection regulation
HLA-DSA	anti-human leukocyte antigen donor specific antibodies
HR	hazard ratio
ICU	intensive care unit
IFTA	interstitial fibrosis and tubular atrophy
k-NN	k-nearest neighbors
LASSO	least absolute shrinkage and selection operator
LMP	last menstruation period
LOOCV	leave-one-out cross-validation
MICE	multiple imputation by chained equations
MSD	mean gestational sac diameter
MYSD	mean yolk-sac diameter
PAC	proportion of ambiguous clustering
PBAC	pictorial blood assessment chart
PCA	principal component analysis
PMM	predictive mean matching
PUL	pregnancy of unknown location
PUQE score	pregnancy-unique quantification of emesis score
PUV	pregnancy of unknown viability
RMST	restricted mean survival time
SHAP	Shapley additive explanations
TCMR	T-cell mediated rejection

List of symbols

d	Number of variables, dimensions of a dataset
N	Number of patients, scans or biopsies in a dataset
k	Number of classes, neighbors or clusters
m	Number of imputed datasets
B	Number of bootstrap samples
$\boldsymbol{X}$	Matrix, tabular dataset
$\boldsymbol{X}^{(b)}$	Bootstrap sample of $\boldsymbol{X}$
$\boldsymbol{x}$	Vector, data-point (e.g. single patient data)
$\boldsymbol{y}$	Outcome variable
$\boldsymbol{z}$	Proxy or incomplete variable
$\boldsymbol{c}$	Hard-clustering partition
C	Generic performance metric
O_C	Optimism estimate of C
$d(\boldsymbol{x}, \boldsymbol{y})$	Distance between $\boldsymbol{x}$ and $\boldsymbol{y}$
$sim(\boldsymbol{x}, \boldsymbol{y})$	Similarity between $\boldsymbol{x}$ and $\boldsymbol{y}$
$\tau^{(t)}(\boldsymbol{x})$	Terminal node of instance $\boldsymbol{x}$ in tree t
$\phi(\boldsymbol{x})$	Non-linear function of $\boldsymbol{x}$
θ	Generic parameter notation
ϕ_j	Feature attribution of variable j
$\mathbb{E}[\boldsymbol{X}_j]$	Expected value of $\boldsymbol{X}_j$

Contents

Abstract iii

Beknopte samenvatting v

Abbreviations vii

List of symbols ix

Contents xi

List of Figures xv

List of Tables xix

1 Introduction 1
- 1.1 What is a clinical decision support system? . . . 2
- 1.2 CDSS taxonomy . . . 3
- 1.3 Diagnostic decision support systems . . . 5
- 1.4 Leveraging local databases for tailored DDSS . . . 6
- 1.5 From medical challenges to data-driven DDSS . . . 7
- 1.6 Additional considerations . . . 7
- 1.7 Medical challenges and objectives . . . 10
- 1.8 Structure of the thesis . . . 11
- 1.9 Data notation . . . 12

2 Flexible patient similarity metrics for incomplete data 15
- 2.1 Introduction . . . 15
- 2.2 Methodology . . . 18
  - 2.2.1 Data . . . 18
  - 2.2.2 k-nearest neighbor classifiers . . . 20
  - 2.2.3 Generic metrics . . . 22

2.2.4 Task-specific metrics . . . 22
2.2.5 Missing values . . . 25
2.2.6 Performance metrics . . . 26
2.2.7 Interpretability . . . 28
2.2.8 Software . . . 28
2.3 Results . . . 28
2.3.1 Impact of the supervision on the neighborhood composition 28
2.3.2 Local Performance . . . 29
2.3.3 Impact of the tree ensemble optimization . . . 34
2.3.4 Impact of the imputation scheme . . . 34
2.3.5 Impact of the data transformation complexity . . . 34
2.3.6 Interpretability . . . 34
2.4 Discussion . . . 37
2.5 Conclusion . . . 39

Appendices **40**
2.A Supplementary tables . . . 40
2.B Post-hoc interpretability with SHapley Additive exPlanations . 48
2.B.1 Additive feature attribution methods . . . 48
2.B.2 The Shapley values . . . 48
2.B.3 SHapley Additive exPlanations (SHAP) . . . 49
2.B.4 Properties of SHAP . . . 50
2.B.5 Interpretability at the variable-level . . . 52
2.B.6 Interpretability at the model-level . . . 52
2.C Impact of the XGBoost's main hyperparameters on the proximity matrix performance . . . 53
2.C.1 Method . . . 53
2.C.2 Results . . . 54

3 Viability prediction in the first trimester of pregnancy **61**
3.1 Problem statement . . . 61
3.2 The EPOS dataset . . . 64
3.2.1 Inclusion criteria . . . 64
3.2.2 Study Design . . . 64
3.2.3 Variables description . . . 65
3.3 Predictive models development . . . 69
3.3.1 Clinical models . . . 69
3.3.2 Augmented clinical models with blood biomarkers . . . 87
3.4 Query pre-processing . . . 100
3.4.1 Variable transformations . . . 100
3.4.2 Online imputation . . . 101
3.5 Patients search engine and case-based reasoning . . . 114
3.5.1 Methods . . . 114

3.5.2 Results . . . 116
3.5.3 Conclusion . . . 123
3.6 Interpretability and visualization . . . 124
3.6.1 Introduction . . . 124
3.6.2 Methodology . . . 125
3.6.3 Results . . . 126
3.6.4 Conclusion . . . 128
3.7 Conclusion . . . 131

Appendices **132**
3.A Supplementary tables . . . 132

4 Reclassification of kidney transplant biopsies **143**
4.1 Introduction . . . 144
4.2 Method . . . 145
4.2.1 Data . . . 145
4.2.2 Data analysis . . . 147
4.3 Results . . . 151
4.3.1 Patient and biopsy characteristics . . . 151
4.3.2 Semi-supervised clustering of rejection phenotypes . . . 151
4.3.3 Comparison of disease clusters with Banff 2019 rules . . . 159
4.3.4 Quantitative visual presentation of disease clusters . . . 159
4.3.5 Association between disease clusters and graft failure . . . 164
4.3.6 External validation . . . 167
4.4 Discussion . . . 169
4.5 Conclusion . . . 171

Appendices **173**
4.A Supplementary results . . . 173
4.A.1 Comparison of clustering vs. Banff classification for association with graft failure . . . 173
4.A.2 External validation . . . 174

5 A generic DDSS framework **191**
5.1 The databases . . . 191
5.1.1 Structured and unstructured data . . . 193
5.1.2 The main characteristics of clinical datasets . . . 195
5.2 Pre-processing . . . 197
5.2.1 Data cleaning . . . 197
5.2.2 Feature selection . . . 198
5.2.3 Feature transformation . . . 199
5.2.4 Feature extraction . . . 200
5.2.5 Missing values . . . 200

5.3 Patient similarity metric . . . 201
5.3.1 Task-specific . . . 202
5.3.2 Robust to incomplete data . . . 202
5.3.3 Adapted to heterogeneous data types . . . 202
5.3.4 Interpretable . . . 203
5.4 Clinical predictive models and inference . . . 203
5.4.1 Regression-based methods . . . 205
5.4.2 Machine learning approaches . . . 205
5.4.3 Case-based reasoning . . . 207
5.5 The interpretability component . . . 210
5.5.1 The notion of model interpretability . . . 210
5.5.2 Interpretability : transparency & explainability . . . 210
5.5.3 Post-hoc interpretability builds on explanations . . . 211
5.5.4 Explanations in DDSS . . . 212
5.5.5 Why are explanations needed ? . . . 212
5.5.6 Interpretability in machine learning: an unsolvable problem ? . . . 213
5.6 Conclusion . . . 214

6 Conclusion **215**
6.1 Contributions . . . 215
6.2 Machine learning in clinical settings: useful or not ? . . . 216
6.3 Challenges associated with DDSS . . . 217
6.4 Future works . . . 218
6.4.1 Additional validation . . . 219
6.4.2 Interpretability . . . 219
6.4.3 DDSS-physicians interactions . . . 220
6.4.4 Semi-supervised clustering and diagnostic continuums . 220

Bibliography **221**

List of publications **245**

List of Figures

1.1 An overview of the main functions of clinical decision support systems 4
1.2 A generic pipeline for the development of data-driven DDSS . 9
1.3 A visual overview of the thesis structure and the dependencies between chapters 13

2.1 Overview of the methodology to study patient similarity metrics. 19
2.2 Proportion of common instances with the neighborhood obtained with the generic SE distance on the EPOS dataset . 29
2.3 Discriminatory performance of k-NN measured by the AUC on the EPOS dataset 30
2.4 Discriminatory performance of k-NN measured by the C-index on the public breast cancer dataset 32
2.5 Discriminatory performance of k-NN measured by the C-index on the Kidney Transplant dataset 33
2.6 Example of a nearest patients search using the proximity matrix of the unoptimized XGBoost model on the EPOS dataset. . . 36
2.A.1 Discriminatory performance of k-NN measured by the IBS on the public breast cancer dataset 46
2.A.2 Discriminatory performance of k-NN measured by the IBS on the Kidney Transplant dataset 47
2.C.1 Impact of the numbers of trees (`n_estimators`) and the maximal depth of each individual tree (`depth`) - Artificial Data 55
2.C.2 Impact of the the numbers of trees (`n_estimators`) and the `subsample` parameter - Artificial Data 56
2.C.3 Impact of the `colsample` and `subsample` parameters - Artificial Data 57
2.C.4 Impact of the numbers of trees (`n_estimators`) and the maximal depth of each individual tree (`depth`) - Wisconsin dataset 58

2.C.5 Impact of the maximal depth (`depth`) and the `subsample` parameters -Wisconsin dataset 59
2.C.6 Impact of the `colsample` and `subsample` parameters - Winsconsin dataset 60

3.1 Overview of the DDSS structure for the prediction of viability in the first trimester of pregnancy 63
3.2 Main sonographic features measured in the first trimester of pregnancy 67
3.3 Flowchart of the inclusion process for the development of clinical models 71
3.4 Influx-outflux plot for missingness analysis 73
3.5 Calibration curves of the clinical models 83
3.6 Distribution of the complete model's risks stratified per outcome 83
3.7 Feature selection with BI-SS 84
3.8 Distribution of the reduced model risks stratified per outcome 85
3.9 Decision curves of the clinical models 85
3.10 Flowchart of the inclusion process for the development of augmented clinical models 88
3.11 Example of various transformations applied to the biomarker CA-125 90
3.12 Overview of the methodology to assess the added value of blood biomarkers in multivariable models. 91
3.13 Changes in predicted probabilities after including the biomarkers variables in a complete model 95
3.14 Changes in predicted probabilities after including the biomarkers variables in the complete model 96
3.15 Changes in predicted probabilities between the two reduced models 97
3.16 Changes in predicted probabilities after including the biomarkers variables in a complete model 98
3.17 Calibration curves of the augmented models 99
3.18 Decision curves of the predictive models with and without biomarkers 99
3.19 Example of online imputation with fixed chained-equations and predictive mean matching. 103
3.20 20 runs of online imputation with fixed chained-equations and predictive mean matching on missing MSD values 104
3.21 Overview of the online imputation process with fixed chained equations 106
3.22 Discrimination performance of Model 2 on various subsets of imputed data 108

3.23 Calibration performance of Model 2 on various subsets of imputed data . . . 109
3.24 Calibration curves of Model 2 after online imputation of single missing variables . . . 110
3.25 Calibration curves of Model 2 after online imputation of randomly selected combinations of variables . . . 111
3.26 Calibration curves of Model 2 after online imputation of predefined combinations of missing variables . . . 112
3.27 Overview of the leave-one-out cross-validation process . . . 117
3.28 Calibration curves of the k-NN classifiers based on different supervised distance metrics at various k . . . 120
3.29 Example of similar cases retrieval with supervised similarity metric . . . 122
3.30 SHAP feature importance of both reduced models . . . 127
3.31 Examples of individual decision plots from the reduced Model 1 129
3.32 Examples of individual decision plots from reduced Model 2 . 130

4.1 Overview of the semi-supervised consensus clustering for kidney transplant biopsies. . . . 148
4.2 Distribution of the individual acute lesion scores in the different clusters, and post-biopsy Kaplan-Meier graft survival curves relative to cluster 1. . . . 155
4.3 Visualization of the Banff classification and the 6 clusters on the whole set of biopsies . . . 157
4.4 Decision tree of the clustering process . . . 159
4.5 Comparison of indication vs. protocol biopsies, as superposed on the polar plot. . . . 160
4.6 Association with graft survival in the polar plot visualization tool. . . . 162
4.7 Association with graft survival in the polar plot visualization tool based on the validation data . . . 168
4.A.1 Distribution of the individual acute lesion scores in the clusters using an unweighted approach withcorresponding post-biopsy Kaplan-Meier graft survival curves . . . 180
4.A.2 Distribution of chronic lesions in the 6 acute lesion clusters . . 181
4.A.3 Relative distances to the closest clusters' boundary . . . 182
4.A.4 Various combinations of lesions scores displayed on the polar plots. . . . 183
4.A.5 Post-biopsy graft survival in the three DSA-/DSA+ pair of clusters. . . . 184
4.A.6 Post-biopsy graft survival in the six clusters, according to the adjustment method for repeated biopsies per patient. . . . 185
4.A.7 Comparison of cluster proportion per center . . . 186

4.A.8 Overlay of the data from Leuven and the external dataset in the polar plot, according to the six clusters identified in the derivation cohort 187
4.A.9 Distribution of the individual acute lesion scores in the different clusters, and post-biopsy Kaplan-Meier graft survival curves relative to cluster 1 of the external validation cohort 188
4.A.10 Distribution of the radius from the polar plot 189
4.A.11 Interface of the web-based application developped for the reclassification of post-transplant kidney biopsies 190

5.1 Overview of the generic DDSS structure 192
5.2 Theoretical representation of the type of clinical data collected in healthcare 194
5.3 Relationship between induction, deduction and transduction . 209

List of Tables

2.1 Summary of the variables types and missingness per dataset . 20
2.2 Summary of the different metrics characteristics 25
2.A.1 Tuning range of the XGBoost hyper-parameters 40
2.A.2 XGBoost discriminatory performance without data imputation 40
2.A.3 Proportion of common neighborhood when comparing the generic SE distance with various metrics on the EPOS dataset . 41
2.A.4 Proportion of common instances with the neighborhood from the generic SE distance on the Kidney Transplant and Breast Cancer public datasets. 42
2.A.5 Discriminative performance of the k-NN classifiers on the EPOS dataset 43
2.A.6 Discriminative performance of the k-NN classifiers on the public breast cancer dataset 44
2.A.7 Performance on the kidney transplant dataset 45

3.1 Summary of the main variables collected during the EPOS study 66
3.2 Univariate analysis of predefined variables - Clinical models . . 81
3.3 Predictive performance of the complete and reduced clinical models 82
3.4 Predictive performance of clinical models trained on consecutive visits 86
3.5 Predictive value of individual biomarkers 93
3.6 Impact of adding the biomarkers variables on the predictive performance of the clinical models 94
3.7 Performance of online imputation with fixed chained equations 107
3.8 Performance of Model 2 after online imputation on predefined variables 108
3.9 Performance of online imputation with fixed chained equations 113
3.10 List of the variables allowed in the query to search for similar cases 115

3.11 Discriminative performance of k-NN classifiers measured by AUC for various supervised metrics and queried database subsets. 118
3.12 Calibration performance of k-NN classifiers measured by ECI for various supervised metrics and queried database subsets. . 119
3.A.1 Characteristics of the dataset used for clinical model development 132
3.A.2 Characteristics of the data subset used to develop the models on consecutive scans 135
3.A.3 Characteristics of the data subset used to develop the biomarker models 139

4.1 Demographic, clinical and histological characteristics of the patients and biopsies included 152
4.2 Contingency tables comparing the Banff 2019 diagnosis and the 6 clusters derived from semi-supervised learning 161
4.3 Graft survival and RMST at 5- and 10-years post-biopsy, according to each cluster and each Banff diagnostic category . 165
4.A.1 Performance indices for a range of k values used to define the optimal number of clusters. 175
4.A.2 Distribution of biopsies among clusters and association with clinical demographics, transplant and biopsy characteristics. . 176
4.A.3 Details of cluster composition according to the individual Banff lesions scores and donor-specific HLA antibodies 177
4.A.4 Demographic, clinical and histological characteristics of the patients and biopsies included in the validation dataset 178
4.A.5 Contingency tables comparing the Banff 2019 diagnosis and the 6 clusters obtained on the external validation dataset 179

5.1 A summary of the major differences between unstructured and structured data 194
5.2 Main regression models based on the type of outcome 205

Chapter 1

Introduction

The technological developments of the last decades have drastically modified the healthcare landscape. The democratization of sensors and hardware, the reduction of data-storage costs, the improved portability and the enhancement of network infrastructures have facilitated the rapid endorsement of new technologies in healthcare. The exome sequencing is a famous example of such rapid progress. The first exome sequencing cost around half a billion US dollars and took several years to process [1]. Nowadays, next-generation sequencers can perform the sequencing of a whole exome in a few hours for a couple of hundred US dollars [2]. Rapid technological advances simplify the creation and the storage of extensive medical databases. In addition, those novel technologies frequently generate high-dimensional data output (e.g. high-resolution images, extensive omics datasets, time series of high temporal resolution). Consequently, the amount of data generated in the healthcare industry has literally exploded[1].

As a result, hospitals have now fully entered the age of *petabytes*[2]. In 2018, an average medical institution produced around 8.5 petabytes of data annually. Besides the generation of large in-house datasets, physicians are also facing another tsunami of information: the proliferation of medical literature. It has been estimated that a physician should read more than 5000 medical articles per day to keep up with the frantic publications pace [4]. Without clever computerized search engines and centralized platforms such as PubMed [5] or Cochrane [6], significant advances could never find their way to the patient bed.

[1] Between 2016 and 2018, the growth of medical data was estimated to more than 850 % by Dell EMC [3].

[2] 1 petabyte = 1,125,899,906,842,624 bytes. To give an order of idea, a petabyte represents more than 220,000 high quality DVDs.

Because of this data revolution, medicine is becoming more and more data-driven [7], even forcing medical faculties to rethink their curriculum [8]. Since medicine relies more than ever on digital information, physicians would rapidly be overwhelmed by the gargantuan masses of data generated on a daily basis without performant tools to assist them. As mentioned in Esteva [9], a single hospitalization represents on average around 150,000 pieces of data, from unstructured textual data to digitalized radiological images and lab results. In developed countries like Belgium, the vast majority of hospitals now implement electronic health record (EHR), which consist of an organized collection of patient health data, stored digitally, often shared through networks and accessed through computers. Consequently, now computers occupy a center place in doctors' daily routine and, more importantly, in their *decision-making process*[3]. The advent of computers is revolutionizing clinical practice just like the stethoscope and X-ray imagery did before. In particular, *clinical decision support systems* (CDSS), as tools to improve clinical decision-making, are now an integral part of doctors' decisional resources.

1.1 What is a clinical decision support system?

Ultimately, the main goal of a clinical decision support system is to improve healthcare delivery [11]. By augmenting a physician's knowledge with patient-tailored information and/or suggestions, a CDSS aims to facilitate complex clinical decision-making. Practically, CDSS are implemented as interactive software, often integrated within the EHR, as stand-alone programs or as web-based applications accessible through web browsers. Typically, CDSS interact with the clinical workflow by providing additional guidance, alerting on potential risks, suggesting alternatives or processing and visualizing complex patient data [12]. The outcomes of CDSS can be automated based on patient EHR data (e.g. allergy alert [13]) or triggered by physician's inputs, usually requested after manually entering additional patient information [14]. By definition, CDSS were initially intended for medical professionals as end-users. However, with the recent development of personal health records and advances in patient empowerment, some CDSS now target patients as end-users.

Although rudimentary and quite limited, the very first computerized clinical decision support systems were introduced in the 1970s [15]. By that time, they were mostly restricted to research or academic purposes, and were by no means integrated in the clinical workflow. Nowadays, CDSS are omnipresent within

[3] A recent extensive review of approximately 100 million of patient encounters demonstrated that a physician in the US spends on average more than 16 minutes on the EHR per encounter [10].

healthcare institutions and extend beyond computers, as smartphones, tablets, or even wearable technologies can easily embed CDSS.

1.2 CDSS taxonomy

CDSS can be formally classified according to various taxonomies. For instance, they can be subdivided based on the timing of an intervention, whether the CDSS is used *before*, *during* or *after* the clinical decision is taken [16]. Another common classification introduced by Berner [12] opposes *knowledge-based CDSS* to *non-knowledge-based CDSS*. The former relies on compiled medical knowledge, usually derived from validated medical studies or expert consensus, while the latter relies on advanced statistical methods or machine learning algorithms to extract relevant outputs for a specific patient. This terminological dichotomization appears, however, more as a view of mind than a real difference of implementation, as any machine learning model actually encodes some knowledge during its training phase. The difference is mainly that the knowledge is explicitly formulated in the knowledge-based CDSS, most of the time as a set of *IF-THEN* rules, whereas it is abstracted and embedded within models in non-knowledge-based CDSS.

CDSS can also be subdivided according to the dominant data modality they exploit. For instance, we can differentiate CDSS devoted to images data (CT-scan, whole slide images,...) from CDSS implementing natural language processing algorithms for free text data (e.g. physicians' notes). Other CDSS can be specifically designed for genetic data while others would deal with extensive time series data (e.g. ICU).

A more practical taxonomy, as described in Figure 1.1, consists in classifying CDSS according to the main function they endorse. Recently, Sutton et al. has listed six main functions for CDSS [14]:

1. **Patient safety**
2. **Clinical management**
3. **Cost containment**
4. **Administrative functions**
5. **Patient-facing decision support**
6. **Diagnostic support**

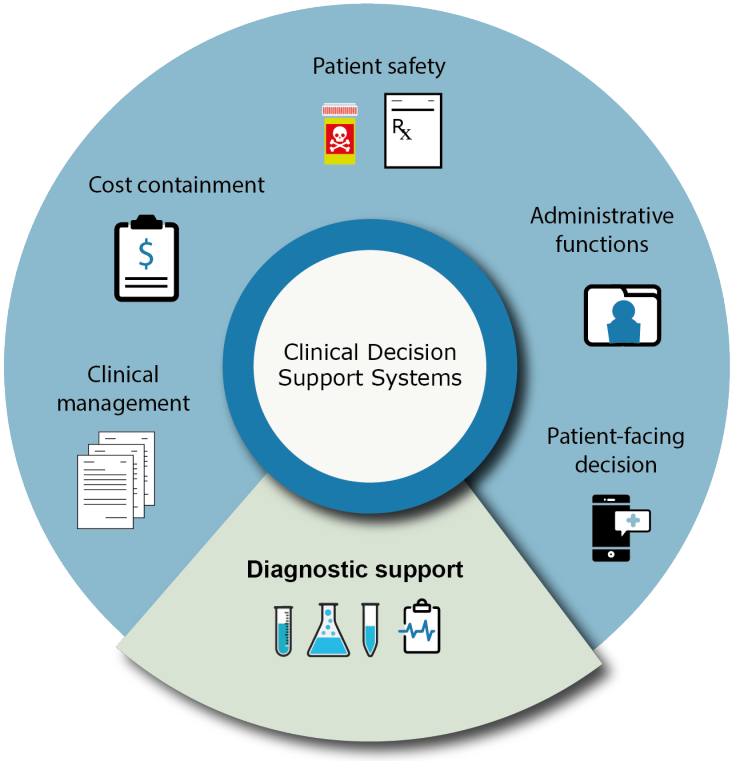


Figure 1.1: An overview of the main functions of clinical decision support systems.
Patient safety targets good prescription practice such as avoiding harmful drug-drug interactions or alerting the prescriber of a serious drug allergy. CDSS for clinical management aim for a better follow-up of the patient status and the current medical knowledge, for instance by enforcing medical guidelines [17]. CDSS can also be used to implement efficient cost-saving measures (e.g. avoiding redundant lab testing [18]). Diagnoses encoding and other administrative tasks can be facilitated with CDSS that automatically suggest relevant codes to the physician [19]. Patient-facing decision support systems target the patients as end-users, in the trend of the emerging personalized health records [14]. Diagnostic decision support systems (DDSS) tackle clinical diagnoses and prognoses by providing accurate suggestions and/or predictions at the patient-level.

Patient safety targets good prescription practice (e.g. avoiding harmful drug-drug interactions or alerting the prescriber of serious drug allergy [13]). Clinical management CDSS aims for a better follow-up of the patients' status and the current medical knowledge. They can enforce medical guidelines in the clinical practice [17]. CDSS can also be used to implement efficient cost-saving measures, for instance by decreasing patient length-of-stay or avoiding redundant lab testing [14, 18]. Tedious administrative tasks, such as diagnoses coding, can be facilitated with CDSS that automatically suggest relevant codes to the physician [19]. Patient-facing decision support systems are CDSS that directly target the patients as end-users, in the trend of the emerging personalized health records [14]. Those patient-facing CDSS can be used to monitor and communicate various medical information with medical practitioners such as treatment observance or blood-pressure monitoring [20].

1.3 Diagnostic decision support systems

CDSS that tackle clinical diagnosis are referred to as *diagnostic decision support systems* (DDSS). Those CDSS are typically tailored for a medical field (e.g. internal medicine, imagery or pathology) and often for specific diseases [12, 14]. Because DDSS must account for the particularities of a given disease, their internal knowledge is often derived from specific databases. Consequently, the composition of a DDSS is difficult to generalize over different settings without heavy re-adjustments and adaptations.

The main goal of a DDSS is to reduce the diagnostic errors by providing accurate suggestions and/or prediction at the patient-level. Those suggestions can be presented as a list of the most probable diagnoses, which results from a mapping of the patient information (e.g. symptoms, clinical signs, lab results, etc.) to the internal medical knowledge embedded in the DDSS. This internal medical knowledge can take various forms, from simple decision-rules to complex deep neural networks models.

DDSS are not restricted to diagnoses as they can also easily implement prognostic predictions. Those prognostic decision support systems can influence therapeutic behavior by providing accurate morbidity and mortality risks [21] and survival estimates, e.g. disease-free survival or graft survival [22, 23].

1.4 Leveraging local databases for tailored DDSS

While many healthcare structures are now endeavoring to centralize and organize the colossal amount of data generated internally, the vast majority of those data remain highly under-exploited for future clinical decision-making purposes.

Although the guidance provided by a DDSS is targeted towards a specific patient, most of the models and scoring systems used in CDSS are derived from one-fits-all models devloped externally on specific patient cohorts. The composition of those cohorts is dictated by the inclusion and exclusion criteria specified in research protocols. Consequently, during external deployment, any departure of the local population from the training cohort carries the risk of producing suboptimal or erroneous outputs. Unrealistic and non-representative training cohorts introduce clear limitations to the model exploitation, which should be explicitly acknowledged during model deployment.

One advantage of using in-house data and tailored models is that the specificities of the local population, e.g. the comorbidities, are directly modeled within the system, circumventing the biases from models trained on non-representative cohorts. This benefit also extends beyond the patient population, as the specificity of the local therapeutic protocols can also be included for instance[4]. The inability of a model to adjust to local behavior in problematic settings is sometimes referred to as *loss of control* and constitutes an additional obstacle to the adoption of predictive models in healthcare [24].

To overcome the limitations of general *one-fits-all* models, effort should be deployed to develop CDSS that leverage local databases of patients and implement personalized approaches, tailored to the local population. Besides developing global models locally, another solution consists in retrieving similar patients from the database to extract tailored knowledge around a given index case. Such approaches, where the knowledge from previous similar cases is used to solve a current problem, is known as case-based reasoning (CBR) [25]. The CBR framework is based on *transductive reasoning*. As previously described by Vapnik [26], transductive learning avoids the unnecessary generalization step based on costly development of population-models, to answer a specific question, here formulated at the patient-level. By inferring direct solutions from specific instances without generating a universal rule, transduction is in fact very similar to how humans, in particular physicians, reason about specific cases [27].

[4]We should note that medical knowledge developed locally might be scarce, especially in the case of rare diseases, where only a few cases are registered in the local databases. In consequence, some CDSS will still need to rely on external knowledge. Additionally, external datasets would be still required for additional validations.

A key concept to deliver personalized solutions resides in the notion of *similarity*, defining to what degree two patients are considered similar or not. The development of meaningful similarity metrics bears crucial importance to deliver accurate results. In a clinical setting, the retrieval of similar cases is often intended to retrieve patients with similar risks or similar prognoses [28]. Consequently, the patient similarity metrics should also be tailored for the medical condition of interest.

1.5 From medical challenges to data-driven DDSS

As we mentioned above, most of the DDSS are developed to tackle a specific medical condition [29]. For instance, the iBox model was developed to predict kidney allograft failure from functional, histological and immunological variables [30]. The ADNEX application [31] provides an interface to compute the risk of ovarian malignancy [32]. Kunhimangalam et al. developed a DDSS specific for the diagnosis of peripheral neuropathy [33]. A DDSS often constitutes the next step after the development and validation of a predictive model[5]. It results that the development of DDSS is largely constituted of applied research centered around a medical challenge and involving medical experts of the field in the loop. A generic pipeline for the development of data-driven DDSS is displayed in Figure 1.2.

1.6 Additional considerations

For the sake of completeness, it is worth mentioning that additional aspects remain crucial for the proper implementation of a DDSS in practice. We briefly mention them here, but those aspects will not be further developed as falling out of the scope of this thesis.

By the sensitive nature of the data collected and their poor security level [35], hospitals have become targets of choice for cyber-attacks. Ensuring the cyber-security of DDSS, for instance with high-standard encryption protocols or extensive pen testing, constitutes an additional challenge during their development. Additionally, the data collected in the healthcare industry are highly sensitive. A typical EHR encodes sensitive data such as drug addictions history, sexual orientation or termination of pregnancy history. Public disclosure of such information could have disastrous consequences on a patient's life. With

[5]Note that an accessible implementation of a predictive model can also facilitate its external validation [34])

regards to the European General Data Protection Regulation [36], healthcare-related data require additional protection safeguards and higher standards than ordinary personal data. Concepts such as privacy by design [37] or de-identification [38] should be included in DDSS to mitigate data privacy issues. Finally, the computational efficiency and the maintenance remain key components for the success and the adoption of DDSS in healthcare infrastructures. An efficient algorithm implementation ensures that DDSS solutions are reliably delivered in a reasonable amount of time. High computational costs would directly impede the clinical usage of DDSS. Therefore, DDSS should be designed to be computationally efficient, reliable and easy to maintain.

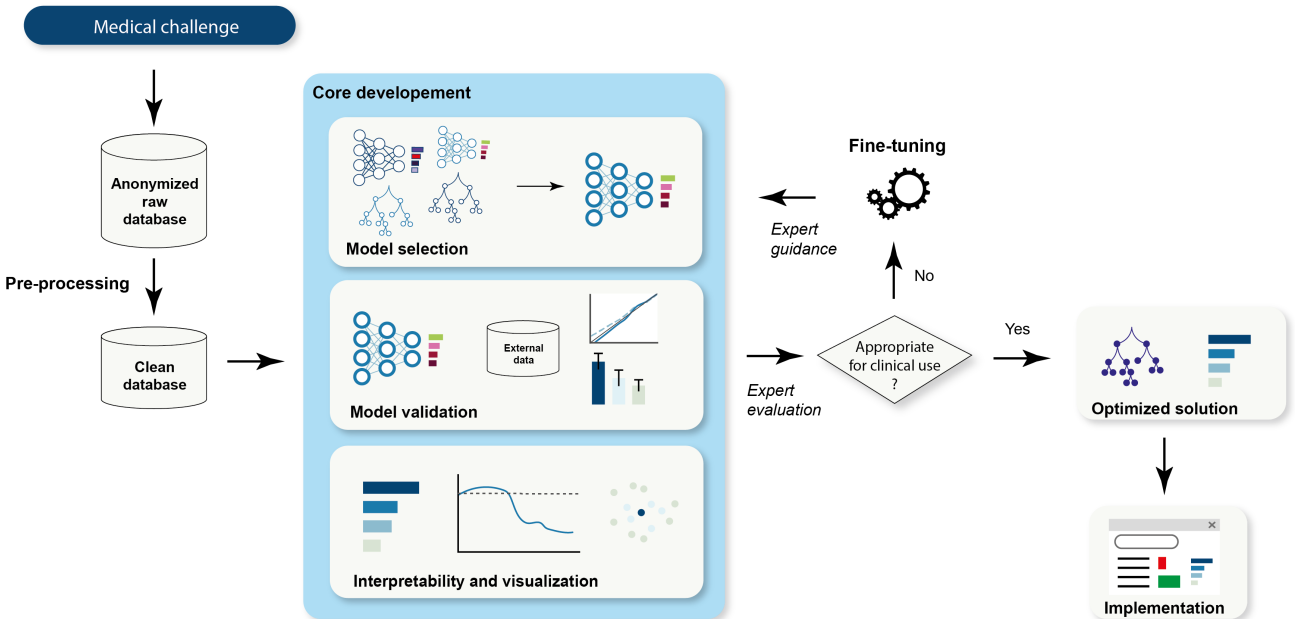


Figure 1.2: A generic pipeline for the development of data-driven DDSS.
Starting from a given medical challenge, a relevant database is first constituted and cleaned. The core development of a DDSS consists then in the selection, optimization and validation of the internal medical knowledge that will be actuated to support future diagnostic decisions. This internal knowledge is often abstracted as predictive models or algorithms that must be evaluated with the adequate performance metrics and validation schemes, depending on the medical problem at hand. Experts play a key role in the loop as they gradually refine the whole process such that the final product meets its clinical expectation. In particular, the communication of the results from a DDSS should be carefully adapted to the targeted end-users in order to maximize the adoption of the DDSS. Finally, once the internal model and interpretation/visualization strategy have been optimized for clinical use, the development of a concrete implementation with an appropriate interface can begin (e.g. web-based application).

1.7 Medical challenges and objectives

In this thesis, we expose the research on DDSS originating from three different medical challenges.

The first challenge focuses on patient similarity. Physicians often solve current cases by referring to previously encountered patients [27]. To exploit valuable extensive clinical databases for similar patient retrieval, and therefore to augment physicians' *memory* with external cases, we must first define a metric to measure inter-patients similarity. This notion defines to what degree two patients can be considered similar or not. In practice, patient similarity is highly problem-dependent, as two cases can be considered similar in one context but very different in another setting. In addition, healthcare-related data are often incomplete, which complicates the computation of a similarity score between data instances. In this project, we study task-specific patient similarity metrics for incomplete clinical databases, with a focus on tree-based ensemble proximity matrices. The aim is to produce flexible metrics that can be adapted to various medical conditions and databases. Although a similarity metric does not constitute a DDSS in its own, they represent an important component for many DDSS where two patients needs to be compared (e.g. case-based reasoning systems).

The second challenge tackles the problem of miscarriage prediction from early pregnancy data. The goal of this DDSS is to reduce the psychological distress associated with uncertain pregnancy outcomes. Populated by a large dataset of patients containing demographics, clinical, ultrasound and biomarkers variables, this DDSS exploits supervised learning methods to accurately predict the viability of the pregnancy at the end of the first trimester. Special attention is paid to the recurring problem of missing values in clinical datasets. In this project, we also study how to improve the interpretability of predictive models with state-of-the-art post-hoc interpretability methods. Finally, this DDSS also implements a case-based reasoning component to perform tailored similar patients retrieval.

Finally, the last project focuses on kidney transplant biopsies. The current categories of kidney transplant rejections are based on standardized descriptions of histological lesions. Therefore, the classification of individual transplant biopsies ignores the association with clinical outcomes such as graft failure. Driven by appropriately re-weighted histological lesions scores, this DDSS aims to reclassify kidney transplant biopsies into clinically and histologically useful categories with data-driven algorithms. In this project, semi-supervised learning with consensus clustering methods are applied on an extensive dataset of kidney transplant biopsies containing clinical and histological variables. Intuitive

visualization tools are also developed to deliver interpretable results of the reclassification process to non-experts physicians.

1.8 Structure of the thesis

This thesis is structured according to the following chapters: **Chapter 2** addresses the notion of a similarity metric in DDSS and expounds on its requirements regarding clinical databases. It presents research on the use of tree-based ensembles to derive patient similarity metrics for incomplete databases.

Chapter 3 describes the development of a DDSS for first trimester viability prediction. After introducing the problem of miscarriage prediction and the EPOS dataset, we describe the developmental steps to create calibrated predictive models, with a focus on the added values of biomarkers variables. This chapter further expounds on the development of an intelligent patient search engine to retrieve similar previous cases using task-specific similarity metrics. We also describe an online imputation method to allow prediction on incomplete data instances. Finally, we elaborate on the application of a post-hoc interpretability method to deliver individual explanations from the developed models.

Chapter 4 describes the development and the validation of a reclassification framework. Based on semi-supervised consensus clustering, this chapter illustrates how to improve the current classification of kidney transplant rejection by developing new phenotypes of kidney transplant biopsies diagnoses that are both histologically and clinically relevant. This chapter also describes the development and application of an intuitive visualization tool to assess both the severity and the new phenotypes on a single plot.

As a synthesis of the use cases research, **Chapter 5** outlines a generic structure for data-driven DDSS, where key elements of a DDSS are described in details. It contains additional descriptions of healthcare-related data and their specificities concerning the pre-processing. This chapter also examines some useful clinical predictive models for DDSS and details the case-based reasoning scheme. Finally, the concepts and challenges around model interpretability are developed in relation to DDSS. To conclude, **Chapter 6** recaps the main contributions described in this work, elaborates on the usefulness of machine learning methods for predictive modelling, discusses some additional challenges associated with DDSS and briefly outlines appealing avenues for future researches.

A visual overview of the thesis structure is displayed in Figure 1.3.

1.9 Data notation

The following data notation of structured tabular datasets is adopted throughout the thesis : a data entry is represented by a vector $\boldsymbol{x} = (x_1, x_2, x_3, ..., x_d) \in \mathbb{R}^d$ where x_j corresponds to the j^{th} variable of $\boldsymbol{x}$ [6]. Due to the structured nature of clinical datasets, it is assumed that the set of variables follows preset specifications in terms of unit and format. Cross-sectional datasets are denoted as $\mathbf{X} = [\boldsymbol{x_1}; \boldsymbol{x_2}; \boldsymbol{x_3}; ...\boldsymbol{x_N}]^\top \in \mathbb{R}^{N \times d}$, a structured tabular dataset of dimensions $N \times d$, where N is the total number of patients and d is the total number of variables registered in the dataset. The notation $\boldsymbol{X}_j$ corresponds to the column vector of the j^{th} variable from the whole set of data. $\boldsymbol{X}_{i,j}$ corresponds to the value of the j^{th} variable of the i^{th} patient.

In case of longitudinal datasets with repeated measurements, a temporal dimension t is added to the dataset. Therefore, $\boldsymbol{X}_{i,j}^{(t)}$ refers to the j^{th} variable of the i^{th} patient at time t, where t can be defined following different conventions (e.g. a date, a pre-specified visit, a time interval,...).

The variable $\boldsymbol{y}$ is a specific variable of $\boldsymbol{X}$ and denotes an *outcome variable*. Typical medical outcome variables consist in *binary outcomes* (e.g. presence or not of a diagnosis, mortality at 5 years,...), *categorical outcomes* (e.g. specific diagnoses or tumor types), *survival information* (e.g. time of kidney graft failure) or *numerical outcomes* (e.g. estimated glomerular filtration rate or eGFR). Depending on the problem at hand, a given variable $\boldsymbol{X}_j$ can be considered as a predictive variable or as an outcome variable $\boldsymbol{y}$.

[6] To fit this notation, categorical variables are one hot encoded : a variable with k different categories is encoded as $k-1$ binary vectors

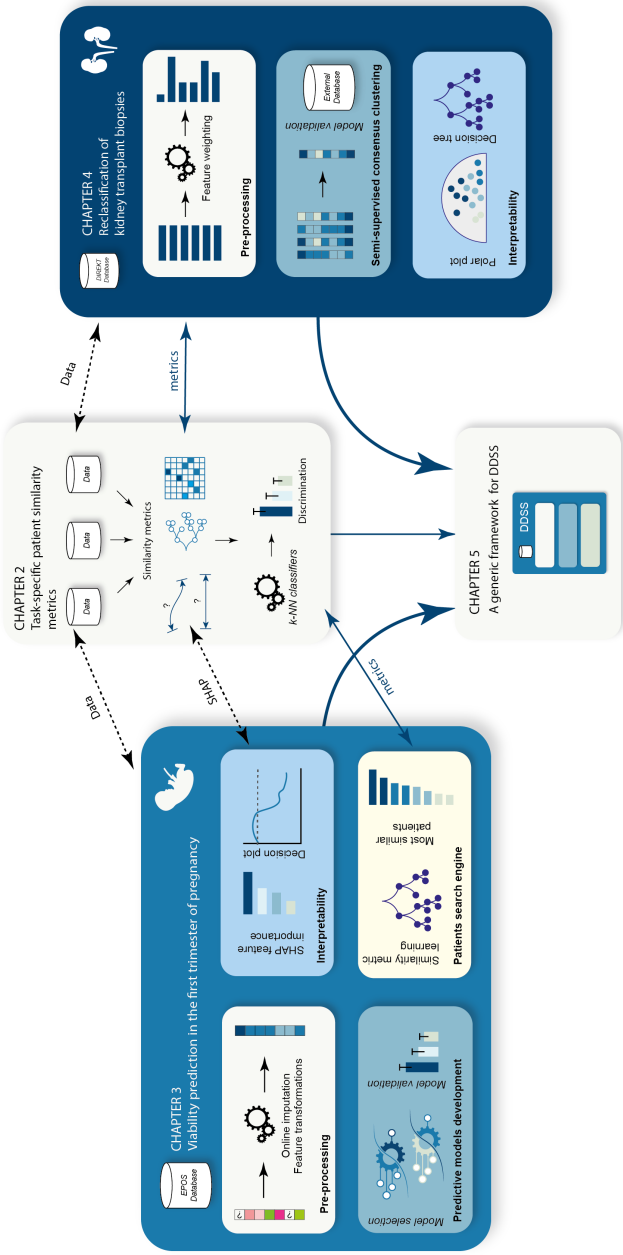


Figure 1.3: A visual overview of the thesis structure and the dependencies between chapters. Chapter 2 describes the research on task-specific patient similarity metrics based on tree-based ensembles to perform patient retrieval on incomplete databases. Chapter 3 explores the development of a DDSS to predict first trimester viability. Beside relevant pre-processing steps, this chapter builds on clinical predictive modelling and post-hoc interpretability framework to deliver explainable individual predictions, even with incomplete queries. Chapter 4 describes the research on the reclassification of kidney transplant biopsies based on data-driven clustering algorithms and extensive research databases. Chapter 5 outlines a generic DDSS structure based on the applied use-case researches and expounds on key DDSS components. Chapter 2 shares common datasets with Chapter 3 and 4. The notion of patient similarity metric is omnipresent across the chapters, from a weighted Euclidean distance in Chapter 4 to the patient proximity matrix of Chapters 3 and 4.

Chapter 2

Flexible patient similarity metrics for incomplete data

This chapter has been submitted to the *Journal of Biomedical Informatics* as: Vaulet, T., Al-Memar, M., Fourie, H., *et al. Retrieving Similar Patients from Incomplete Databases with Tree-Based Ensembles*

Objective

Develop a task-specific patient similarity metric to retrieve similar cases on incomplete datasets.

2.1 Introduction

Clinical decision-making is the core intellectual process of physicians. Although complex and not completely understood, it has been demonstrated that prior examples play a crucial role in clinical reasoning [27], as physicians often solve current cases by referring to their own previous experience. This process assumes a perfect long-term memory with unlimited storage and optimal recall. However, the human memory is full of cognitive biases [40] which renders the retrieval of information inconsistent.

To overcome these limitations, individual physician's memory can be augmented by extensive clinical databases where the retrieval of similar patients is deterministic and unbiased [25]. The digitalization of healthcare has facilitated the creation of valuable structured clinical databases [41], which constitute overlooked resources for insightful patient retrieval systems.

The cornerstone of these systems is the notion of *similarity*, defining to what degree two cases are considered similar or not. In a medical context, the similarity is frequently tailored for each medical condition regarding the risk of an external outcome [28] (e.g. risk of developing a disease or the estimated disease-free survival [42]). A variable can be deemed important in one setting but completely irrelevant for another medical condition. For that reason, generic metrics, e.g. Euclidean distance, which consider variables as equally relevant, often underperformed compared to task-specific distance metrics obtained with supervised metric learning [43–45].

In the vast literature on metrics learning (see Bellet et al. [46] for an extensive review), many methods have been developed to learn task-specific metrics. They encompass a large variety of algorithms with various degrees of complexity where the supervision often relies on explicit must-link/cannot link constraints formulations [46], which can be difficult to transpose to a clinical context. Additionally, many of them do not accommodate the presence of missing values. In real life, the missing values are ubiquitous in medical databases [47]. Naïve approaches to tackle this problem rely on complete cases only or simple imputation with the variables mean/median values. However, these simple approaches can discard significant portion of valuable information and/or introduces severe biases in subsequent analysis [48–50]. In addition, missingness can also appear in the query itself. It is therefore fundamental to develop metrics that accommodate missing values. Beside incomplete data, clinical repositories often gather data from various sources, which can be very heterogeneous in term of variable types. Therefore, the similarity metric should also account for the large heterogeneity of clinical variables encountered.

With regard to those considerations, we studied potential solutions emerging from decision tree ensembles. Tree-based ensembles encompass ensembles of individual trees, based on the bagging (e.g. Random Forest[51]) or the boosting (e.g. XGBoost [52], see Box) frameworks. Besides their excellent performance in various domains [53, 54], a less well-known aspect of tree-based ensembles, is their ability to provide a similarity matrix, which summarizes the degree of paths similarity between two instances amongst the whole ensemble of individual trees[55]. Given their flexibility to accommodate various data types and learning objectives (e.g. classification, regression or survival analysis), their native handling of missing values, their ease of hyper-parameters tuning and the variety of well-maintained implementations available (e.g. [52, 56]), tree-based

ensembles constitute suitable candidates to provide meaningful task-specific similarity metrics for clinical dataset.

Gradient boosted tree – XGBoost

To overcome the instability of a single decision tree, a solution consists in combining multiple trees together to form an *ensemble* or *committee* of trees. The predictions are then obtained from a consensus process over the ensemble (e.g. average or majority voting). As a result, one of the main advantages of using an ensemble of single classifiers is to lower the variance of the predictions. Most of the tree ensembles rely on one of the two following concepts: *bagging* or *boosting* [a]. In bagging, all individual trees are constructed in parallel and independently of each other. With boosting, the ensemble is developed in a greedy stage-wise fashion where each newly added tree intends to correct the short comings of the previous model.

A notable example of tree-based ensembles is the XGBoost algorithm, which is based on gradient boosting [52]. With gradient boosting, a special form of boosting, the structure of the next tree to add to the current ensemble F is determined through the optimization of a loss function L [57]. More specifically, each new tree f is fit to the negative gradients of L evaluated on $F(\boldsymbol{X})$ and given by $-g(\boldsymbol{X}) = \frac{\partial L(\boldsymbol{y},\boldsymbol{X})}{\partial F(\boldsymbol{X})}$. The ensemble F is then updated as $F = F + \rho f$, where ρ is an important parameter called the *learning rate* which controls the contribution of each new tree to the final ensemble and acts as a regularizer to further prevent overfitting.

However, converting a decision tree learning algorithm into an optimization problem is not straightforward as decision trees are not differentiable. With special loss function formulations, it is possible to optimize the construction of a new tree such that for each node split, the best split is chosen, taking into account the model complexity. Turning the loss function into a splitting criterion avoids the intractable problem of constructing all possible trees at each iteration.

Additional stochasticity can be introduced by a subsampling parameter which controls the amount of original data used to fit a single learner and leads to what is called *stochastic gradient boosting* [58]. It can force the algorithm to construct a tree using, for instance, only 50% of the original data to prevent overfitting.

[a] Note that bagging and boosting are general concepts and are not limited to tree ensembles.

Curiously, despites its advantages, the proximity matrix has rarely been exploited

in clinical-related literature. For instance, the study of Gray et al. [59] exploited proximity matrices from Random Forest algorithm to train a multi-dimensional scaling algorithm for Alzheimer's disease diagnosis. They demonstrated that proximity matrices provided consistent pairwise similarity measures for multiple modalities.

In this paper, we compare the performance of several supervised and interpretable approaches to retrieve similar experienced examples to new cases in the presence of missing values at both the query and the database levels. We highlight the advantages of tree-based ensemble algorithms as a flexible supervised metric learning approach, particularly suited for clinical datasets. Finally, we propose a simple modification of the post-hoc interpretation SHAP framework based on Shapley values to derive explanations for non-linear similarity metrics.

2.2 Methodology

An overview of the methodology to study patient similarity metrics is represented in Figure 2.1.

2.2.1 Data

The subsequent analysis are based on three extensive datasets, which reflect the reality of clinical data in terms of dimensionality, missingness and variables heterogeneity. The first dataset consists in public de-identified clinical and histological data on breast cancer with survival outcome [60], obtained from the CBIO portal [61, 62]. The two other datasets are proprietary data and consist in 1.) the EPOS dataset, an extensive dataset on first trimester viability, prospectively collected at the Queen Charlotte's & Chelsea Hospital, London, between March 2014 and May 2019. These data combine demographics, clinical information and ultrasound data. The main outcome is the viability of the pregnancy at the end of the first trimester; and 2.) the Kidney Transplant dataset, an extensive dataset on kidney transplant biopsies, performed at the UZ Leuven Hospital, Belgium, between March 2004 and February 2015. This dataset combines histological data with clinical information and markers of the kidney function. The main outcome is the graft survival. Additional information on those two datasets can be obtained in [63] and [64], respectively. A summary of the datasets used in the present study is displayed in Table 2.1. Continuous variables were initially scaled to the unit interval with a *min-max* transformation.

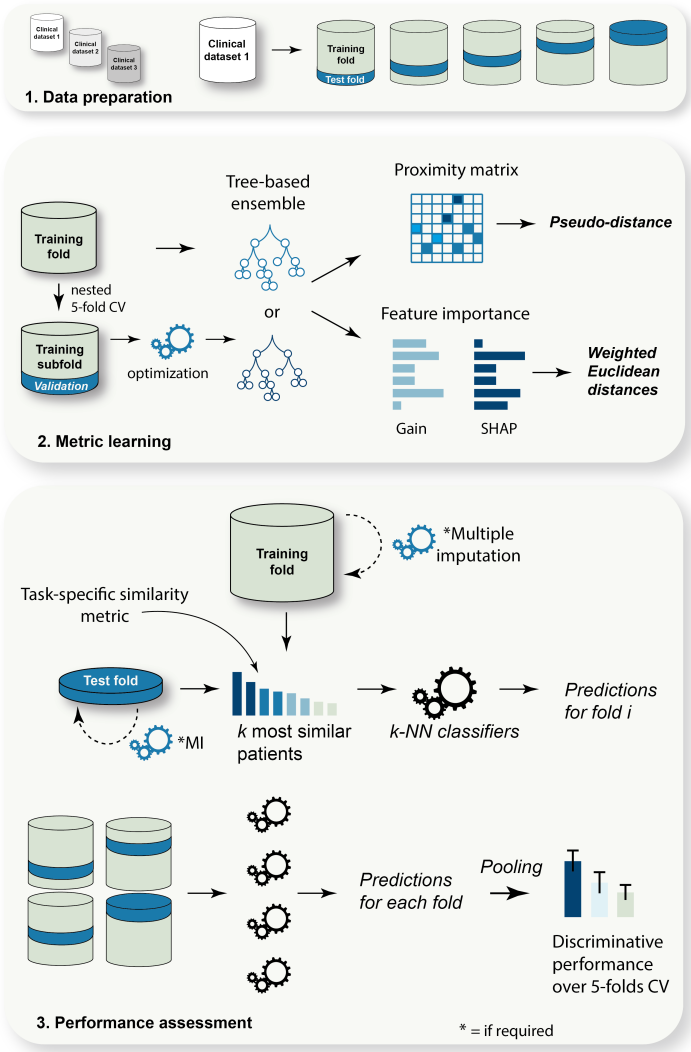


Figure 2.1: Overview of the methodology to study patient similarity metrics. The whole process is divided into three main steps. 1. *Data preparation*: each of the three clinical datasets is first divided into 5 folds for the stratified cross-validation. 2. *Metric learning*: a tree-based ensemble (here XGBoost) is trained on the training folds. The tree-based ensemble is either optimized or set to pre-defined default values. From the trained model, the variable weights based on feature importance measures and the proximity matrix are extracted to produce (pseudo-) distance metrics. 3. *Performance assessment*: For each instance of the test fold, the k most similar patients are retrieved from corresponding training folds based on the different distance metrics. Missing data are imputed with multiple imputation (m=20) for Euclidean distances-based approaches. The discrimination performance of the k-NN classifiers is assessed over the 5 test folds of the cross-validation and averaged to report the final performance estimate.

Table 2.1: Summary of the variable types and missingness per dataset

	Public Breast Cancer	**EPOS**	**Kidney Transplant**
Objective for supervision	Survival	Classification	Survival
Dimensions*			
N	1981	3419	3622
N patients	1981	1272	949
d	25	25	35
d continuous	6	18	9
d categorical	5	1	1
d ordinal	4	0	18
d binary	10	6	7
Missingness			
Missingness per variable (mean)	0.0510	0.0453	0.0111
Range	0.000-0.260	0.000-0.163	0.000-0.215

* N= number of instances, d= number of variables

Ordinal variables were transformed into numerical variables and categorical variables were one-hot encoded.

2.2.2 k-nearest neighbor classifiers

For each metric, we computed the nearest neighborhood for a size $k \in \{5, 10, 20, 50\}$[1]. The data were analyzed in a 5-folds cross-validation strategy, where the metric was learned on the training folds and evaluated on the test fold. In case of repeated data at the patient-level, all the data-points from the same patient were strictly allocated to either the training folds or the test fold to avoid data leakage.

For classification tasks, the binary outcomes retrieved from the corresponding k-NN were averaged to produce local risk estimates. For the survival tasks, similarly to the approach used by Kroupa et al. [65], a local Kaplan-Meier estimator was computed based on the k-nearest patients at various predefined time points (1, 5 and 10 years). If the same patient was retrieved several times in case of repeated measurements, only the nearest data-point per patient was used to compute the local Kaplan-Meier estimator.

[1]Note that k was not optimized as the main goal of a patient retrieval engine is often to retrieve a fixed k number of patients and not to produce the best k-NN estimator.

Distance and similarity metrics

Formally, a distant metric d is a function $d : \mathbf{X} \times \mathbf{X} \longmapsto [0, +\infty)$ that satisfied the following axioms:

- Identity
$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbf{X}, d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$$
- Symmetry
$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbf{X}, d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$$
- Triangle inequality
$$\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbf{X}, d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$$

The most famous example is the Euclidean distance (Eq. 2.1).

$$d_{Eucl.}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{d} (\boldsymbol{x}_j - \boldsymbol{x}_j)^2} \tag{2.1}$$

Distances that are not distinguishable, i.e. $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ even if $\boldsymbol{x} \neq \boldsymbol{y}$, are referred as *pseudo-metric*.
The *similarity metric*, sometimes known as *proximity measure*, is a concept that reverses the distance relationship between a pair of data points. A similarity is often expressed in a closed interval, e.g. [0,1]. A famous example is the *cosine similarity* (Eq. 2.2), which measures the cosine of the angle formed by two non-zero vectors in a multi-dimensional space. This similarity metric is often used in natural language processing to assess the degree of similarity between two documents based on their textual contents[66].

$$sim_{cos}(\boldsymbol{x}, \boldsymbol{y}) = cos(\theta) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{\sum_{j=1}^{d} \boldsymbol{x}_j \boldsymbol{y}_j}{\sqrt{\sum_{j=1}^{d} \boldsymbol{x}_j^2} \sqrt{\sum_{j=1}^{d} \boldsymbol{y}_j^2}} \tag{2.2}$$

Distance and *similarity* are often used interchangeably, as similarity can be derived from a distance and inversely. Common monotonic functions that perform the distance-similarity mapping include : $1/(1 + d(\boldsymbol{x}, \boldsymbol{y}))$, $e^{-d(\boldsymbol{x}, \boldsymbol{y})}$ and $K(\boldsymbol{x}, \boldsymbol{y}) = exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right)$, known as the Radial Basis Function kernel.

2.2.3 Generic metrics

We compared the performance of various distance metrics d as measure of patient similarity sim, using the following relation: $sim(\boldsymbol{x}, \boldsymbol{y}) = 1 - d(\boldsymbol{x}, \boldsymbol{y})$. As baselines, we used the Euclidean distance (Eq. 2.3) and a standardized version of the Euclidean distance (SE), where variables are weighted based on their variance (Eq. 2.4).

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{d} (\boldsymbol{x}_j - \boldsymbol{y}_j)^2} \tag{2.3}$$

$$d_{SE}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{d} \frac{1}{s_j^2} (\boldsymbol{x}_j - \boldsymbol{y}_j)^2} \tag{2.4}$$

s_j represents the standard deviation of the variable j.

2.2.4 Task-specific metrics

Because each medical condition differs in terms of relevant predictive variables and clinical outcome, we studied tailored similarity metrics from different supervised metric learning approaches.

Linear transformations

We first studied various weighted versions of the standardized Euclidean (WE) distance, which consist in linear transformations of the original Euclidean distance (Eq.2.5).

$$d_{WE}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{d} \frac{w_j}{s_j^2} (\boldsymbol{x}_j - \boldsymbol{y}_j)^2} \tag{2.5}$$

We compared different methods to estimate those weights w_j. For classification tasks, univariate variable weights were estimated by the AUC obtained from a

univariate logistic regression model (Eq. 2.6).

$$w_j^{AUC} = |\frac{AUC_j - 0.5}{0.5}| \tag{2.6}$$

Similar weights were obtained from the C-index [67] of Cox models for the survival tasks.

Alternatively, we also studied weights derived from feature importance (FI) of the tree-based ensemble. Those weights were obtained in two ways : 1.) FI based variable's gain at node split and 2.) FI based on the SHapley Additive exPlanations (SHAP) approach (see below), defined as the averaged absolute SHAP value per variable.

The calculation of weighted Euclidean distances is straightforward but it does not consider potential inter-variable correlations. To study more elaborated linear transformations, we applied the generalized Mahalanobis distance, parametrized by $\boldsymbol{M}$, a square matrix of size $d \times d$ (Eq. 2.7). Note that the weighted Euclidean distance can be seen as a special case of the generalized Mahalanobis distance where $\boldsymbol{M}$ is constrained to a diagonal matrix.

$$d_M(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \mathbf{M}^{-1} (\boldsymbol{x} - \boldsymbol{y})} \tag{2.7}$$

$\boldsymbol{M}$ was obtained with the Least Margin Nearest Neighbors (LMNN) algorithm from [43] with k=3 (not optimized). LMNN is designed for classification tasks; in consequence, it cannot be directly applied to survival outcomes.

Non-linear transformations

Proximity matrix Finally, we studied the use of tree-based ensembles, and in particular, the XGBoost algorithm [52], to deliver task-specific metrics. As described by Breiman in his seminal work on random forest [51, 55], tree-based ensembles provide proximity matrices, which indicates the path similarity between instances. The proximity matrix is a symmetric square matrix $\boldsymbol{P} \in \mathbb{R}^{N \times N}$ where $\boldsymbol{P}_{i,j}$ indicates the similarity between the i^{th} and j^{th} instances, and is computed as follows :

For each tree t of the ensemble, the $\boldsymbol{P}_{i,j}$ entry is incremented by 1 when the two corresponding instances end up in the same terminal node, given by $\tau^{(t)}(\boldsymbol{x})$. This process is repeated along the whole set of trees constituting the ensemble, and ultimately the total count is normalized to the unit interval by dividing

with the total number of trees T (Eq. 2.8). Therefore, a similarity score of one indicates that two instances systematically follow the exact same path in the whole ensemble. Note that this proximity matrix provides only a pseudo-metric, as it does not validate the zero-property axiom of the metric definition. In the present manuscript the term *metric* is used in a broad meaning.

$$sim_{prox}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{T} \sum_{t=1}^{T} I(\tau^{(t)}(\boldsymbol{x}) = \tau^{(t)}(\boldsymbol{y})), \tag{2.8}$$

where $\tau^{(t)}(\boldsymbol{x})$ is the terminal node of $\boldsymbol{x}$ in tree t. When comparing test instances to a training database, this proximity matrix is given by $\mathbf{P} \in \mathbb{R}^{\mathbf{N} \times \mathbf{N}'}$, where N' refers to the size of the test set.

We studied two approaches to derive such proximity matrices using the XGBoost algorithm [52]. First, the XGBoost hyperparameters were optimized with tree-structured parzen estimators [68] through 5-folds cross-validation on the training fold. The list of hyperparameters tuned can be found in Table 2.A.1. Because the goal is not to use the XGBoost model for classification directly, a second unoptimized version of the tree ensemble was used, with fixed parameters as reported in Table 2.A.1. The corresponding pseudo-distances were derived from the similarity matrix as $d(\boldsymbol{x}, \boldsymbol{y}) = 1 - sim(\boldsymbol{x}, \boldsymbol{y})$. Note that the proximity matrices were obtained on the original data without scaling.

SHAP Finally, we applied the SHAP interpretability framework for tree-based ensembles [54] on the optimized XGBoost model to derive individual SHAP values. The SHAP framework, as an additive feature attribution method based on the collaborative game theory, fairly decomposes individual model predictions among the set of variables (see Appendix 2.B for additional information on post-hoc interpretability and the SHAP framework). The Shapley value of each variable represents then the variable's contribution to the current prediction. Under the SHAP framework, the FI of a variable j is computed as the mean absolute SHAP value of that variable on the training fold. The SHAP FI were used to provide weights w_j for the weighted Euclidean distance.

We also computed the Euclidean distance directly on the SHAP values space as a special non-linear transformation $\phi(\boldsymbol{x})$ of the original data (Eq. 2.9), where $\phi(\boldsymbol{x}_j)$ corresponds to the SHAP value of $\boldsymbol{x}_j$, noted ϕ_j. A summary of the studied metrics and their characteristic is displayed in Table 2.2.

Table 2.2: Summary of the different metrics characteristics

Metric	Task-specific	Inter-pretable	Missing values	Input transfor-mation	Comments
Euclidean	-	✓	- [1]	$\boldsymbol{X}$	
Standardized Euclidean (SE)	-	✓	- [1]	$\boldsymbol{X}/s^2$	s^2= variance
Weighted Euclidean (WE)	✓	✓	- [1]	$\boldsymbol{wX}$	Various approaches to compute $\boldsymbol{w}$
Generalized Mahalanobis (LMNN)	✓	-	-	$\boldsymbol{MX}$	Only for classification problem
Proximity matrix from tree-based ensemble	✓	(✓) with SHAP	✓	Non-linear	Model-dependent, pseudo-distance
Euclidean distance (in SHAP values space)	✓	(✓)	✓	Non-linear	Model-dependent

[1] Incomplete queries can be proceeded using partial distance.

$$d_\phi(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{d} (\phi(\boldsymbol{x}_j) - \phi(\boldsymbol{x}_j))^2} \tag{2.9}$$

2.2.5 Missing values

Most of the previously described distance metrics $d(\boldsymbol{x}, \boldsymbol{y})$ require complete information from both $\boldsymbol{x}$ and $\boldsymbol{y}$ instances to be computed. In practice however, missing values can occur at the database- and/or the query-level. To alleviate this issue, the training folds, acting as *queried databases*, were imputed where appropriate with multiple imputation, similarly to Löw et al. [48]. Partial queries were addressed in two different ways: 1) computing the distance only on the observed variables in the query (*partial distance*) and 2) completing the query using multiple imputation based on the training data. In detail, 20 datasets were imputed with MICE [69] and the predictive mean matching method (see Box). The training folds were imputed alone to avoid data leakage during the imputation process. The corresponding test folds were imputed

jointly with the training data (*stacked imputation*) with the outcome variable deliberately set as missing [70, 71].

The aim of a patient retrieving system is to retrieve a predefined number k of similar cases. Therefore, the results of the nearest neighbors search from the various imputed datasets were pooled at the neighborhood level[2]. Specifically, we retrieved the k most frequent instances from the pooled set of neighborhoods resulting from the m imputed datasets.

Unlike the other approaches, proximity matrices can directly be computed in the presence of missing values [55]. XGBoost implements a default split for each node [52], optimized on the training data. Therefore, instances with missing values can still find their way down to a terminal node. Subsequently, SHAP values derived from XGBoost models can be obtained in the presence of missing values. The proximity matrix approaches were performed both on the original data without imputation and on the imputed datasets.

2.2.6 Performance metrics

To study the impact of the supervision during metric learning on the neighborhoods composition, we computed the proportion of common neighbors between the k-nearest neighbors obtained with a standardized Euclidean distance and the k-nearest neighbors obtained with a task-specific metric. If the same set of patients constituted both neighborhoods, resulting in a proportion of common neighbors equals to 1.00, the supervision did not affect the composition of the k-nearest neighbors. For the classification task, the AUC on the aggregated predictions from the 5-CV was used as discriminatory index for the various k-NN approaches. For the survival task, the concordance index for right-censored data (C-index)[76] and the integrated Brier score (IBS) [77] were used. AUC and the C-index both rely on the rank of the predictions and not their magnitude. They vary from 0 to 1 with 0.5 reflecting performance as good as random ordering and 1 perfect ordering of the instances. When events are rare, C-index might be problematic, as many survival intervals will result in tie. Therefore, in addition, we computed IBS, a censored version of the Brier score. Finally, although it was not the goal to produce a predictive model, the performance of the models used to learn the proximity matrices were also reported.

[2]Löw et al. used pooling on the similarities [48], i.e. before constituting the neighborhood.

Multiple imputation

Single imputation method consists in assigning a non-missing element to a missing value. Although, it is relatively easy to implement, single imputation suffer from several problems. First, the standard errors derived from single constant imputation are typically underestimated. Second, it can introduce a serious loss of variation in the imputed data. Finally, it is at risk of introducing severe biases in the analysis [72].

To overcome the major drawbacks of single imputation, Rubin developed the concept of *multiple imputation* [73], where several datasets are first imputed, analyzed in parallel then aggregated together to obtain *pooled results*. Therefore, multiple imputation operates in three steps: *imputation, analysis* and *pooling*. In the imputation phase, m datasets are imputed based on single unbiased imputation strategy (i.e. not mean or median imputation). Stochasticity can be introduced by stochastic regression or random donor draws (e.g. predictive mean matching) [74]. Once imputed, the datasets can be used with any regular statistical method requiring complete data. Each of the m imputed datasets is analyzed separately. Finally, the individual results are aggregated together in the pooling stage. Unlike single imputation, multiple imputation does take the uncertainty of imputation into account when delivering estimates. Using a special formulation of standard errors that includes the imputation variability, multiple imputation produces accurate estimates of standard errors under non-informative missingness. One of the most commonly used algorithms is the Multiple Imputation by Chained Equations (MICE)[75].

In the MICE algorithm, incomplete variables are imputed individually based on fully conditional models. Each variable benefits from its own imputation model. However, in practice, many variables often contain missing values. Therefore, incomplete variables must be imputed from other variables that are potentially incomplete. In MICE, the imputation phase follows an iterative process: the imputation starts with a random draw from the observed values to obtain a first complete dataset. The imputation models are fit to this dataset. Then, in a posterior step, new estimates of the models' parameters are sampled using previously completed dataset. This process is repeated until some convergence criteria are met or until a pre-specified number of iteration cycles. The several random draws introduce stochasticity in the final imputed dataset.

MICE uses PMM during the single imputation process for continuous values. Unlike traditional regression imputations, PMM samples real values from observed cases, i.e. *donors*, with similar predicted values from the imputed models. Using donors guarantees that the range of the imputed values remains realistic (e.g. no negative age) and better accommodates for variable heteroscedasticity. However, this approach cannot extrapolate outside of the observed values range.

2.2.7 Interpretability

The interpretation of the results from a query search is straightforward in case of Euclidean distance as both the query and the database are expressed in original units. For the weighted versions of the Euclidean distance, visual explanations of the results can be communicated through simple *trajectory plots* where the individual variables are scaled and ordered proportionally to their importance. Similarly, results based on the SHAP values space can be visually explained with *decision plots* where the variables' SHAP values are ordered by importance.

For a query search based on proximity matrices, we derived a simple approach, based on the SHAP framework [54], to explain the obtained results.

The initial SHAP framework provides individual additive explanations for a prediction y of a model $f(\boldsymbol{x})$, where $y = f(\boldsymbol{x})$. In a relative setting, we replace the function $f(\boldsymbol{x})$, by a similarity function $sim_{\boldsymbol{q}}(\boldsymbol{x})$, tailored for the query $\boldsymbol{q}$ such that $sim_{\boldsymbol{q}}(\boldsymbol{x})$ equals $sim_{prox}(\boldsymbol{x}, \boldsymbol{q})$, the value of the proximity matrix between $\boldsymbol{x}$ and $\boldsymbol{q}$ (Eq. 2.8). The obtained SHAP value ϕ_j of a data point $\boldsymbol{x}$ represents then the influence of the variable j on the total similarity score defined by $sim_{prox}(\boldsymbol{x}, \boldsymbol{q})$. The expected value $\mathbb{E}[sim_{\boldsymbol{q}}]$ represents the average similarity towards the query $\boldsymbol{q}$ from the whole database $\boldsymbol{X}$ (Eq. 2.10). In other words, the SHAP values of $\boldsymbol{x}$ explain the departure of the similarity $sim_{\boldsymbol{q}}(\boldsymbol{x})$ compared to the averaged similarity $\mathbb{E}[sim_{\boldsymbol{q}}]$.

$$\mathbb{E}[sim_{\boldsymbol{q}}] = \frac{1}{N} \sum_{x \in X} sim_{\boldsymbol{q}}(\boldsymbol{x}) \tag{2.10}$$

2.2.8 Software

The multiple imputation of the data was performed with the `mice` R package[78]. All subsequent analysis were performed with Python 3.6.6 [79].

2.3 Results

2.3.1 Impact of the supervision on the neighborhood composition

The proportion of common instances between a SE distance-based neighborhood and task-specific distance-based neighborhoods are reported in Figure 2.2 and Tables 2.A.3 and 2.A.4. These results demonstrated that metric supervision

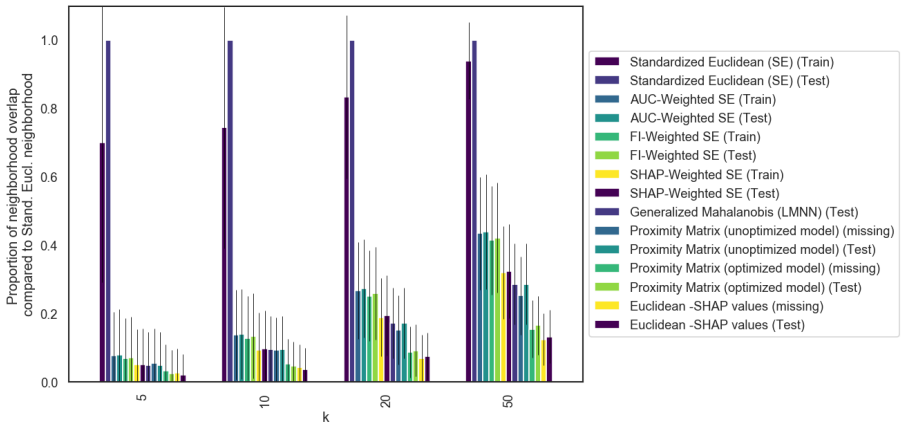


Figure 2.2: Proportion of common instances with the neighborhood obtained with the generic SE distance for various metrics on the EPOS dataset.
The imputation scheme is labeled as follows: *missing* refers to no imputation, *train* refers to imputation performed on the queried database only, i.e. the training folds, and *test* indicates that the imputation was performed on both the database and the query. For the generic reference metric, the imputation was performed on the database and the query.
All task-specific metrics produced drastically different neighborhoods compared to the neighborhood based on the standardized Euclidean distance.

drastically affects the composition of a neighborhood compared to a unsupervised approach. These results also demonstrated that the probability of a case being common to the unsupervised and the supervised nearest neighborhood increases with k, the size of the neighborhood. This is explained by the finite amount of instances composing each datasets. However, even with k as large as 50, all nearest neighborhoods computed with task-specific metrics displayed less than 50% of common instances with the neighborhood obtained with a standardized Euclidean metric. These results also reflect the degree of transformation of the original input space, where weighted Euclidean distance demonstrated a higher overlap with SE distances compared to more complex non-linear transformations, which resulted in poorer overlap in all datasets.

2.3.2 Local Performance

As expected from k-NN models, moderate increases of k improved the performance for all datasets compared to very low values of k, even though

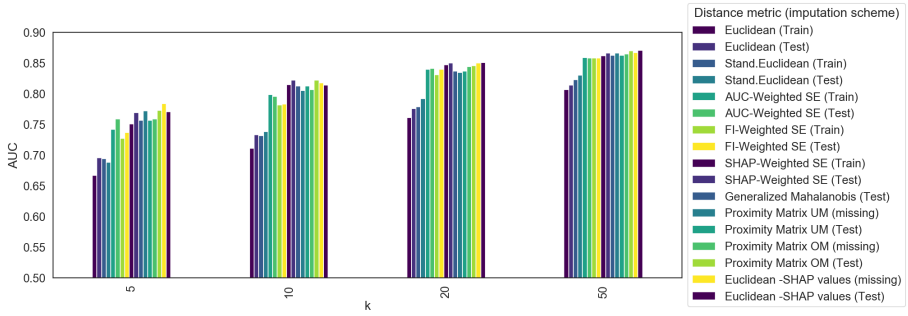


Figure 2.3: Discriminatory performance of k-NN measured by AUC on the EPOS dataset for various metric at k equals 5, 10, 20 and 50. (Averaged from a 5-folds cross-validation). The imputation scheme is labeled as follows: *missing* means that no imputation was performed, *train* means that imputation was performed on the queried database only and *test* means that imputation was performed both at the database and at the query levels.

local modelling is not the main interest of similar patients retrieval.

Binary outcome (EPOS dataset)

The discrimination performance of the k-NN classifiers for various values of k on the EPOS dataset is displayed in Figure 2.3 (the corresponding results are reported in Table 2.A.5). For a given k, the performance of the task-specific metrics is globally similar, except the WE distance with weights obtained from the model's features importance, which slightly underperformed with low values of k. The average AUC gains of the task-specific metrics approaches compared to the generic SE distance with multiple imputation, were 0.062 (+/- 0.016 SD), 0.074 (+/- 0.013), 0.066 (+/- 0.006) and 0.050 (+/- 0.004) for k equals to 5,10, 20 and 50 respectively. The results demonstrated the clear benefit of using supervised metric learning against unsupervised approaches.

Survival outcome (Breast Cancer and Kidney Transplant datasets)

The discrimination performance in terms of C-index for the Breast Cancer dataset and the Kidney Transplant dataset is displayed in Figures 2.4 and 2.5 respectively. The corresponding results are reported in Tables 2.A.6 and 2.A.7. On both datasets, for the large majority of combinations of k and time limit,

the highest C-index were achieved by the proximity matrix approach based on unoptimized XGBoost, without imputation. The average C-index gains from supervised approaches at 5 years for the Breast Cancer dataset, were 0.033 (+/- 0.021 SD), 0.040 (+/- 0.024), 0.040 (+/- 0.023) and 0.037 (+/- 0.018) for k equals to 5,10, 20 and 50 respectively. Similarly, the average C-index gains at 5 years for the Kidney Transplant dataset, were 0.037 (+/- 0.030 SD), 0.038 (+/- 0.030), 0.029 (+/- 0.027) and 0.018 (+/- 0.018) for k equals to 5, 10, 20 and 50 respectively. The performance measured by IBS for the two datasets is displayed in Figures 2.A.1 and 2.A.2, with the corresponding results reported in Tables 2.A.6 and 2.A.7. Overall, the various forms of metrics extracted from the tree-based ensemble performed similarly, except for WE distance based on the native model's feature importance, which performed notably worse than the rest. The unoptimized proximity matrix approach performed the best on both datasets in the vast majority of the k and time limit combinations.

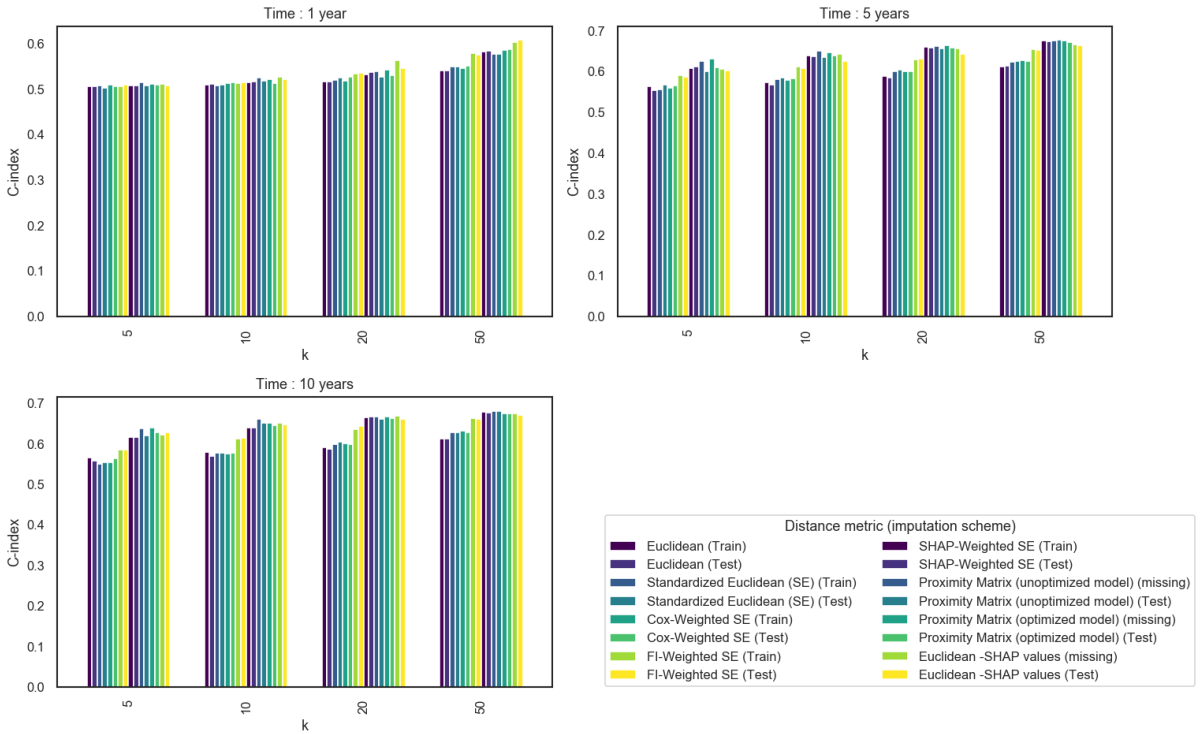


Figure 2.4: Discriminatory performance of k-NN measured by the C-index on the public breast cancer dataset for various metrics at k equals 5, 10, 20 and 50. The imputation scheme is labeled as follows: *missing* means that no imputation was performed, *train* means that imputation was performed on the queried database only and *test* means that imputation was performed both at the database and at the query levels.

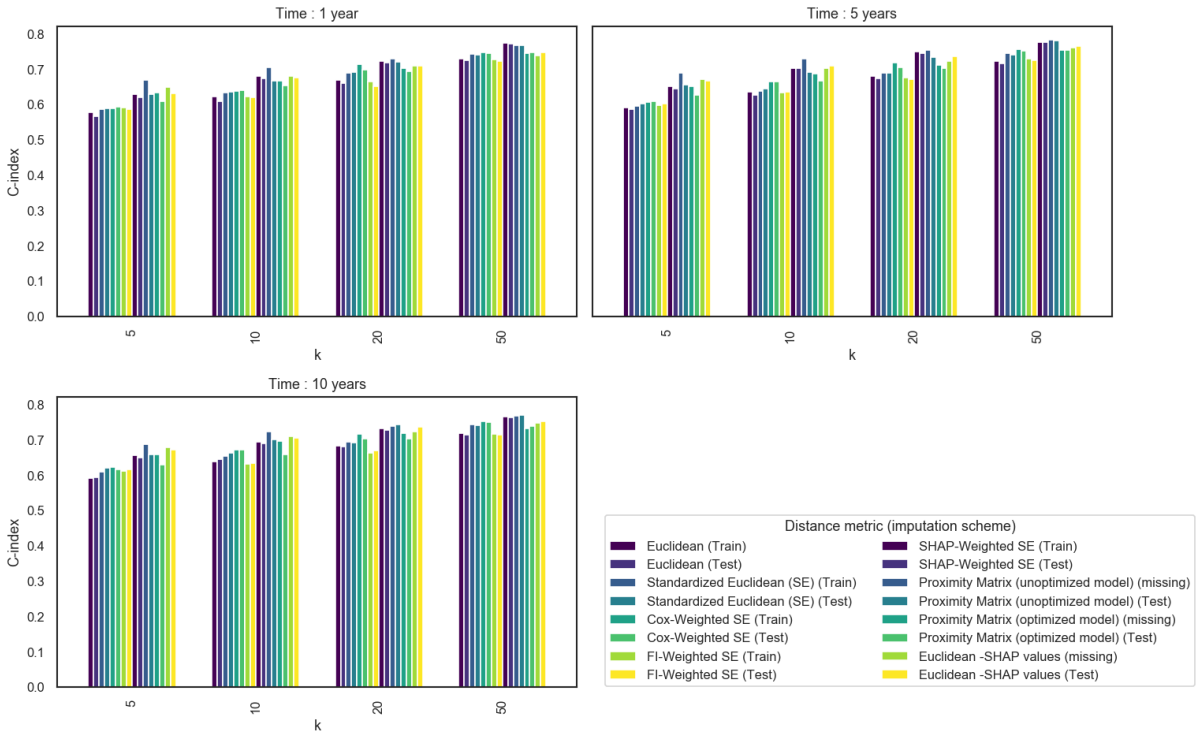


Figure 2.5: Discriminatory performance of k-NN measured by the C-index on the Kidney Transplant dataset for various metrics at k equals 5, 10, 20 and 50. The imputation scheme is labeled as follows: *missing* means that no imputation was performed, *train* means that imputation was performed on the queried database only and *test* means that imputation was performed both at the database and at the query levels.

2.3.3 Impact of the tree ensemble optimization

For all three datasets, the performance based on the proximity matrices approach was similar whether the matrices were extracted from a carefully optimized model or from an unoptimized version with fixed parameters (see additional analyses on the impact of the individual XGBoost parameters in Section 2.C in appendices).

2.3.4 Impact of the imputation scheme

Overall, completing incomplete queries with multiple imputation prior to the nearest neighbors search did not improve the results compared to partial queries based on the available variables. More importantly, the proximity matrix approaches using the original data without imputation of the query nor the database produced similar results, if not better, than the supervised approaches relying on cumbersome multiple imputation.

2.3.5 Impact of the data transformation complexity

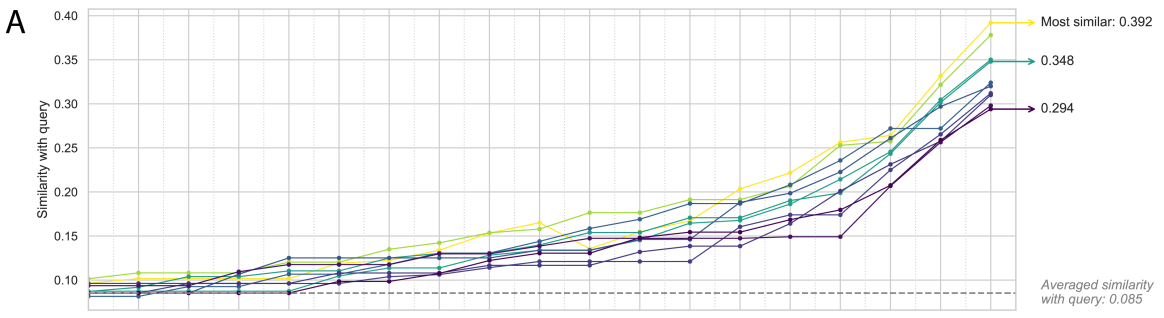
On the various datasets, complex non-linear pseudo-distances did not necessarily outperform weighted Euclidean distance with appropriate weighting schemes. In particular, the weighted Euclidean distance based on SHAP features importance performed well in the three datasets.

For comparison purposes, the performance of both unoptimized and optimized XGBoost models used to derive the proximity matrices is reported in Table 2.A.2 for each dataset. As expected, the optimization of the hyper-parameters produced slightly more discriminant ensembles.

2.3.6 Interpretability

Figure 2.6 demonstrates how the results of a nearest patients search using a proximity matrix can be translated visually. This figure displays the 10 most similar scans to a randomly draws scan (*query*) from the EPOS dataset using the proximity matrix learned on the unoptimized XGBoost model. The similarity scores are decomposed into sums of additive variable contributions expressed as SHAP values, as developed in Section 2.2.7. This decision plot displays the cumulative sum of the SHAP values for each of the 10 most similar cases. It starts from a baseline value corresponding to the average similarity between the query and the whole dataset. The SHAP values explain the gradual increase

of the similarity along the most relevant variables compared to the baseline similarity value. The corresponding retrieved cases are displayed in a data table with conditional formatting.



B

	Bleeding days	Gravida	Para	Fetal heartbeat	# previous misc.	Certainty of LMP date	GA by CRL	GA by LMP	Presence of amnion	PUQE score	MYS	MSD	Difference in GA	Maternal Age	CRL	Similarity	Outcome
Query	2	4	1	0	2	10	-	46	0	-	2.9	13.8	2.2	40	0		
1421	0	4	1	0	2	10	33.079	49	0	-	2.2	12.7	6.3	38	1.3	0.392	Viable
886	1	5	1	0	3	10	-	41	0	3	3.2	13.3	-2.3	40	0	0.378	Viable
320	0	4	1	1	2	10	38.1669	48	0	9	3.4	13.7	4.3	39	3.1	0.35	Viable
1449	1	3	1	0	1	10	-	43	0	5	3.2	10	3	41	0	0.348	Viable
381	2	2	1	0	0	10	-	44	0	3	3.3	10.7	3.3	35	0	0.324	Viable
1520	2	3	2	1	0	10	43.1338	47	0	-	2.4	15.8	1.2	37	5.6	0.32	Viable
1920	-	4	1	0	2	10	-	41	-	-	3.2	8.5	2.5	37	0	0.312	Viable
134	0	5	2	1	2	10	37.2033	46	0	4	2.9	12.7	3.3	36	2.7	0.31	Viable
2729	2	1	1	0	0	10	-	37	-	3	2	5	2	38	0	0.298	Miscarriage
2530	0	3	0	0	2	10	-	46	0	3	2.3	11.3	4.7	39	0	0.294	Miscarriage

Figure 2.6: Example of a nearest patients search based on the proximity matrix of unoptimized XGBoost model on the EPOS dataset. The query was randomly drawn from the dataset prior to supervised metric learning.
A Decision plot: the similarity of the 10 nearest cases to the patient query is decomposed into special additive SHAP values. The variable contributions sum up from left to right to reach the final similarity score on the right. The horizontal dashed line represents the baseline value corresponding to the average similarity between the query and the whole dataset. The SHAP values explain the gradual increase of the similarity along the most relevant variables compared to the baseline similarity value. Only the variables with an average absolute contribution greater than 0.05 SHAP are displayed. **B** The corresponding cases from the database are displayed in a HTML table with a conditional formatting based on the individual SHAP values. The scans ID are reported on the leftmost column. The similarity scores and the viability outcomes are reported on the rightmost columns.

2.4 Discussion

In this study, we corroborated previous findings that task-specific patient similarity metric outperformed unsupervised approaches. We also showed that proximity matrices from tree-based ensembles can be used to provide efficient task-specific metrics for similar patients retrieval on incomplete databases. Individual explanations of the similarity between two cases can be provided with post-hoc interpretability based on the SHAP framework. Alternatively, efficient retrieval of similar patients from incomplete data can be performed using weighted Euclidean distances but require carefully imputed datasets.

As previously pointed out in the literature [43–45, 80], our results demonstrated that task-specific metrics outperform the generic Euclidean distance in producing more discriminant metrics, even after proper standardization of the variables. The change in the neighborhood composition compared to generic distance neighborhoods was the highest with the Euclidean distance computed on the SHAP values space for all datasets. In comparison, WE distance approaches, which are less *supervised*, showed a higher overlap with generic distance approaches.

Our results demonstrated that proximity matrices from tree-based ensembles, although providing only pseudo-distances, are very efficient in retrieving similar patients from incomplete databases. By taking advantage of the ensemble's optimal default path learned on the training data, proximity matrices can also easily perform nearest neighbors search on incomplete queries. Moreover, we demonstrated that the need of hyper-parameters tuning was not always necessary. With a sufficient number of individual trees of reasonable complexity and a stochastic component [58], all models provided excellent proximity matrices for each dataset. These settings, by introducing a sufficient level of complexity in the ensemble, mitigated the impact of the pseudo-distance by reducing the probability of $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ when $\boldsymbol{x} \neq \boldsymbol{y}$.

In addition to handling missing values, the proximity matrices of tree-based ensembles hold several advantages for patient retrieval systems. They leverage the internal features selection mechanism and the implicit features weighting of the tree-based ensembles such that features irrelevant for model prediction are not considered to compute the similarity between patients [81]. Furthermore, the flexibility of tree-based ensemble architecture allows them to tackle various learning tasks, including classification, regression and survival analysis. Therefore, the supervision of the nearest search can easily be optimized for various external outcomes, not limited to binary classification tasks. Unlike (weighted) Euclidean distances, the approaches based on tree-based ensembles do not require any variables transformations such as scaling or standardization.

Finally, while most of the existing metrics are tailored for a specific data type, the proximity matrices inherit from the flexibility of decision trees to accommodate different variables types [81].

In spite of their efficiency, tree-based ensembles approaches suffer from a lack of transparency in delivering queries of similar patients, due to the heavy non-linear transformations of the input space through the ensemble. However, we demonstrated that the SHAP post-hoc interpretability framework, initially designed to explain model predictions [82], can be adapted to similar patients queries. By delivering additive explanations of the overall similarity, it becomes interpretable why two patients are considered similar by the algorithm. This explanation method can easily be extended to other non-linear computations of patient similarity, e.g. based on neural network [83]. Euclidean distance on the SHAP values space also provided excellent performance but required additional post-hoc computations of individual SHAP values. The interpretation of a query results can be delivered with decision path plots, where the query and the nearest patients share very similar decision paths along the trees ensemble.

The main limitation of metrics based on tree-based ensemble is the model dependency. Those approaches require the training of a model on the queried database beforehand, with the risk of overfitting the data. Secondly, computing and storing the similarity matrix for large datasets might be problematic [55], although in the specific setting of a CBR, typically only one single query is addressed at a time.

Although it was not the aim to build k-NN competing with tree-based ensembles, we should note that with a sufficiently large neighborhood of size k, k-NN based a proper task-specific distance metric can reach similar performance as XGBoost. Additionally, k-NN based on proximity matrices could even outperformed the very same unoptimized XGBoost models used to derive these matrices.

Beside non-linear transformations, the present results also highlighted the importance of the weighting scheme chosen to compute the WE distance. In all three datasets, the SHAP-FI weighting approach provided greater discrimination performance compared to other weighting schemes based on either AUC, C-index or XGBoost‘s internal FI. WE distance with SHAP-FI weights remains interpretable and competitive with complex non-linear transformations of the input space. However, this approach is also model-dependent and requires first to compute the individual SHAP values. Furthermore, although variables weighting interpretation is straightforward as it simply involves scaling the original variables, WE distances also require careful imputation of the incomplete database, which can be computationally intensive, especially in the presence of nested structures in the dataset (e.g. longitudinal data) [75]. Our results demonstrated that retrieving similar cases based on partial queries delivered

similar results to fully imputed queries.

Future works Future works should investigate the scalability of the present approach, both in terms of dimensionality (d) and size (N) of the dataset. Although XGBoost can natively handle the presence of missing value [52], its default path setting might not always be optimal (especially for crucial predictors with missing values). Additional works should investigate the impact of this default behavior. Finally, in the same vein as multiple kernels learning [84], similarity matrices from different data sources can easily be combined together. Therefore, future works could extend the present approach to enable multimodal queries.

2.5 Conclusion

Tree-based ensembles provide efficient task-specific distance metrics for patient retrieval systems. In particular, the similarity metric derived from the proximity matrix provides a powerful pseudo-distance for nearest neighbors searches on incomplete datasets. The flexibility of tree-based ensembles in terms of learning objectives, variables types and their facility to accommodate missing values constitute serious advantages over traditional methods based on Euclidean distance which require carefully scaled and imputed datasets. Post-hoc explanations based on SHAP values can provide individual explanations of the similarity between two cases and mitigate the lack of transparency of non-linear metrics.

Appendices

2.A Supplementary tables

Table 2.A.1: Tuning range of the XGBoost hyper-parameters

	Generic unoptimized ensemble	**Public Breast Cancer**	**EPOS**	**Kidney Transplant**
`n_estimators`	500	700*	110*	810*
`max_depth`	7	2-7	2-7	2-7
`learning_rate`	0.1	0.01-0.5	0.05-1.0	0.01-0.5
`subsample`	0.5	0.3-1.0	0.3-1.0	0.3-1.0
`colsample`	1	1	0.1-1	1
`reg-`α	0	0-1	0-1	0-1
`reg-`λ	0	0-5	0-5	0-5
Objective function	survival:cox binary:logistic	survival:cox	binary:logistic	survival:cox

*determined by early stopping

Table 2.A.2: XGBoost discriminatory performance without data imputation

	Public Breast Cancer (C-index)	**EPOS** (AUC)	**Kidney Transplant** (C-index)
Unoptimized GBM	0.674	0.857	0.779
Optimized GBM	0.687	0.883	0.798

Table 2.A.3: Proportion of common neighborhood when comparing the generic SE distance with various metrics on the EPOS dataset.
The imputation scheme is labeled as follows: *None* refers to no imputation, *Train* refers to imputation performed on the queried database only and *Train & Test* indicates that the imputation was performed on both the database and the query.

Distance Metric	**Imputation scheme**	***k*** **5**	**10**	**20**	**50**
Standardized Euclidean (SE)	Train	0.700 ±0.407	0.744 ±0.354	0.833 ±0.239	0.940 ±0.111
	Train & Test	**1.000**	**1.000**	**1.000**	**1.000**
AUC-Weighted SE	Train	0.077 ±0.129	0.138 ±0.131	0.267 ±0.141	0.434 ±0.165
	Train & Test	0.080 ±0.132	0.140 ±0.132	0.273 ±0.143	0.439 ±0.168
FI-Weighted SE	Train	0.069 ±0.118	0.128 ±0.123	0.252 ±0.133	0.415 ±0.159
	Train & Test	0.071 ±0.121	0.134 ±0.125	0.260 ±0.135	0.422 ±0.160
SHAP-Weighted SE	Train	0.050 ±0.104	0.094 ±0.109	0.189 ±0.114	0.320 ±0.136
	Train & Test	0.051 ±0.105	0.098 ±0.111	0.195 ±0.117	0.324 ±0.137
Generalized Mahalanobis	Train & Test	0.049 ±0.097	0.096 ±0.097	0.172 ±0.103	0.286 ±0.119
Proximity Matrix unoptimized model	None	0.055 ±0.102	0.093 ±0.097	0.152 ±0.101	0.253 ±0.115
	Train & Test	0.049 ±0.097	0.096 ±0.097	0.172 ±0.103	0.286 ±0.119
Proximity Matrix optimized model	None	0.032 ±0.078	0.052 ±0.074	0.087 ±0.075	0.155 ±0.085
	Train & Test	0.024 ±0.069	0.047 ±0.071	0.092 ±0.076	0.166 ±0.086
Euclidean -SHAP values	None	0.026 ±0.072	0.043 ±0.067	0.070 ±0.067	0.124 ±0.076
	Train & Test	0.020 ±0.061	0.037 ±0.063	0.075 ±0.069	0.132 ±0.078

Table 2.A.4: Proportion of common instances with the neighborhood from the generic SE distance for various metrics on the Kidney Transplant and Breast Cancer public datasets.
The imputation scheme is labeled as follows : *None* refers to no imputation, *Train* refers to imputation performed on the queried database only and *Train & Test* indicates that the imputation was performed on both the database and the query.

Distance Metric	**Imputation scheme**	**Kidney Transplant dataset** k**=5**	**10**	**20**	**50**
Standardized Euclidean (SE)	Train	0.781 ±0.364	0.819 ±0.312	0.891 ±0.19	0.987 ±0.032
	Train & Test	**1.000**	**1.000**	**1.000**	**1.000**
C-index-Weighted SE	Train	0.290 ±0.277	0.407 ±0.271	0.646 ±0.169	0.932 ±0.038
	Train & Test	0.318 ±0.286	0.435 ±0.278	0.674 ±0.166	0.938 ±0.032
FI-Weighted SE	Train	0.051 ±0.107	0.097 ±0.114	0.195 ±0.118	0.339 ±0.114
	Train & Test	0.053 ±0.110	0.100 ±0.118	0.200 ±0.12	0.341 ±0.114
SHAP-Weighted SE	Train	0.063 ±0.117	0.11 ±0.123	0.218 ±0.120	0.368 ±0.112
	Train & Test	0.063 ±0.120	0.115 ±0.125	0.224 ±0.121	0.370 ±0.113
Proximity Matrix unoptimized model	None	0.049 ±0.101	0.083 ±0.097	0.139 ±0.093	0.249 ±0.094
	Train & Test	0.040 ±0.095	0.074 ±0.100	0.147 ±0.099	0.258 ±0.097
Proximity Matrix optimized model	None	0.040 ±0.094	0.067 ±0.093	0.115 ±0.089	0.207 ±0.090
	Train & Test	0.030 ±0.085	0.055 ±0.086	0.114 ±0.090	0.203 ±0.088
Euclidean -SHAP values	None	0.016 ±0.062	0.032 ±0.064	0.054 ±0.065	0.101 ±0.072
	Train & Test	0.013 ±0.054	0.028 ±0.062	0.056 ±0.066	0.103 ±0.071
Distance Metric	**Imputation scheme**	**Breast cancer public dataset** k**=5**	**10**	**20**	**50**
Standardized Euclidean (SE)	Train	0.655 ±0.419	0.707 ±0.362	0.819 ±0.227	0.961 ±0.065
	Train & Test	**1.000**	**1.000**	**1.000**	**1.000**
C-index-Weighted SE	Train	0.300 ±0.294	0.417 ±0.272	0.661 ±0.17	0.942 ±0.057
	Train & Test	0.361 ±0.312	0.487 ±0.281	0.727 ±0.155	0.966 ±0.022
FI-Weighted SE	Train	0.041 ±0.089	0.076 ±0.093	0.147 ±0.098	0.256 ±0.116
	Train & Test	0.043 ±0.091	0.080 ±0.095	0.149 ±0.097	0.255 ±0.115
SHAP-Weighted SE	Train	0.048 ±0.099	0.089 ±0.105	0.178 ±0.126	0.293 ±0.159
	Train & Test	0.049 ±0.100	0.095 ±0.108	0.18 ±0.128	0.294 ±0.159
Proximity Matrix unoptimized model	None	0.055 ±0.102	0.093 ±0.095	0.155 ±0.098	0.248 ±0.101
	Train & Test	0.044 ±0.093	0.079 ±0.092	0.155 ±0.099	0.258 ±0.107
Proximity Matrix optimized model	None	0.045 ±0.094	0.078 ±0.091	0.132 ±0.103	0.219 ±0.121
	Train & Test	0.038 ±0.086	0.065 ±0.090	0.123 ±0.101	0.213 ±0.125
Euclidean -SHAP values	None	0.031 ±0.080	0.049 ±0.076	0.079 ±0.087	0.134 ±0.108
	Train & Test	0.022 ±0.067	0.040 ±0.069	0.077 ±0.085	0.140 ±0.114

Table 2.A.5: Discriminative performance in AUC of the k-NN classifiers on the public breast cancer dataset with a 5-folds cross-validation, for various k and distance metrics.

Distance Metric	**Imputation scheme**	**k**			
		5	**10**	**20**	**50**
Euclidean	Train	0.667	0.711	0.761	0.807
	Train & Test	0.696	0.733	0.776	0.814
Standardized Euclidean (SE)	Train	0.694	0.732	0.779	0.823
	Train & Test	0.688	0.738	0.792	0.830
AUC-Weighted SE	Train	0.742	0.799	0.840	0.859
	Train & Test	0.759	0.796	0.841	0.858
FI-Weighted SE	Train	0.727	0.782	0.831	0.858
	Train & Test	0.737	0.783	0.840	0.858
SHAP-Weighted SE	Train	0.751	0.815	0.847	0.862
	Train & Test	0.769	0.822	0.850	0.866
Generalized Mahalanobis	Train & Test	0.757	0.813	0.837	0.863
Proximity Matrix (unoptimized model)	None	0.772	0.805	0.835	0.866
	Train & Test	0.757	0.813	0.837	0.863
Proximity Matrix (optimized model)	None	0.759	0.807	0.844	0.865
	Train & Test	0.773	0.822	0.846	0.870
Euclidean -SHAP values	None	0.784	0.818	0.850	0.867
	Train & Test	0.771	0.814	0.851	0.871

Table 2.A.6: Discriminative performance of the k-NN classifiers on the public breast cancer dataset, for various k and distance metrics.

		C-index											
		12 months				60 months				120 months			
Distance metric	**Imputation scheme**	k=5	10	20	50	5	10	20	50	5	10	20	50
Euclidean	Train	0.506	0.509	0.516	0.541	0.563	0.574	0.589	0.613	0.565	0.579	0.592	0.613
	Train and Test	0.507	0.512	0.516	0.541	0.555	0.567	0.586	0.614	0.558	0.570	0.588	0.613
Standardized Euclidean (SE)	Train	0.508	0.508	0.521	0.549	0.557	0.581	0.601	0.624	0.55	0.577	0.599	0.627
	Train and Test	0.503	0.509	0.526	0.550	0.567	0.585	0.604	0.625	0.555	0.578	0.604	0.627
C-index-Weighted SE	Train	0.509	0.514	0.518	0.547	0.560	0.580	0.601	0.628	0.554	0.576	0.601	0.631
	Train and Test	0.507	0.515	0.527	0.551	0.565	0.583	0.600	0.626	0.563	0.577	0.599	0.628
FI-Weighted SE	Train	0.507	0.513	0.534	0.579	0.590	0.612	0.629	0.655	0.586	0.613	0.636	0.662
	Train and Test	0.510	0.515	0.536	0.576	0.587	0.609	0.631	0.653	0.586	0.614	0.643	0.660
SHAP-Weighted SE	Train	0.508	0.515	0.532	0.583	0.609	0.639	0.660	0.675	0.617	0.639	0.664	0.679
	Train and Test	0.508	0.517	0.538	0.584	0.613	0.638	0.658	0.674	0.616	0.639	0.666	0.677
Proximity Matrix (unoptimized model)	None	0.515	0.525	0.539	0.578	0.625	0.651	0.663	0.676	0.638	0.660	0.667	0.681
	Train and Test	0.508	0.519	0.527	0.578	0.600	0.636	0.657	0.677	0.620	0.651	0.661	0.681
Proximity Matrix (optimized model)	None	0.511	0.522	0.542	0.586	0.631	0.646	0.664	0.675	0.639	0.651	0.666	0.675
	Train and Test	0.509	0.513	0.530	0.588	0.610	0.640	0.659	0.672	0.627	0.646	0.663	0.674
Euclidean -SHAP values	None	0.511	0.528	0.564	0.604	0.606	0.643	0.657	0.666	0.622	0.652	0.668	0.675
	Train and Test	0.508	0.522	0.547	0.608	0.602	0.625	0.643	0.664	0.628	0.648	0.660	0.670

		IBS											
		12 months				60 months				120 months			
Distance metric	**Imputation scheme**	k=5	10	20	50	5	10	20	50	5	10	20	50
Euclidean	Train	0.006	0.005	0.005	0.005	0.088	0.082	0.078	0.076	0.160	0.147	0.141	0.137
	Train and Test	0.0054	0.005	0.005	0.005	0.089	0.082	0.079	0.076	0.162	0.149	0.142	0.137
Standardized Euclidean (SE)	Train	0.006	0.005	0.005	0.005	0.087	0.080	0.076	0.074	0.163	0.147	0.140	0.135
	Train and Test	0.006	0.005	0.005	0.005	0.0883	0.080	0.076	0.074	0.164	0.147	0.120	0.135
C-index-Weighted SE	Train	0.006	0.005	0.005	0.005	0.085	0.079	0.076	0.074	0.160	0.146	0.139	0.135
	Train and Test	0.006	0.005	0.005	0.005	0.085	0.079	0.0761	0.074	0.159	0.146	0.139	0.135
FI-Weighted SE	Train	0.005	0.005	0.005	0.005	0.084	0.078	0.075	0.073	0.156	0.141	0.135	0.131
	Train and Test	0.0055	0.005	0.005	0.005	0.085	0.079	0.075	0.073	0.155	0.142	0.137	0.132
SHAP-Weighted SE	Train	0.005	0.005	0.005	0.005	0.082	0.075	0.073	0.072	0.148	0.135	0.130	0.128
	Train and Test	0.005	0.005	0.004	0.004	0.081	0.0756	0.0732	0.072	0.149	0.136	0.130	0.128
Proximity Matrix (unoptimized model)	None	0.005	0.005	0.005	0.005	0.079	0.073	0.072	0.071	0.142	0.133	0.130	0.127
	Train and Test	0.005	0.005	0.005	0.005	0.085	0.077	0.073	0.071	0.152	0.138	0.132	0.128
Proximity Matrix (optimized model)	None	0.006	0.005	0.005	0.005	0.082	0.075	0.072	0.072	0.145	0.135	0.130	0.129
	Train and Test	0.006	0.005	0.005	0.005	0.082	0.076	0.072	0.072	0.148	0.136	0.130	0.129
Euclidean -SHAP values	None	0.005	0.005	0.005	0.005	0.088	0.076	0.074	0.073	0.152	0.137	0.132	0.130
	Train and Test	0.006	0.005	0.005	0.005	0.084	0.078	0.075	0.073	0.153	0.140	0.135	0.131

Table 2.A.7: Discriminative performance of the k-NN classifiers on the kidney transplant dataset, for various k and distance metrics.

		C-index											
		12 months				60 months				120 months			
Distance metric	**Imputation scheme**	k=5	10	20	50	5	10	20	50	5	10	20	50
Euclidean	Train	0.578	0.623	0.670	0.731	0.592	0.635	0.681	0.723	0.592	0.640	0.684	0.720
	Train and Test	0.568	0.610	0.661	0.725	0.588	0.627	0.674	0.716	0.595	0.645	0.681	0.715
Standardized Euclidean (SE)	Train	0.587	0.635	0.690	0.743	0.595	0.638	0.690	0.746	0.611	0.656	0.696	0.743
	Train and Test	0.588	0.637	0.691	0.740	0.603	0.645	0.690	0.742	0.622	0.663	0.693	0.741
C-index-Weighted SE	Train	0.590	0.639	0.715	0.748	0.608	0.665	0.718	0.757	0.623	0.672	0.717	0.753
	Train and Test	0.593	0.641	0.700	0.745	0.610	0.665	0.705	0.752	0.616	0.672	0.705	0.750
FI-Weighted SE	Train	0.591	0.623	0.664	0.728	0.599	0.634	0.676	0.731	0.613	0.632	0.664	0.718
	Train and Test	0.586	0.621	0.651	0.724	0.603	0.636	0.672	0.725	0.617	0.634	0.671	0.715
SHAP-Weighted SE	0.629	0.681	0.724	0.774	0.653	0.703	0.749	0.776	0.656	0.695	0.733	0.766	
	Train and Test	0.620	0.674	0.720	0.773	0.646	0.704	0.745	0.776	0.650	0.692	0.728	0.765
Proximity Matrix (unoptimized model)	None	0.671	0.705	0.730	0.767	0.690	0.730	0.754	0.783	0.689	0.725	0.741	0.769
	Train and Test	0.629	0.668	0.721	0.769	0.656	0.691	0.735	0.781	0.658	0.701	0.744	0.770
Proximity Matrix (optimized model)	None	0.634	0.667	0.704	0.745	0.653	0.687	0.713	0.755	0.659	0.696	0.719	0.732
	Train and Test	0.610	0.655	0.693	0.748	0.627	0.667	0.704	0.755	0.629	0.660	0.704	0.739
Euclidean -SHAP values	None	0.649	0.682	0.711	0.739	0.673	0.703	0.722	0.762	0.679	0.712	0.724	0.748
	Train and Test	0.632	0.677	0.709	0.748	0.667	0.709	0.736	0.767	0.674	0.706	0.738	0.752

		IBS											
		12 months				60 months				120 months			
Distance metric	**Imputation scheme**	k=5	10	20	50	5	10	20	50	5	10	20	50
Euclidean	Train	0.047	0.045	0.043	0.042	0.071	0.067	0.063	0.061	0.117	0.107	0.101	0.097
	Train and Test	0.048	0.046	0.043	0.042	0.071	0.067	0.064	0.061	0.117	0.108	0.101	0.097
Standardized Euclidean (SE)	Train	0.046	0.043	0.042	0.041	0.068	0.064	0.061	0.059	0.113	0.105	0.099	0.094
	Train and Test	0.046	0.043	0.042	0.041	0.067	0.063	0.061	0.06	0.111	0.103	0.099	0.095
C-index-Weighted SE	Train	0.046	0.043	0.041	0.04	0.069	0.064	0.060	0.058	0.114	0.104	0.097	0.093
	Train and Test	0.046	0.043	0.041	0.04	0.068	0.063	0.061	0.059	0.112	0.103	0.098	0.093
FI-Weighted SE	Train	0.047	0.045	0.043	0.041	0.070	0.066	0.063	0.06	0.117	0.109	0.102	0.096
	Train and Test	0.046	0.045	0.043	0.041	0.069	0.066	0.064	0.061	0.117	0.108	0.102	0.096
SHAP-Weighted SE	Train	0.044	0.041	0.039	0.037	0.065	0.061	0.058	0.055	0.111	0.102	0.095	0.090
	Train and Test	0.045	0.042	0.04	0.038	0.067	0.062	0.059	0.055	0.112	0.102	0.096	0.090
Proximity Matrix (unoptimized model)	None	0.035	0.034	0.035	0.035	0.055	0.053	0.053	0.053	0.100	0.093	0.091	0.089
	Train and Test	0.040	0.039	0.037	0.036	0.06	0.058	0.055	0.053	0.102	0.098	0.093	0.089
Proximity Matrix (optimized model)	None	0.039	0.038	0.037	0.038	0.061	0.058	0.057	0.056	0.107	0.100	0.096	0.093
	Train and Test	0.042	0.040	0.039	0.037	0.063	0.060	0.058	0.055	0.109	0.103	0.098	0.092
Euclidean -SHAP values	None	0.039	0.037	0.035	0.035	0.060	0.056	0.054	0.053	0.103	0.097	0.093	0.090
	Train and Test	0.041	0.039	0.037	0.035	0.061	0.056	0.055	0.053	0.105	0.098	0.095	0.091

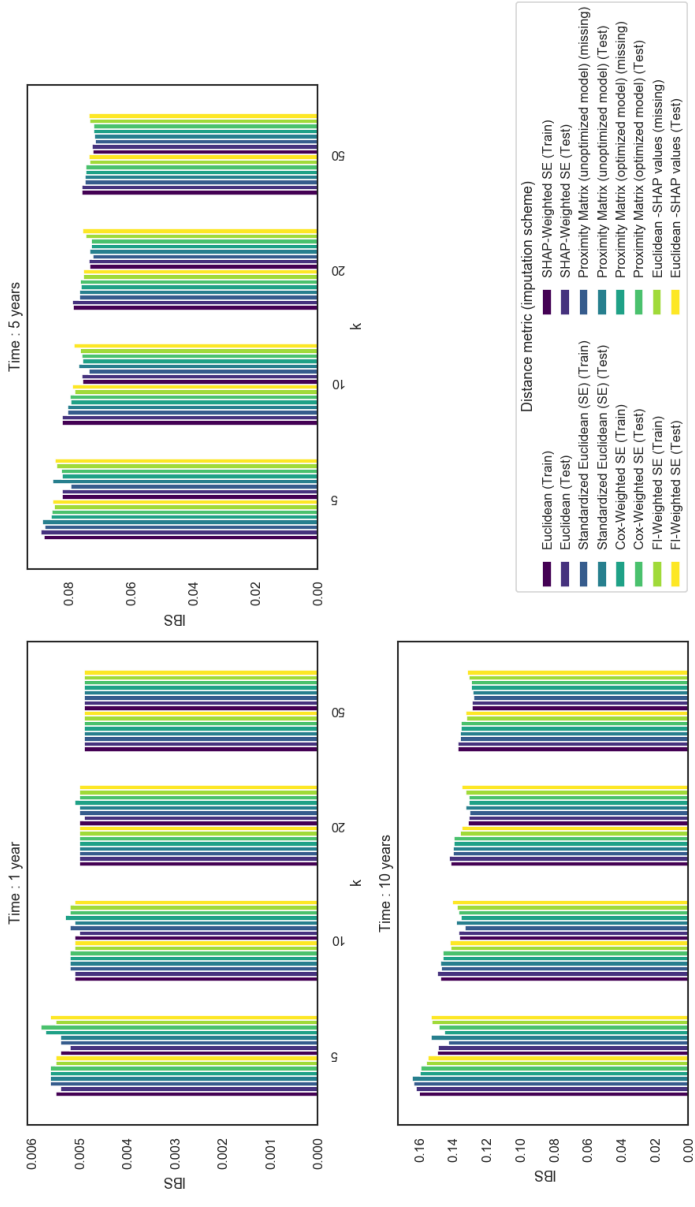


Figure 2.A.1: Discriminatory performance of k-NN measured by the IBS on the public breast cancer dataset for various metrics at k equals 5, 10, 20 and 50. The imputation scheme is labeled as follows: *missing* means that no imputation was performed, *train* means that imputation was performed on the queried database only and *test* means that imputation was performed both at the database and at the query levels.

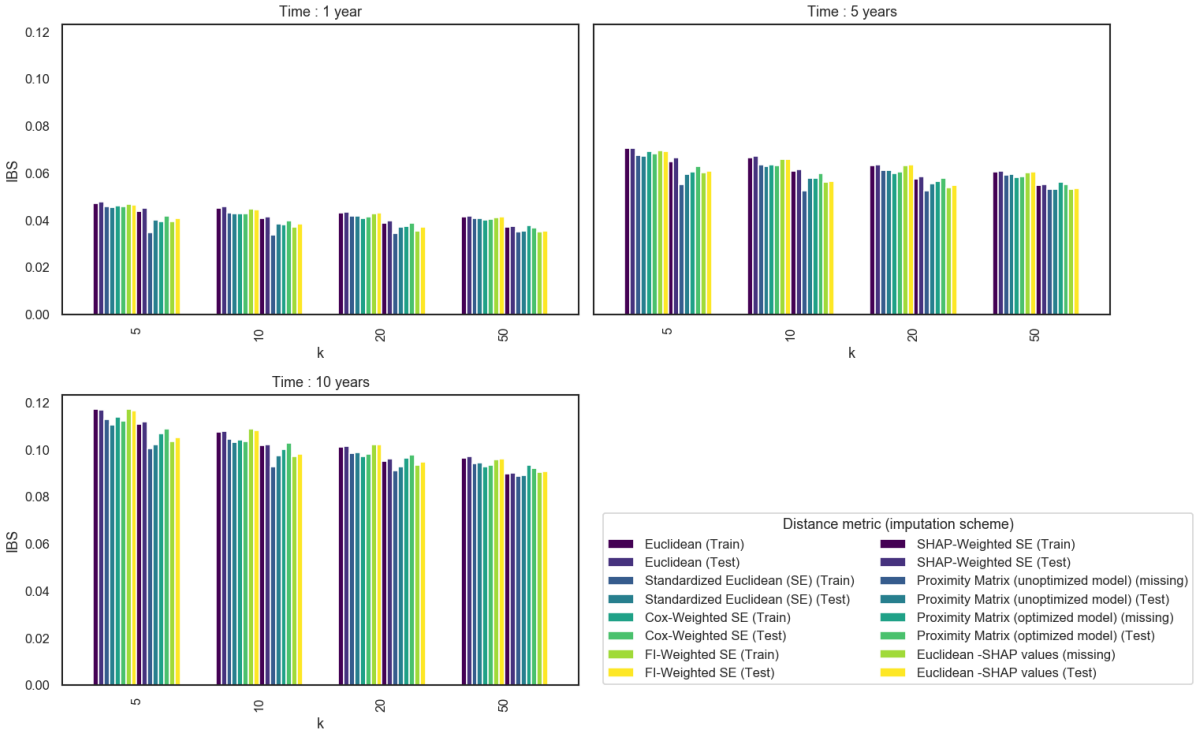


Figure 2.A.2: Discriminatory performance of k-NN measured by the IBS on the Kidney Transplant dataset for various metrics at k equals 5, 10, 20 and 50. The imputation scheme is labeled as follows: *missing* means that no imputation was performed, *train* means that imputation was performed on the queried database only and *test* means that imputation was performed both at the database and at the query levels.

2.B Post-hoc interpretability with SHapley Additive exPlanations

Post-hoc patient-level explanations aim to explain why the combination of a given set of values resulted in a specific prediction from a *black box* model. One way to compute those patient-level explanations is to use *additive feature attribution methods*.

2.B.1 Additive feature attribution methods

Those model-agnostic methods rely on explanation models to decompose each prediction as a sum of individual feature contributions. The explanation models often use simplified and interpretable inputs $\boldsymbol{x}'$ (e.g. binary vectors that simulate the presence or absence of certain variables in $\boldsymbol{x}$). $\boldsymbol{x}'$ can be mapped to the original input space with a mapping function $h_{\boldsymbol{x}}(\boldsymbol{x}')$, specific for the instance $\boldsymbol{x}$ such that $h_{\boldsymbol{x}}(\boldsymbol{x}') = \boldsymbol{x}$. Following the notation of Lundberg et al. [82], the explanation model g of additive feature attribution methods takes the form:

$$g(\boldsymbol{z}') = \phi_0 + \sum_{j=1}^{d} \phi_j \boldsymbol{z}'_j \tag{2.11}$$

where $\boldsymbol{z}' \in \{0,1\}^d$ is a binary vector of dimension d which simulates the subsets of variables by indicating the presence and the absence of the variables. $\boldsymbol{z}'_j$ represents therefore the presence (=1) or absence (=0) of feature j in $\boldsymbol{z}$. $\phi_j \in \mathbb{R}$ represents the feature attribution of the j^{th} variable of $\boldsymbol{z}$.

Whereas some additive methods (e.g. LIME [85]) approximate the global function locally, one can demonstrate that SHapley Additive exPlanations (SHAP) is the only additive attribution method that remains locally faithful to the black box model prediction $y = f(\boldsymbol{x}) = g(\boldsymbol{x}')$ [82]. In order to do so, SHAP produces explanations as *Shapley values*.

2.B.2 The Shapley values

The concept of Shapley values [86] was introduced by Lloyd Shapley (who won the Nobel Prize in Economics for it in 2012) in the cooperative game theory. It provides a solution to the question: *what is the contribution of each individual player with regard to a common payoff in a collaborative game?* The difficulty being that, in cooperative games, players are not independent of each other's,

they can create synergies or antagonisms depending on the sets of players. In the cooperative game theory, a given set of players is called a *coalition*.

When applied to model explanation, the Shapley values assure that the individual contributions to the model output are fairly distributed among the variables (acting here as the players). For a given prediction, the Shapley value of the variable j estimates the impact of the variable on the model output. It is computed as the weighted average of the all the differences between predictions with- and predictions without using the variable j. Therefore, to compute the exact Shapley value of a variable, all possible coalitions of variables must be estimated with and without the j^{th} variable.

Formally, for a given variable j (set at a defined value), its contribution to the final prediction is given as:

$$\phi j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(d - |S|! - 1)}{d!} [f(\boldsymbol{x}_{S \cup \{j\}}) - f(\boldsymbol{x}_S)] \tag{2.12}$$

where F is the set of all input variables and S represents all possible subsets of variables excluding j.

2.B.3 SHapley Additive exPlanations (SHAP)

In practice, two major problems occur that prevent the computation of exact Shapley values:

1. Most of the models f cannot natively handle the presence of missing variables (and therefore compute predictions on subsets of variables such as $\boldsymbol{x}_S$)
2. Because of the combinatory nature of Equation 2.12 (2^d possible combinations of variables), the computation of the exact Shapley values becomes rapidly impracticable when d increases.

Before developping how SHAP tackles these two problems, let's first introduce some notations:
$\boldsymbol{z}' \in \{0, 1\}^d$ is a coalition vector.
S is the set of non-zero entries in $\boldsymbol{z}'$, C is the set of zero entries in $\boldsymbol{z}'$ (the complement of S, with all the "missing" variables.)
$h_{\boldsymbol{x}}(\boldsymbol{x}')$ is the mapping function, specific to instance $\boldsymbol{x}$, that converts the binary coalition vector $\boldsymbol{x}'$ into the original input space. $h_{\boldsymbol{x}}(\boldsymbol{x}') = \boldsymbol{x}$.
With $h_{\boldsymbol{x}}(\boldsymbol{z}') = \boldsymbol{z}$, $\boldsymbol{z}$ contains value of $\boldsymbol{x}$, when $\boldsymbol{z}' = 1$ and sampled values from

the training data $\boldsymbol{X}$ where $\boldsymbol{z}' = 0$.

To overcome 1.), SHAP simulates the dropping of variables by marginalizing over the set of unused variables C [82] [3]. In practice, SHAP does the marginalization through an *interventional* approach designed to break the dependence between the variables [87]. In this process, the variables are deliberately set to values sampled from a dataset with the mapping function $h_x(\boldsymbol{z}')$, irrespective of the other variables [4]. Because $\boldsymbol{z}_C$ and $\boldsymbol{z}_S$ are treated as independent with the interventional approach, $f(h_{\boldsymbol{x}}(\boldsymbol{z}')) = f(\boldsymbol{z}_S)$ is approximated by $E_{\boldsymbol{z}_C}[f(\boldsymbol{z})]$, where E is the marginal expectation over $\boldsymbol{z}_C$, i.e. the marginal expectation over the "missing" variables. As a drawback, this approach requires a dataset to provide the sampled values for the mapping $h_{\boldsymbol{x}}(\boldsymbol{z}') = \boldsymbol{z}$.

To solve problem 2.), SHAP uses a sampling of the coalitions. Instead of computing every possible coalition, it approximates the Shapley value with a clever weighting scheme (see Equation 2.13) and an appropriate sampling strategy, which focuses on high-weight coalitions first. Ultimately, the approximated SHAP values are computed with a weighted linear regression that minimizes the loss function of Equation 2.14. The resulting coefficients ϕ_j's of the optimized models constitute the Shapley values.

$$\pi_{\boldsymbol{x}'}(\boldsymbol{z}') = \frac{(d-1)}{(d \text{ choose } |\boldsymbol{z}'|)|\boldsymbol{z}'|(d-|\boldsymbol{z}'|)} \tag{2.13}$$

$$L(f, g, \pi_{\boldsymbol{x}'}) = \sum_{\boldsymbol{z}' \in \boldsymbol{X}} [f(h_{\boldsymbol{x}}(\boldsymbol{z}')) - g(\boldsymbol{z}')]^2 \pi_{\boldsymbol{x}'}(\boldsymbol{z}') \tag{2.14}$$

where $\boldsymbol{X}$ is the training dataset.

2.B.4 Properties of SHAP

SHAP satisfies three important properties :

Local accuracy (efficiency)

[3] Note that conceptually, the creators of SHAP suggested to use a conditional approach to sample the dropped variables [82]. However, as highlighted by Janzing et al. [87], this lead to biased estimations as it attribute non-zeros SHAP values to non-influential variables, i.e. variables that are not explicitly used by the model. The reason is that conditional expectation on a given variable, e.g. $E[y|\boldsymbol{X}_1]$, does not only condition on $\boldsymbol{X}_1$, but also implicitly conditions on other variables correlated to $\boldsymbol{X}_1$, as in real life, variables are rarely completely independent of each other's.

[4] To avoid the confusion, some authors recommend to use the Pearl *do* operator, i.e. the variable $\boldsymbol{X}_1$ is *manually* set to a value x, not just *observed* being x, as it would be with regular conditioning $y|\boldsymbol{X}_1 = x$ [87].

$$y = f(\boldsymbol{x}) = g(\boldsymbol{x'}) = \phi_0 + \sum_{j=1}^{d} \phi_j \boldsymbol{x'_j}$$

The explanation model corresponds to the original model when $\boldsymbol{x} = h_{\boldsymbol{x}}(\boldsymbol{x}')$. Because of the local accuracy (or efficiency) property, SHAP is one of the rare explanation methods to deliver a complete explanation. As a consequence, selective explanations are not possible. ϕ_0 corresponds to a baseline prediction often set to $E[f(\boldsymbol{X})]$. Therefore, the SHAP values estimate the contribution of each variable to explain the difference between $f(\boldsymbol{x})$ and $E[f(\boldsymbol{X})]$, the averaged prediction of the model.

Missingness

$\boldsymbol{x}'_j = 0 \longrightarrow \phi_j = 0$

Missing variables in the original input has no impact, therefore the corresponding SHAP value is zero.

Consistency

$f_{\boldsymbol{x}}(\boldsymbol{z}') = f(h_x(\boldsymbol{z}'))$ and $\boldsymbol{z}'\backslash j$ corresponds to $\boldsymbol{z}'_j = 0$.

For any two models f and f', if $f'_{\boldsymbol{x}}(\boldsymbol{z}') - f'_x(\boldsymbol{z}'\backslash j) \geq f_{\boldsymbol{x}}(\boldsymbol{z}') - f_{\boldsymbol{x}}(\boldsymbol{z}'\backslash j)$ for all input $\boldsymbol{z}' \in 0, 1^d$, then $\phi j(f', \boldsymbol{x}) \geq \phi_j(f, \boldsymbol{x})$

If a model is modified such that one variable contributes equally or more to the output, its variable attribution should not decrease.

In contrast with LIME , which can also be an additive feature attribution method, the SHAP framework builds on *contrastive* explanations, as each explanation is expressed relative to a baseline prediction. Therefore, SHAP better complies with the implicit notion of *explanation* (see Chapter 5).

SHAP benefits from several implementations tailoring specific algorithms. In particular, TreeShap is tailored for tree-based algorithms such as random forest or gradient boosted trees and exploits their specific internal structure to facilitate the computation of SHAP values [54]. Although the large computational cost of SHAP can hinder its implementation, it remains an attractive approach for low-dimensionality data, which constitute the vast majority of current clinical datasets. The complex process behind SHAP might constitute an obstacle to its adoption in DDSS, as it is not very intuitive at first side and difficult to explain. However, as we will see, clever visualizations techniques can facilitate the communication of SHAP explanations to a wide audience.

2.B.5 Interpretability at the variable-level

Variable-level explanations are interesting to understand the behavior of a given variable on the model outcome. However, because it restricts the explanations to a single or a few variables, those methods provide only a partial view of the model. Variables can be investigated with *feature importance* (FI) which assigns an importance score to each feature of the model regarding a pre-defined objective. There exist several approaches to compute FI. The permutation FI is model-agnostic. It is therefore easy to compare various algorithms together. Although computationally intensive, it produces more reliable estimates than traditional FI measures [88, 89]. The idea is to assess the model performance (e.g. accuracy) after randomly permuting a variable values. If a model is highly dependent on a given variable, its performance should significantly drop if the variable values are randomly shuffled. In this setting, individual FI is expressed in terms of prediction errors (e.g. 1-AUC). SHAP also produces a special measure of feature importance defined as the mean absolute SHAP value of that variable on the training data.

2.B.6 Interpretability at the model-level

Because post-hoc interpretability deals mostly with local explanations, interpretability at the model-level is not straightforward. However, it is possible to derive some understanding of a model using variable-level and patient-level explanations. The SHAP method proposes the aggregation of individual SHAP values to depict an overview of the model behavior on a given dataset [90]. Recent work by Covert et al. [91] tackles the issue to derive global feature importance associated with individual predictive power and accounting for feature interactions.

2.C Impact of the XGBoost's main hyperparameters on the proximity matrix performance

In this supplementary material, we report the impact of various hyper-parameters of XGBoost on the performance of the produced similarity metric. We trained an XGBoost algorithm on artificial data and on the Wisconsin breast cancer dataset [92] to produce a proximity matrix, which was subsequently used in a k-NN classifier.

2.C.1 Method

Artificial data were generated using the `make_classification` function of Sklearn [93]. The following settings were used: `n_samples`= 800, `n_features`= 50, `shuffle`= True, `n_clusters_per_class`= 2, `n_informative`= 40, `n_redundant`= 3, `random_state`= 42.

We used 20 repeated 5-folds cross-validation to estimate the performance of the similarity metric. At each of the 20 iterations, the data were split under a random 5-folds partition. Subsequent cross-validation was performed: a similarity matrix was produced with XGBoost on the training folds following the approach described in Section 2.2.4 and used in a k-NN classifiers on the corresponding testing fold. The discriminatory performance assessed with AUC for $k \in \{5, 10, 20, 50\}$. The predictions of the k-NN classifiers were aggregated over the 5-folds of the cross-validation prior to derive a unique performance estimate. This approach produced therefore 20 different results for a given combination of XGBoost hyper-parameters.

We focused on the following four hyper-parameters : the numbers of trees (`n_estimators`), the maximal depth of each individual tree (`depth`), the `subsample` and the `colsample` parameters. The `subsample` parameter indicates the proportion of original data that is used to create an individual tree. The algorithm can randomly sample the training data in a subset of the original data given by the `subsample` proportion. This parameter introduces a stochastic component in the process. Similarly, `colsample` parameter indicates the amount of variables subsampling to construct each individual tree. For instance, a `colsample` of 0.5 uses only half of the variables.

2.C.2 Results

Larger numbers of trees and more complex trees (i.e. higher `depth`) provided the best similarity metrics as indicated by Figures 2.C.1 and 2.C.4. Large ensembles of complex trees resulted in a finer granularity of the proximity matrices. The difference between `depth` 6 and 7 appeared negligible and corroborated the findings by Hastie [81] that increasing the complexity of individual trees beyond a certain extent is probably not necessary to produce good results. The stochastic component implemented by the `subsample` parameter acts as a regularizer parameter to prevent overfitting of the tree ensemble. The best proximity matrices were produced with values 0.25 and 0.50 as demonstrated on Figures 2.C.2, 2.C.5, 2.C.3 and 2.C.6.

The `colsubsample` parameters produced the best metrics when set to its maximum value as shown in Figures 2.C.3 and 2.C.6. Intuitively, increasing the amount of predictors also increases the granularity of the measurements, as potentially more variables contribute to the definition of the similarity. These results demonstrated the importance of the four parameters tested in producing good proximity matrices.

To summarize, efficient similarity metrics can be produced with the following settings: a sufficiently large number of complex trees (`n_estimators`=500-1000 and `depth`=6-7), with a significant amount of data subsampling (`subsample=0.25-0.5`) covering the whole set of variables (`colsample`=1).

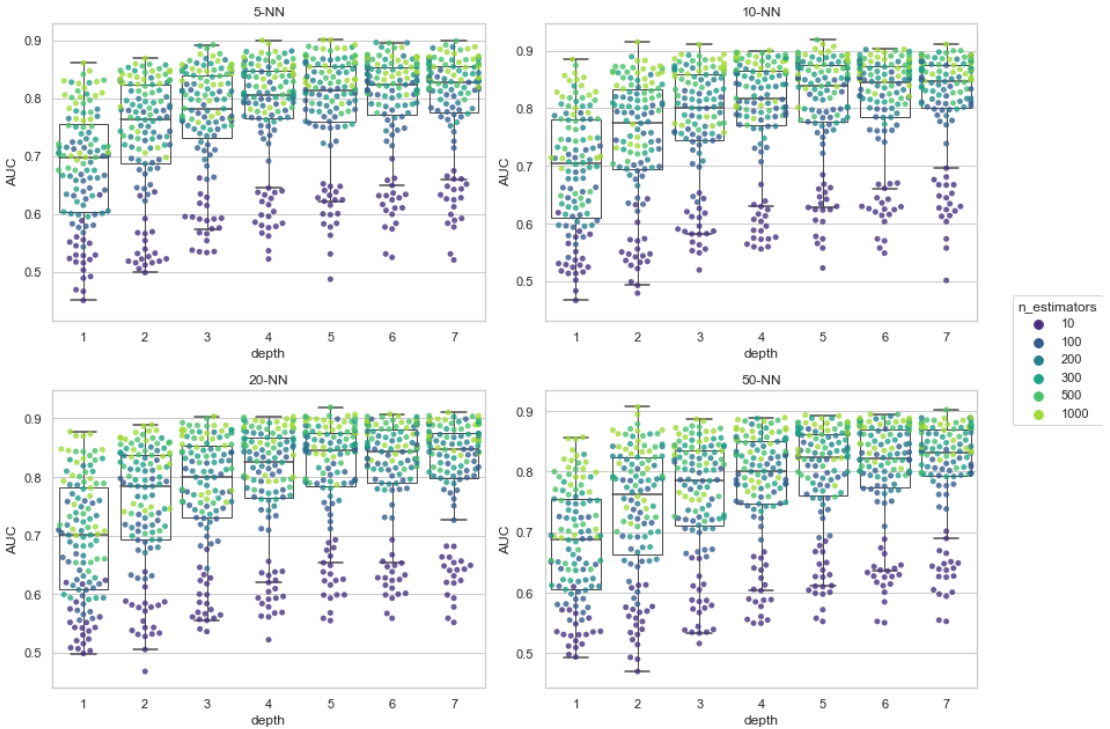


Figure 2.C.1: Impact of the numbers of trees (`n_estimators`) and the maximal depth of each individual tree (`depth`) - Artificial Data.
The `subsample` and the `colsample` parameters were fixed to 0.5 and 1 respectively. Other parameters were set to their default values. Larger numbers of trees and trees that are more complex provide better metrics as the resulting proximity matrices have a finer granularity.

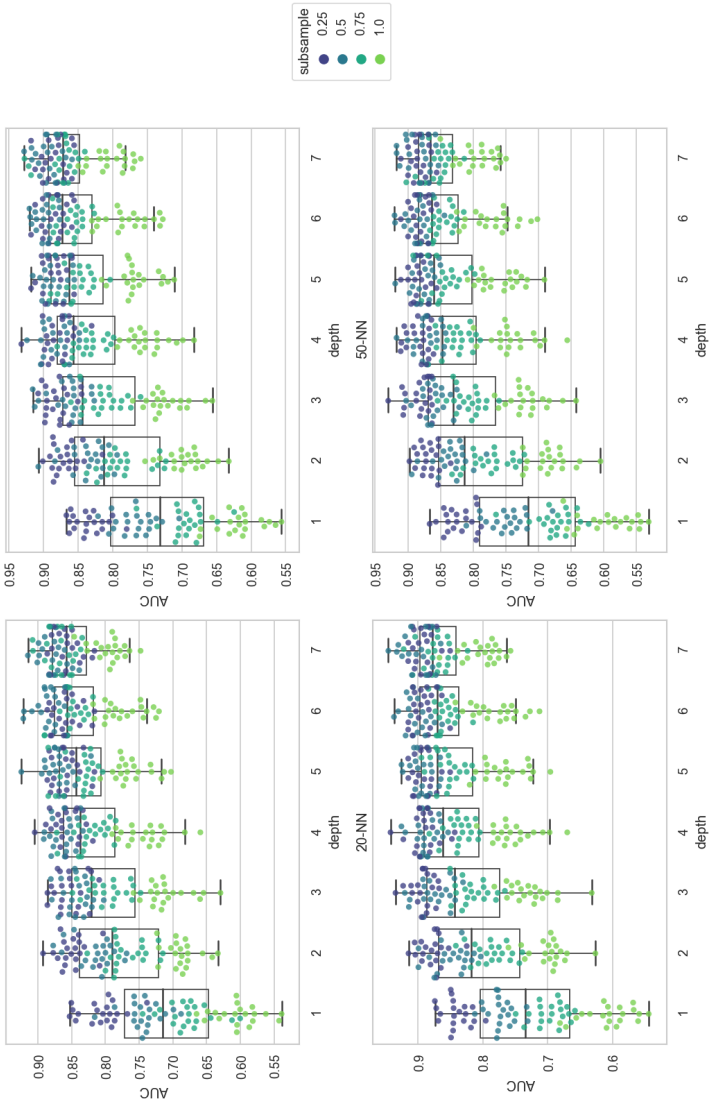


Figure 2.C.2: Impact of the the numbers of trees (`n_estimators`) and the `subsample` parameter - Artificial Data. The `depth` was set to 7 and the `colsample` parameter was set to 1. Other parameters were set to their default values. The stochastic component, i.e. when `subsample` value is less than 1, is crucial to derive good metrics. The more complex are the individual base trees, the better the metric as a higher complexity increases the fine granularity of the proximity matrix.

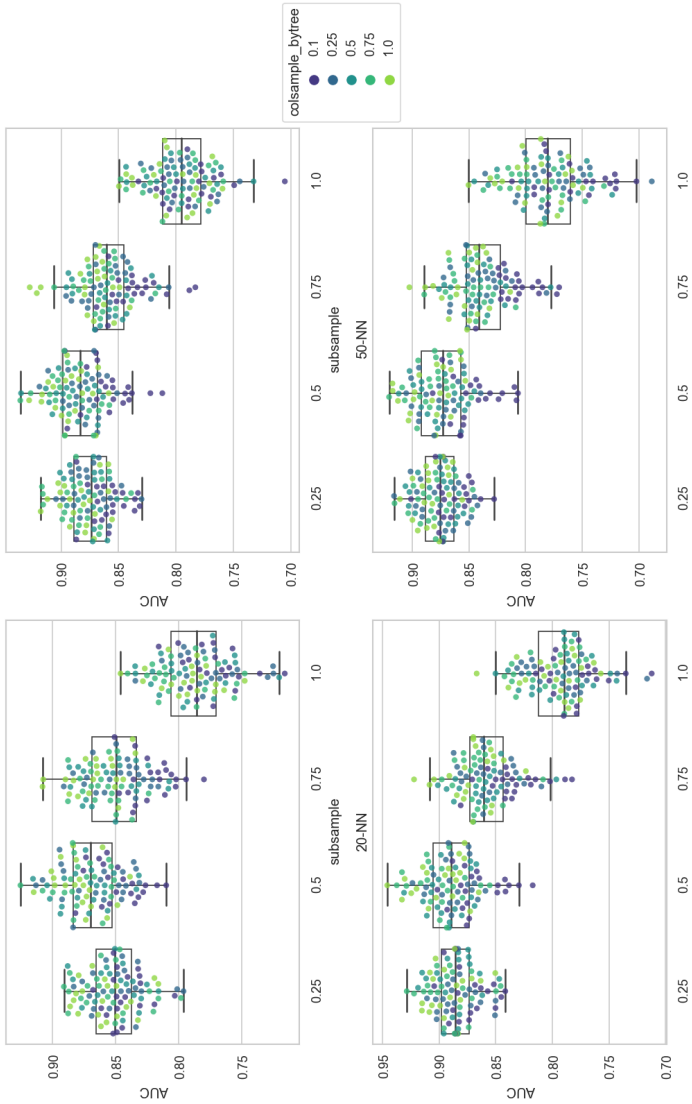


Figure 2.C.3: Impact of the `colsample` and `subsample` parameters - Artificial Data. The number of `estimators` was set to 500 trees and the maximal `depth` parameter was set to 7. Other parameters were set to their default values. As demonstrated above, adding stochasticity to the model by lowering the `subsample` parameter improved the discrimination of the metric. On the other hand, subsampling the variables was less effective. Especially, including only a very low fractions (e.g. 0.1) of the variables produced noticeably worse results.

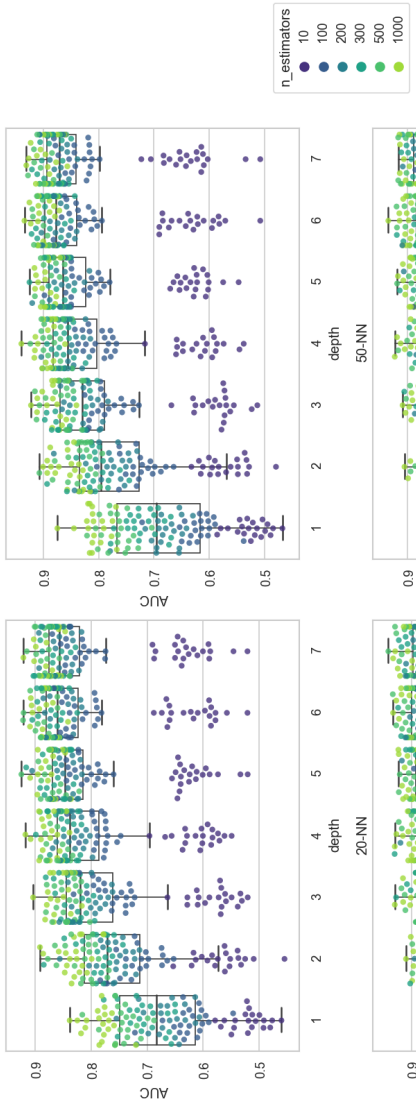


Figure 2.C.4: Impact of the numbers of trees (`n_estimators`) and the maximal depth of each individual tree (`depth`) - Wisconsin dataset.
The `subsample` and the `colsample` parameters were fixed to 0.5 and 1 respectively. Other parameters were set to their default values. Larger numbers of trees and more complex trees provide better metrics as the resulting proximity matrices have a finer granularity.

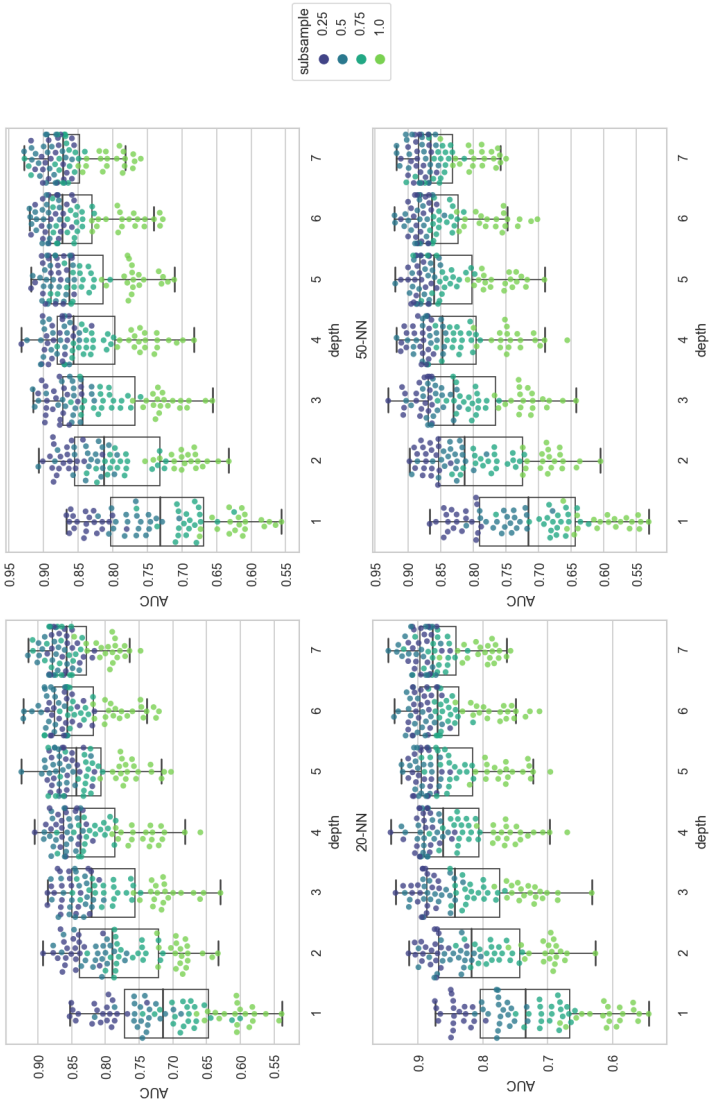


Figure 2.C.5: Impact of the maximal depth (`depth`) and the `subsample` parameters -Wisconsin datase. The number of `estimators` was set to 500 trees and the `colsample` parameter was set to 1. Other parameters were set to their default values. The stochastic component, i.e. when `subsample` value is less than one, is crucial to derive good metrics. The more complex are the individual base trees, the better the metric as a higher complexity increases the fine granularity of the proximity matrix.

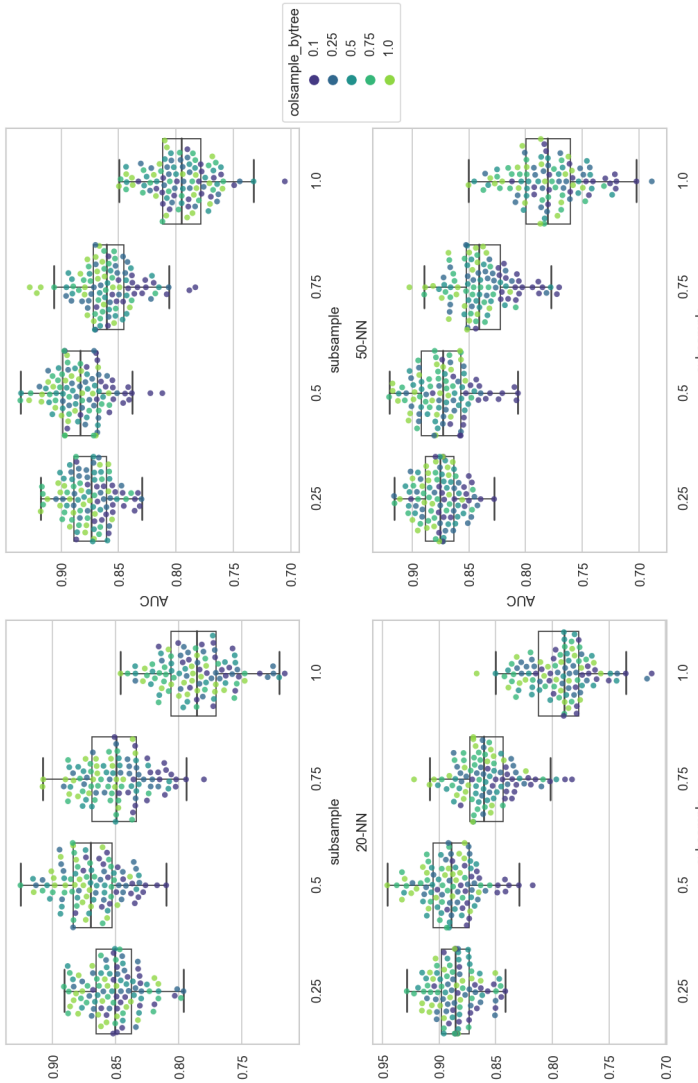


Figure 2.C.6: Impact of the `colsample` and `subsample` parameters - Winsconsin dataset. The number of `estimators` was set to 500 trees and the maximal `depth` parameter was set to 7. Other parameters were set to their default values.
As demonstrated above and similarly to the artificial data, adding stochasticity to the model by lowering the `subsample` parameter improves the discrimination of the metric. On the other hand, subsampling the variables was less effective. Especially, including only a very low fractions (e.g. 0.1) of the variables produced noticeably worse results.

Chapter 3

Viability prediction in the first trimester of pregnancy

Part of this chapter has been submitted to the *American Journal of Obstetrics and Gynecology* as:
Al-Memar, M., Fourie, H., Vaulet, T., *et al. Prediction of First Trimester Viability Using Simple Clinical and Ultrasound Variables*

3.1 Problem statement

Early pregnancy loss or first trimester miscarriage is defined as the loss of a viable pregnancy before the 13^{th} week of gestation [95][1]. It is the most common complication of early pregnancy [97]. Although its precise incidence is difficult to assess as many miscarriages occur before the pregnancy is confirmed, recent studies estimates its incidence in recognized pregnancies around 13-17% [98–100]. Early pregnancy losses result in potentially devastating psychological impact on the parents to be with psychological distress that can last for several months [101–103].

Due to the advent of sensitive home pregnancy testing kits, many women attend ultrasound scans at early stages of pregnancy. However, for early

[1]The current ultrasound-based criteria to make a miscarriage diagnosis are an empty gestational sac measuring > 25 mm in mean diameter, or an embryo with crown–rump length > 7 mm and no evidence of cardiac activity [96].

gestational ages (GA), the initial ultrasound scans are often inconclusive with a significant number of pregnancies labeled as *pregnancies of unknown viability* (PUV) [104]. This situation occurs when no embryo is demonstrated and/or its cardiac activity is not visible under transvaginal echography. Because the vast majority of pregnancy losses occur during the first trimester [2], the outcome of a sonographic diagnosis at 13 weeks is a good indicator of the future pregnancy viability. For women left with an uncertain diagnosis, this situation causes great anxiety that justifies the need for accurate predictive models on early pregnancies. Predicting first trimester viability would also enable resource allocation to pregnancies deemed at risk of miscarriage.

By combining information from demographics, clinical symptoms, ultrasound data and/or serum hormone levels, several multivariable predictive models have been developed [105–112]. Despite good predictive accuracy (AUC between 0.75-0.95) and for some, external validation [107], those studies had several limitations. Some of those models have been trained on an outdated definition of miscarriage [106–109], while others addressed specifically viable pregnancies and therefore discarded a large proportion of initial PUV which prevents them from being generalizable to early gestational ages [110–112]. Finally, some studies targeted specific subpopulations with restrictive inclusion criteria (e.g. pregnancies from assisted reproductive technique [105]).

Ultimately, most of those models have not reached clinical practice. Beside the aforementioned limitations, we also postulate that the lack of an accessible and interpretable interface to deploy those models has potentially contributed to their poor dissemination.

Objectives

The main objective of this project is to design a DDSS that integrates calibrated and interpretable models to predict the pregnancy outcome at the end of the first trimester. This DDSS should also include a case-based reasoning engine to retrieve similar cases and further exploit a high-quality dataset on early pregnancy. In addition, this DDSS should handle the presence of missing values to allow prediction and similar cases retrieval despite incomplete queries.
An overview of the DDSS structure is displayed in Figure 3.1.

[2]Subsequent losses are called second trimester miscarriages (between 13 and 20 weeks of GA) or stillbirths (after 20 weeks of GA).

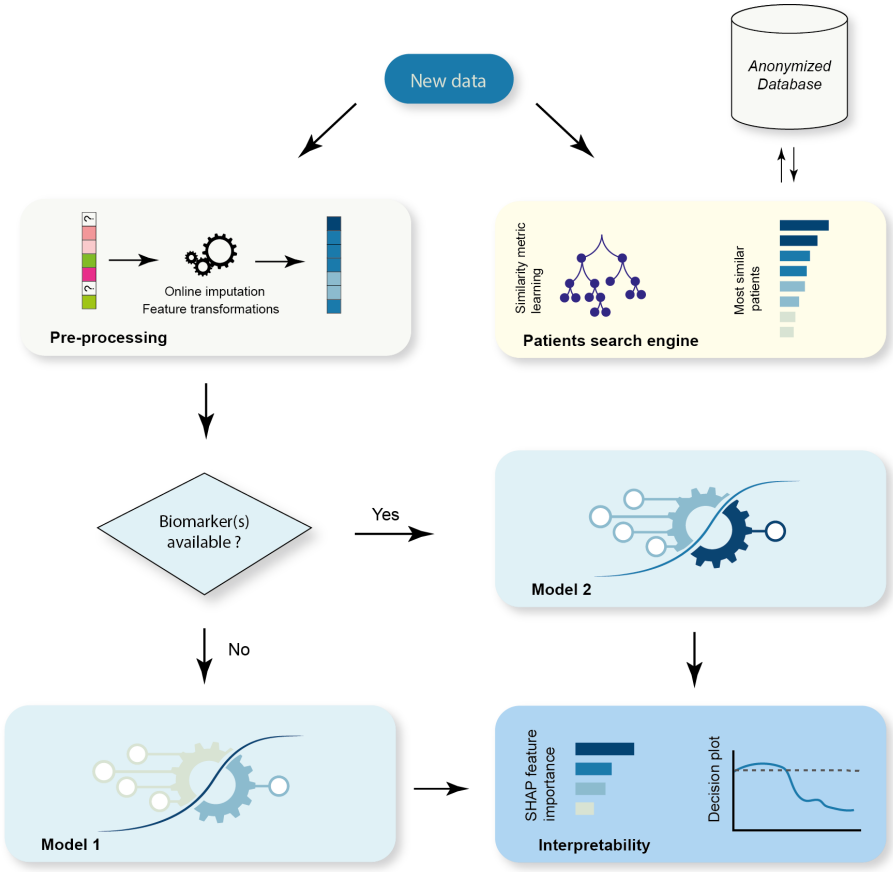


Figure 3.1: Overview of the DDSS structure for the prediction of viability in the first trimester of pregnancy.
When a prediction needs to be performed on a new data entry, the query first undergoes relevant pre-processing steps (Section 3.4), including potential online imputation. Additionally, the provided data can also trigger the retrieval of similar patients (Section 3.5) from the EPOS database (Section 3.2). Based on the availability of biomarkers values at the time of the prediction, the appropriate predictive model is selected (Section 3.3). Finally, explanations about the prediction are provided by the last component (Section 3.6). The development of each component is described in details in the following sections of this chapter.

3.2 The EPOS dataset

The dataset used to construct this DDSS was obtained from a prospective observational cohort study based at Queen Charlotte's & Chelsea Hospital, London, conducted between March 2014 and May 2019. The study had been approved by NHS National Research Ethics Service (NRES) Riverside Committee London (REC 14/LO/0199) and NHS North East – Newcastle and North Tyneside 2 Research Ethics Committee (17/NE/0121). All participants provided written informed consent. Details on the study design and recruitment criteria can be found in [113].

3.2.1 Inclusion criteria

Women in the first trimester of pregnancy (<14 weeks by last menstrual period (LMP) or ultrasonography dating of embryo crown rump length measurements (CRL) when LMP was not known), with intrauterine pregnancies were invited to participate. An intrauterine pregnancy was defined as an ultrasound scan showing an intrauterine gestation sac. The absence of an embryo and/ or a visible heartbeat defined PUV. The presence of an embryo and heartbeat confirmed a viable pregnancy. Women aged less than 16 and over 50 years were excluded. Participants were recruited via advertisements in local GP surgeries and hospitals, and at the university where the study was conducted. The majority of women were recruited following an attendance at the Early Pregnancy Assessment Unit.

3.2.2 Study Design

Serial ultrasound scans were performed until the routine first trimester dating scan at 11-14 weeks' gestation. Ultrasonography measurements were taken at each visit (see ultrasound features below). Participants were encouraged to contact the research team if they had any complications, such as bleeding. If necessary, women were invited to attend an additional ultrasound scan. Pregnancy outcomes (first trimester miscarriage or viable pregnancy) were collected using hospital medical records. Diagnosis of miscarriage was based on strict sonographic criteria [96]

Out of 1757 patients screened, 1403 patients were recruited for a total of 3813 scans (with an average of 3.8 scans per patient). Reasons for declining to take part included: inability to attend follow up, choosing to book antenatal care elsewhere and personal choice.

3.2.3 Variables description

A summary of the main variables registered in the EPOS dataset project is reported in Table 3.1. This dataset consists of panel data constituted of baseline features related to the parents and the current pregnancy and of longitudinal variables repeatedly measured at each visit. Note that the interval between visits was not predefined and the reasons for visit were either for reassurance or because of the presence of pelvic pain and/or vaginal bleeding.

Baseline variables

Those features were recorded during the initial visit and were considered constant throughout the pregnancy. They include demographics (age, ethnicity of parents, body-mass index (BMI)), addiction habits (smoking, alcohol usage), medical and surgical history (e.g. number of previous miscarriage(s), history of previous surgery), medications (e.g. folic acid supplementation, metformine treatment). The data also included the estimated date of last menstruation period (LMP) along with the certainty of the recall, varying from 0 (unknown) to 10 (100% certain). A precise knowledge of the LMP is important in first trimester studies as it allows to assess potential growth retardation by comparing the observed GA based on ultrasound features to the expected GA, based on LMP.

Longitudinal variables

Clinical features Pelvic pain and vaginal bleeding are very common symptoms in early pregnancy and often a reason for referral. At each visit, the patients were asked to evaluate their symptoms with validated symptoms scores. Vaginal bleeding was evaluated with the pictorial blood assessment chart (PBAC), a numerical scale from 0 (no bleeding) to 4 (clots/ flooding) [114]. Pelvic pain was assessed using the visual analogue score (scale of 0 to 10) [115]. Both bleeding and pain symptoms scores were reported for the day of attendance as well as the maximal scores experienced. The total number of days experiencing pain and/or bleeding were also reported. Additional clinical variables include the pregnancy-unique quantification of emesis score (PUQE score), which assesses the nausea symptoms on a range from 3 to 15.

Ultrasound features The ultrasound variables measured were part of the standard biometric measurements performed during the first trimester of pregnancy. The included: the crow-rump length (CRL), i.e. the length of the embryo, the mean sac diameter (MSD), the mean yolk-sac diameter (MYSD),

Table 3.1: Summary of the main variables collected during the EPOS study

Variables	Range	Unit
Baseline		
Maternal age	18-45	years
Paternal age	18-45	years
Maternal BMI	15.8-53.9	kg/m^2
Maternal ethnicity	White, black, asian, mixed, others	-
Paternal ethnicity	White, black, asian, mixed, others	-
Gravida	1-15	
Para	1-6	
Number of previous caesarian section(s)	0-4	-
Number of previous first trimester miscarriage(s)	0-11	-
Folic acid supplementation	yes-no	-
Smoking status	yes-no	-
Certainty of LMP	0-10	-
Longitudinal		
Clinical		
Current bleeding score	0-4	-
Max bleeding score	0-4	-
Total bleeding days	0-68	days
Current pain score	0-10	-
Max pain score	0-10	-
Total pain days	0-30	days
PUQE score	0-15	-
GA by LMP	18-209*	days
Ultrasound		
CRL	0.0-92.2	mm
MSD	0.0-68.9	mm
MYSD	0.0-34.0	mm
Fetal heart presence	yes-no	-
Amnion presence	yes-no	-
Presence of subchorionic hematoma	yes-no	-
GA by CRL	0-120	days
GA by MSD	0-99	days
Biomarkers		
Progesterone	0-135	ng/mL
β-HCG	0-429720	mIU/mL
E2	0-6414	pg/mL
CA-125	0-554	U/mL
α-FP	0-95	ng/mL

Additional variables : Treatment with aspirin, metformine and/or progesterone; preconceptional acid folic supplementation, alcohol usage, previous termination(s) of pregnancy, number of previous second trimester miscarriage(s), location, content and dimensions of subchorionic hematoma, amnion size (mm)

*Some scans were performed after the first trimester

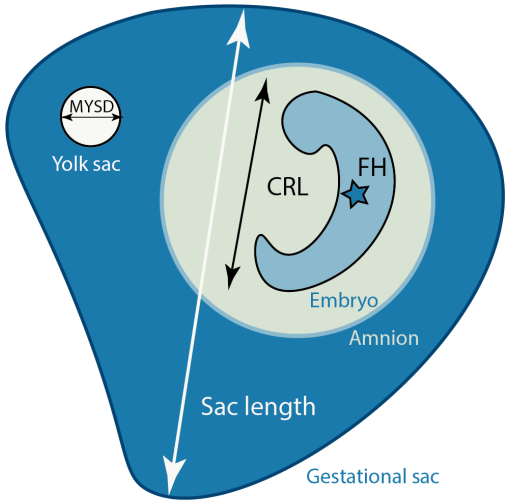


Figure 3.2: Main sonographic features measured in the first trimester of pregnancy. The mean sac diameter (MSD) is computed as the mean of the diameters of the gestational sac in the three axes, i.e. (length + height + width)/3. CRL: crown-rump length, FH: fetal heartbeat, MYSD: mean yolk sac diameter.

the presence of fetal heart beat (FH) and the presence of the amnion. The absence of the embryo was encoded as a CRL of 0. A visual description of the main ultrasound markers is displayed in Figure 3.2.

Blood biomarkers For a subset of randomly selected patients, blood samples were taken at the time of visit. The following maternal serum markers were measured: progesterone, β-human chorionic gonadotropin (β-hCG), estradiol (E2), α-foeto-protein (α-FP) and cancer antigen-125 (CA-125). Those blood biomarkers were selected for their acknowledged or potential predictive value [116]. For instance, low values of progesterone and β-hCG during the first trimester are known to be associated with lower chances of viability [117, 118]. Estradiol is a steroid hormone whose serum concentration increases throughout the pregnancy due to placental production. Although its role is still unclear, a study on baboons suggested a crucial role of estradiol in maintaining pregnancy [119]. A meta-analysis emphasized the potential predictive value of estradiol for first trimester viability studies [116]. Some studies have found a potential interest in CA-125, a non-specific cancer marker, to predict first trimester

viability [120–123]. However, the clinical utility of dosing CA-125 in early pregnancy has not been sufficiently evaluated. α-FP is a protein produced by the yolk sac and later on by the fetal liver. An increased α-FP value is associated with several fetal malformations such as neural tube defects [124] and Down syndrome [125]. Its potential association with pregnancy loss has been discussed in the literature with contradictory findings [126–128]

3.3 Predictive models development

Although adding blood biomarkers can potentially improve the predictive accuracy of clinical models [108, 109], additional blood tests are traditionally not part of the clinical routine in most countries. The need for biomarker-free models remains therefore important for a large clinical deployment. The following section describes the development of two different models: one model including solely commonly measured clinical and ultrasound variables and another model, developed on a subset of the EPOS data, which leverages blood biomarkers. The advantages of adding blood biomarkers are discussed in the last section.

Objective

Develop accurate and calibrated predictive models for first trimester viability based on early pregnancy data.

3.3.1 Clinical models

This section focuses on the development of predictive models based on common clinical and ultrasound variables measured during the first trimester of pregnancy.

Methodology

The main outcome was defined as *pregnancy viability* at 14 weeks of gestation. For each patient, the first scan demonstrating an intrauterine pregnancy was selected (PUV or viable pregnancy).

Data subset Women who underwent termination of pregnancy (n=28), withdrew from the study (n=7), and were lost to follow-up (n=40) were excluded from the analysis. Women whose LMP was unknown or not accurate (certainty score less than 7/10) were also excluded (n=223). The final analysis (n=1105) compared first trimester miscarriages (n=160) with viable control pregnancies (n=945). A flowchart of the inclusion process is displayed in Figure 3.3. The descriptive statistics for patient characteristics, first trimester symptoms and ultrasound variables for those 1105 patients are reported in Table 3.A.1. Comparison of the cohort characteristics with regard to first trimester

viability was assessed with the Student's t-test for continuous variables and the χ^2 test for binary or categorical variables.

A restricted set of potential predictors was first defined by expert knowledge and previously published studies (referenced in Section 3.1) to lower the risk of overfitting the data in the multivariate model [129]. This set included : *maternal age, maternal ethnicity, BMI, number of previous miscarriages, folic acid supplementation, smoking status, CRL, MSD, FH, MSD*FH, worst bleeding score, PUQE score, GA by LMP* and *discrepancy in GA*. The variable MSD*FH is an interaction term combining the presence of a fetal heartbeat (FH) and MSD measurements and has been implemented in a previous published model [106]. We used the GA by MSD, not CRL so that we could use this variable for PUV as well as viable pregnancies. Discrepancy in gestational age (GA) was also examined. Expected GA is based on the LMP and ultrasonography estimated GA uses a formula combining the CRL and MSD measurements to create an ultrasound-mean. Using both CRL and MSD to estimate the GA allowed us to include more patients than using CRL alone. Little difference has been observed between CRL alone and ultrasound-mean in terms of calculating estimated gestation [130].

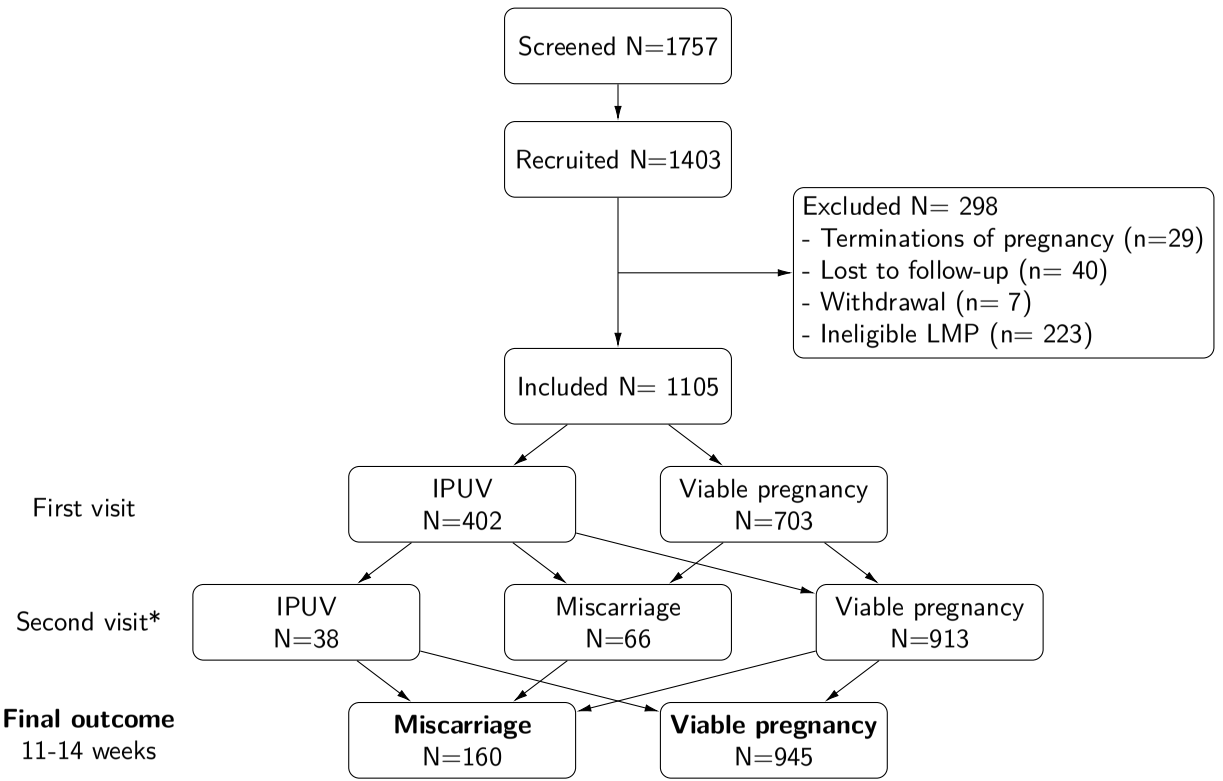


Figure 3.3: Flowchart of the inclusion process for the development of clinical models.
* 88 patients with an ultimately viable pregnancy had no second scan.

Missing values The presence of missing values was handled with multiple imputation [75] using 20 imputed data sets. The number of missing values per variable is reported in Table 3.A.1. The missing values were estimated with predictive mean matching. Data were assumed to be missing at random.

A convenient approach to analyse variable missingness is to display the variables on an influx-outflux plot [75]. These indices are based on pairs of variables (X_j, X_k) and indicate how incomplete and complete variables relate to each other. The influx coefficient I corresponds to the number of variable pairs (X_j, X_k) where X_j is missing and X_k is observed, divided by the total number of observed data. Completely observed variables have an influx of 0, while completely missing variables have an influx of 1. For a given missingness proportion, variables with a higher influx are easier to impute [75].

The outflux coefficient O corresponds to the number of variable pairs (X_j, X_k) where X_j is observed and X_k is missing, divided by the total number of missing data. The outflux index constitutes therefore a good indicator on the usefulness of the variable X_j to impute the other variables. The input-output plot is displayed in the Figure 3.4.

After applying the inclusion criteria described above, none of the variables had unworkable missingness on the remaining subset of patients. The variable *amnion* had the worst combination of outflux-influx. Although its proportion of missing values (25%) remained manageable, its predictive power is potentially limited. The variable was therefore excluded from further analysis.

Feature transformation To lower the influence of extreme values on the regression coefficients, MSD and the discrepancy in GA were winsorized using percentile 0.01 and 0.99, i.e. extreme values were truncated to the 0.01 or 0.99 percentiles value [131]. The other variables did not require winsorizing as their distribution did not demonstrate extreme values. The need for non-linear transformation of continuous predictors was assessed in a univariate setting with the likelihood ratio tests.

Univariate analysis The association with viability of each variable was tested with odds ratios and p-values from logistic regression models with a single independent variable. The discrimination performance was assessed with the area under the curve (AUC) of ROC curves and corrected for optimism using enhanced bootstrap (B=2000, see below Internal Validation).

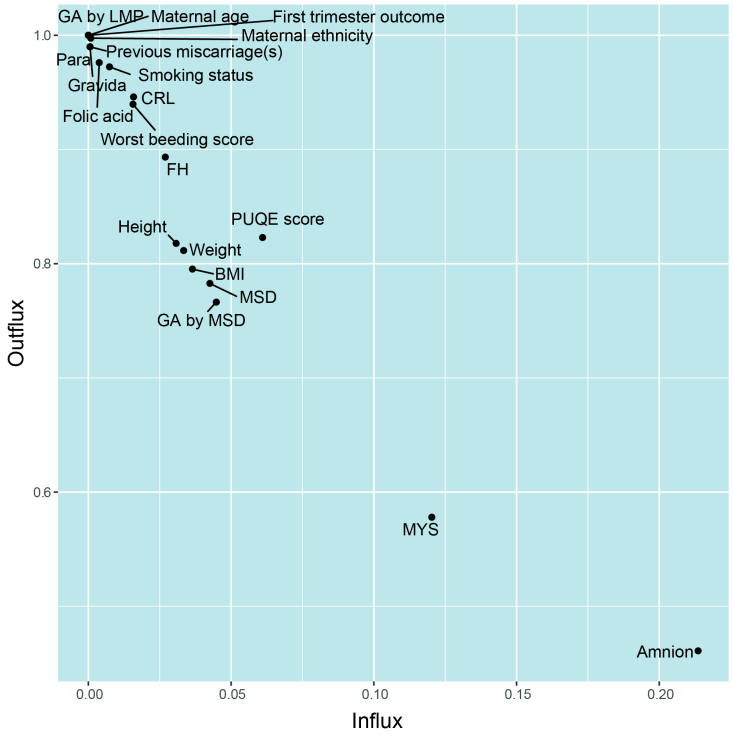


Figure 3.4: Influx-outflux plot of the variables used during model development. The influx represents how the missing data of a variable relate to the observed data of the other variables. The outflux is indicative of the usefulness of a variable to impute the other variables. Most of the variables had excellent influx-outflux combinations suggesting an efficient imputation. The variables *amnion* and *MYS* demonstrated a lower outflux indicative of a lower predictive power. Note that auxiliary variables are also present. These additional variables were not included in the predictive model but were used to improve the imputation process.

Multivariable models In order to properly validate our models given the limited number of patients, we restricted the choice of the predictive models to logistic regression models (see Box) with prior knowledge. Under the logistic regression setting, internal validation based on the enhanced bootstrap procedure to correct for optimism can be applied [132] (see Internal Validation below). As a result, the predictive models can be trained on the whole set of patients, which increases the statistical power compared to hold-out validation procedures [129]. This method is particularly suited for logistic regression models[3].

Multivariable logistic regression was used to build two predictive models for first trimester viability. The first one included all the predefined variables reported in Table 3.2, irrespective of their univariate performance, aiming to maximize the predictive power. To reduce the number of variables and to facilitate ease of clinical use, a second model was developed using a robust feature selection method described below.

Feature selection We performed feature selection with stability selection with bootstrap imputation (BI-SS) (with parameters alpha=0.5, selection threshold =0.7, the number of bootstrap samples B =2000) [137]. This method uses randomized LASSO over each bootstrap sample with a single imputation and keeps track of the variables selected in this L1 regularization setting. The final set of selected variables contains all variables whose inclusion proportion is above the predefined selection threshold. This selection process is summarized in the Algorithm 1. This approach alleviates the instability inherent to a single run of LASSO. Ultimately, a more parsimonious model was fitted using the selected subset of predictors.

Internal validation The enhanced bootstrap procedure [129, 132] was used as internal validation instead of the traditional training-test sets approach to take advantage of the whole dataset during the modelling process. This method computes the optimism needed to correct the apparent performance indices that result from evaluating a model on its training data The pseudo-code for bootstrap optimism correction is reported in Algorithm 2. To combine the enhanced bootstrap with the multiple imputation framework, we opted for single imputation nested in the bootstrap procedure [138]. Practically, we sampled with replacement 2000 datasets from the original data and then performed a single imputation on each bootstrap dataset. Due to the heavy computational cost of BI-SS, the feature selection step was not reproduced on each sample of

[3]Some studies demonstrated however that enhanced bootstrap fails to accurately assess the optimism of flexible machine learning approaches as the adjusted results remain overoptimistic even after correction [133, 134].

The logistic regression model

Logistic regression is a statistical model used to perform regression analyses on binary outcomes. Because many medical outcomes are binary (e.g. presence/absence of a disease), logistic regression models are widely used in medicine.
Generally speaking, a *regression* analysis estimates the association between a *dependent* variable and one or several *independent* variable(s). The general formula for multivariable regression models is given by :

$$\boldsymbol{y}_i = f(\boldsymbol{X}_i\boldsymbol{\beta}) + \boldsymbol{e}_i \tag{3.1}$$

where $\boldsymbol{y}_i$ is the dependent variable, i.e. the predicted outcome or response, $\boldsymbol{X}_i$ is the vector of independent predictors, also known as explanatory variables, for patient i. $\boldsymbol{\beta}$ is the vector of parameters optimized during model training, f is a pre-specified function that models different distributions of responses and $\boldsymbol{e}_i$ is the error term, also known as *residual*. Those models rely on a linear combination of variables $\boldsymbol{X}_i$ with the learned coefficients $\boldsymbol{\beta}$. The function f can model non-linear relations between the predictors and the outcome variable(s). More specifically, logistic regression is a *generalized linear model* with the *logit* link function and is defined as :

$$logit(\boldsymbol{p}_i(y=1)) = \boldsymbol{X}_i\boldsymbol{\beta}, \text{ where } \boldsymbol{p}_i(y=1) \in [0,1] \tag{3.2}$$

The *logit* link function defined as :

$$logit(p) = log(\frac{p(y=1)}{1-p(y=1)}) \tag{3.3}$$

where,

$$\boldsymbol{p}_i(y=1) = \frac{exp(\boldsymbol{X}_i\boldsymbol{\beta})}{(1+exp(\boldsymbol{X}_i\boldsymbol{\beta}))} \tag{3.4}$$

The *logit* function of Eq. 3.3 guarantees to produce outcomes within the $[0,1]$ interval, ideal to estimate probabilities.
As a generalized linear model, logistic regression does not require a normal distribution of the residuals. In addition, unlike ordinary linear regression, logistic regression models do not rely on homoscedasticity. Maximum likelihood techniques are usually used to estimate the coefficients $\boldsymbol{\beta}$.
The additive constraint of multivariable regression restricts the model capacity but facilitates the understanding of the prediction process. To overcome the lack of flexibility from the linear combination of variables, additional complexity can be introduced with interactions and/or non-linear transformations of the independent variables. Unlike, data-driven machine learning alternatives, those model specifications must often be manually specified upfront by the statistician [129, 135], although some (semi)-automatic procedures exist (e.g. multivariable fractional polynomial [136]).

Algorithm 1: stability selection with bootstrap imputation (BI-SS).
$(\mathbf{X}, y)$ is the incomplete dataset with outcome y, B is the number of bootstrap samples, Λ is a set of feasible regularizators λ values, α is a sampling parameter $\in (0.2, 0.8)$, π is the inclusion threshold chosen $\in (0.6, 0.9)$ and $\hat{S}_\pi$ is the final set of selected variables

Input: $(\mathbf{X}, y)$
parameters: B, Λ, α, π
Output: $\hat{S}_\pi$

begin
- $S = \emptyset$
- **for** *b= 1,...,B* **do**
  1. Draw a bootstrap sample $(\mathbf{X}^{(b)}, y^{(b)})$ from $(\mathbf{X}, y)$
  2. Impute $\mathbf{X}^{(b)}$ as $\mathbf{Z}^{(b)}$ with a single imputation
  - **for** λ *in* Λ **do**
    1. Fit a randomized LASSO estimator parametrized by α on $(\mathbf{Z}^{(b)}, y^{(b)})$:
       $$\hat{\beta}_\lambda^{(b)} = \underset{\beta}{\operatorname{argmin}}(\|y^{(b)} - \mathbf{Z}^{(b)}\beta\| + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\omega_j^{(b)}}),$$
       where $\omega_j^{(b)}$ is i.i.d in $[\alpha, 1]$
    2. Get $\hat{S}_\lambda^{(b)}$, the active set of variables (non-zero parameters estimates) from $\hat{\beta}_\lambda^{(b)}$ and append it to S
  - **end**
- **end**
- Compute the final active set of variables as:
  $$\hat{S}_\pi = \left\{ j : \max_{\lambda \in \Lambda}(\Pi_j^\lambda) \geq \pi \right\} \text{ where } \Pi_j^\lambda = \frac{1}{B} \sum_{b=1}^{B} I(j \in \hat{S}_\lambda^{(b)})$$

end

the enhanced bootstrap for internal validation. However, the nature of BI-SS, including a heavy resampling scheme, makes it more stable and less prone to overfit the training data compared to other feature selection mechanisms, which need to be repeated within each bootstrap sample. The discrimination ability of the models was estimated with the optimism-corrected AUC (and 95% CI) and the calibration was estimated with non-parametric calibration curves, the optimism-corrected calibration slope and the optimism-corrected estimated calibration index (ECI) [139].

Decision curve analysis The clinical usefulness of the models was compared with decision curves and net benefit (see Box). Decision curves are mostly used in settings involving well-defined therapeutic interventions (e.g. performing surgery in case of cancer) [140]. Unfortunately, in the first trimester, no such therapeutic intervention exists. However, we can evaluate the impact of our models on the need for subsequent monitoring of patients predicted at risk of miscarriage. Although our models were trained to predict viability, we reported the decision curves for predicting miscarriage, as opposed to the generic policies of subsequent monitoring for all patients and an absence of any follow-up, i.e. no monitoring for any patient. Obviously, a model that performs even slightly better than a random guess will provide some benefit over such generic policies. However, depending on the discriminative and calibration performance of a model, the net benefit will be more or less substantial. More importantly, the net benefit of useful models will cover a larger range of clinically relevant decision thresholds.

Consecutive scans To utilise the longitudinal sonographic data, we built models on patients who had at least two consecutive scans without demonstrating a miscarriage. New aggregated features were constructed to summarize longitudinal features from both scans. To provide sound performance comparisons, we fitted three models on the same subset of patients: a model trained on first scan only, a model trained on second scan only and a model trained using aggregated features from both scans (Table 3.A.2). These models were trained on a subset of 951 patients of whom 94 had an outcome of miscarriage and 857 had a viable pregnancy at 11-14 weeks (Table 3.A.2). The average GA by LMP at the time of the first scan was 47.9 days and 60.4 days at the time of the second scan (Table 3.A.2).

All analyses have been performed with R 3.6.0 (R Core Team 2017) [144].

Clinical usefulness : Decision Curve Analysis

Even if a model is accurate and well-calibrated, it does not mean it is clinically useful [142]. A common tool to estimate the impact of a model in clinical practice is the *decision curves analysis* [143].
Because it is difficult to define a single optimal cut-off applicable for every patients, it is of interest to evaluate the performance of a model over a large range of clinically attractive decision thresholds.

Decision curve analysis builds upon positives cases, i.e. the cases that would be considered positive given a cut-off threshold. Medical decisions following the identification of a positive case can be justified for a true positive (TP) but are unnecessary, if not harmful, for a false positive (FP). Hence, decision curves rely on the notion of *net benefit*, which is defined as : $\frac{(TP-wFP)}{N}$, where w is a threshold–dependent weight given by $\frac{p}{(1-p)}$, i.e. the odds of the probability threshold. The net benefit estimates the benefit of correctly treating those who really have the disease (true positive) minus the cost of treating those who are erroneously diagnosed to have the disease (false positive)[a]. As specified by Vickers [140] : *"Net benefit differs from accuracy metrics such as discrimination and calibration because it incorporates the consequences of the decisions made on the basis of a model or test".*

Decision curves plot the net benefit of a model over a range of possible decision thresholds and compared the curve to two default intervention strategies, namely *treat all* and *treat none*. Those default interventions assume that all cases are either positive (treat all) or negative (treat none), irrespective of the model outputs (e.g. perform a surgery for every patients, even if the model predict a very low score of malignancy for some of them). Decision curves analysis is also useful to compare the clinical benefit of different models on the same plot.

[a]Note that additional costs can be subtracted from the net benefit formula, e.g. in case of invasive procedures to gather additional data.

Algorithm 2: Bootstrap correction for optimism
For each bootstrap sample $\boldsymbol{X}^{(b)}$ of the original data $\boldsymbol{X}$, a model $M^{(b)}$ is fitted to $\boldsymbol{X}^{(b)}$. For a given performance metric C, the performance of $M^{(b)}$ is measured for both $\boldsymbol{X}^{(b)}$ and $\boldsymbol{X}$, leading to $C_{orig}^{(b)}$ and $C_{boot}^{(b)}$. The overall optimism O_C is computed as the average between $C_{boot}^{(b)}$ and $C_{orig}^{(b)}$ over B and subtracted from the apparent performance C_{app} to produce an optimism-corrected performance C_{corr}. Note that pre-processing steps such as feature selection or feature transformation should be repeated at each iteration b [141].

Input: $\boldsymbol{X}, \boldsymbol{y}$
parameters: B, model specifications, choice of C
Output: C_{corr}

Train a model M on $\boldsymbol{X}$ and estimate C_{app} the apparent performance of M on $(\boldsymbol{X}, \boldsymbol{y})$
for *b =1, 2,...,B* **do**
    Take a bootstrap sample $\boldsymbol{X}^{(b)}$ from $\boldsymbol{X}$
    Optional : pre-processing steps (feature selection, feature transformations,...)
    Train a model $M^{(b)}$ on $\boldsymbol{X}^{(b)}$ and estimate its performance C_{boot}^{b} on $(\boldsymbol{X}^{(b)}, \boldsymbol{y}^{(b)})$
    Estimate $C_{original}^{b}$, the performance of $M^{(b)}$ on $(\boldsymbol{X}, \boldsymbol{y})$
end
Estimate the optimism $O_C = \frac{1}{B} \sum_{1=b}^{B} (C_{boot}^{b} - C_{original}^{b})$
Adjust the apparent performance $C_{corr} = C_{app} - O_C$
return C_{corr}

Results

Univariate analysis The odds ratios, p-values and the AUC from the univariate analysis are reported in 3.2. The majority of the preselected variables demonstrated good predictive value. The five most discriminant variables were MSD*FH (AUC: 0.783, 95% CI: 0.764-0.802), CRL (AUC: 0.737, 95% CI: 0.703-0.771), discrepancy in GA (AUC: 0.735, 95% CI: 0.698-0.772), MSD (AUC:0.726, 95% CI: 0.687-0765) and FH (AUC: 0.715, 95% CI: 0.677-0.752). The significance of GA by LMP was borderline (p-value=0.04 with AUC: 0.549, 95% CI: 0.499-0.598) which is the result of similar GA for both outcomes at first scan. Out of the baseline variables, maternal age was identified as the main non-modifiable predictor for miscarriage (AUC: 0.602, 95% CI: 0.552-0.652). It is common knowledge that rates of first trimester miscarriage increase with advancing maternal age [98, 145]. BMI also demonstrated some predictive value (AUC:0.572 95% CI: 0.528-0.616). This results corroborated the finding of previous large cohort studies, which identified obesity as a risk factor for miscarriage [146, 147].

Interestingly, the supplementation in folic acid was also significantly associated with the first trimester viability (OR: 2.63, 95% CI: 1.54-4.49; p-value < 0.0001). The World Health Organisation recommends all women take folic acid 400µg prior to conception until 12 weeks' gestation. This recommendation exists to reduce the incidence of neural tube defects such as anencephaly [148]. While the impact of folic acid supplementation on neural tube defects is clear, the impact on first trimester miscarriage remains controversial. Low plasma folate levels increase the incidence of miscarriage [149]. However a large population based study did not show a difference in miscarriage between women who were taking folic acid supplementation and those who did not [150]. Other studies have shown an association between lower folic acid intake and miscarriage [151–153]. Another possible explanation for this association in our study is that folic acid use is a surrogate for general health and wellbeing.

Table 3.2: Summary of all potential variables included in the predefined variable selection. The odds ratios, p-values and the AUC result from univariate logistic regression models. The AUC is corrected for optimism using enhanced bootstrap (B=2000).

Variables	**OR (95% CI)**	**p-values**	**AUC (95% CI)**
Baseline variables			
Maternal Age	0.93 (0.90-0.97)	<0.0001	0.602 (0.552-0.652)
Maternal ethnicity		<0.001	0.509 (0.468-0.549)
Intercept	6.00 (3.47-10.36)		
Asian	1.44 (0.67-3.10)	0.35	
Caucasian	0.95 (0.53-1.70)	0.86	
Mixed	0.67(0.26-1.72)	0.40	
Others	0.93(0.38-2.27)	0.87	
BMI	0.97 (0.94-1.00)	0.09	0.572 (0.528 -0.616)
Number of previous miscarriages	0.87 (0.76-1.00)	0.05	0.529 (0.479-0.579)
Folic acid suppl. (yes or no)	2.63 (1.54-4.49)	<0.001	0.554 (0.517-0.591)
Smoking status (yes or no)	0.71 (0.41-1.22)	0.21	0.516 (0.487-0.545)
Longitudinal variables			
CRL (mm)	1.19 (1.13-1.25)	<0.0001	0.737 (0.703-0.771)
MSD (mm)	1.10 (1.07-1.12)	<0.0001	0.726 (0.687-0.765)
FH (presence or absence)	6.12 (4.18-8.98)	<0.0001	0.715 (0.677-0.752)
MSD*FH (interaction effect between MSD and FH)	1.23 (1.15-1.31)	<0.0001	0.783 (0.764-0.802)
Worst bleeding score at initial scan	0.93 (0.78-1.11)	0.42	0.502 (0.453- 0.551)
PUQE score	1.23 (1.10-1-37)	<0.0001	0.630 (0.592-0.668)
GA difference (days)	0.91 (0.89-0.93)	<0.0001	0.735 (0.698-0.772)
GA by LMP (days)	1.02 (1.00-1.03)	0.04	0.549 (0.499-0.598)

Table 3.3: Apparent and corrected predictive performance of the complete and reduced clinical models. Reducing the number of variables in the model did not drastically affect its performance, as reflected by similar AUC and calibration indices. After correction for optimism, the discrimination and the calibration of the models remained excellent.

	Apparent metric	**Metric corrected for optimism***
Complete clinical model		
AUC	0.855 (0.826-0.883)	0.837 (0.807-0.868)
Calibration slope	1.013 (0.829-1.197)	0.943 (0.763-1.122)
ECI	0.005 (-0.063 - 0.074)	0.036 (-0.032 - 0.104)
Reduced clinical model		
AUC	0.841 (0.811-0.871)	0.836 (0.806-0.866)
Calibration slope	1.007 (0.826-1.188)	0.969 (0.799- 1.149)
ECI	0.004 (-0.064 - 0.073)	0.018 (-0.050 - 0.087)

*using the enhanced bootstrap with B=2000

Complete model (single scan) The complete clinical model including the whole set of predefined variables demonstrated an apparent AUC of 0.855 (95% CI 0.826-0.883), and AUC of 0.837 (95% CI 0.807-0.868) when corrected for optimism. The optimism-corrected calibration slope of this model was 0.943 (95% CI 0.763-1.122) and the optimism-corrected ECI was 0.036 (95% CI -0.032 - 0.104) (Table 3.3). The non-parametric calibration curves are displayed on Figure 3.5. The corresponding distributions of the model's predictions are displayed in the histograms in Figures 3.6. The vast majority of viable pregnancies are accurately and confidently predicted (see for instance the amount of viable pregnancies predicted with a probability > 0.70). A proportion of miscarriages remains unpredictable and is erroneously predicted as viable pregnancies, with a high degree of confidence (with probability > 0.75).

Reduced model (single scan) After feature selection using BI-SS, a parsimonious model was built using maternal age, folic acid use, the GA difference and MSD*FH. The results of the stability selection based on BI-SS is displayed on Figure 3.7. This reduced clinical model demonstrated similar discrimination performance (AUC 0.841; 95% CI 0.811-0.871: optimism-corrected AUC 0.836; 95% CI 0.806-0.866). The optimism-corrected calibration

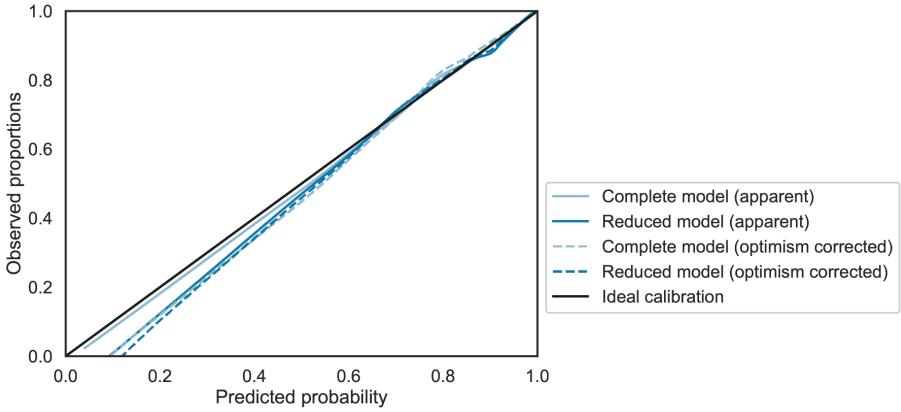


Figure 3.5: Apparent and optimism-corrected non-parametric calibration curves of the complete and reduced clinical models. Both models demonstrate excellent calibration performance, even after correction for optimism.

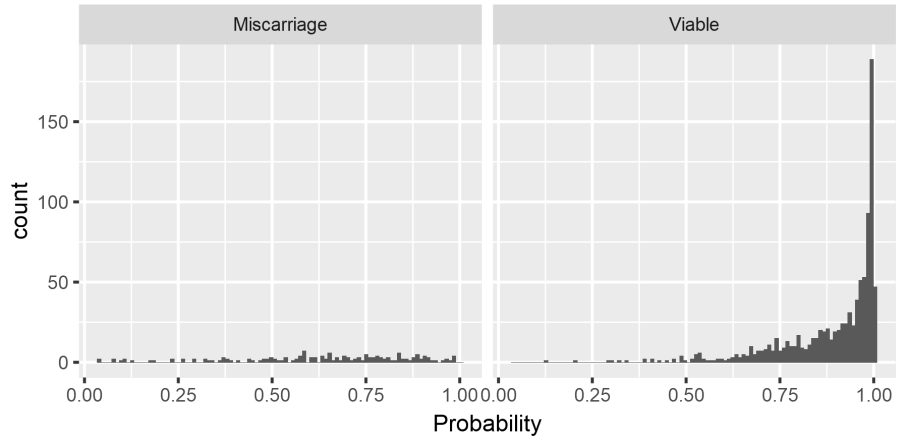


Figure 3.6: Distribution of the complete model's risks stratified per outcome. The vast majority of viable pregnancies are accurately and confidently predicted, as demonstrated by the large numbers of viable cases predicted with a probability greater than 0.70. A proportion of miscarriages remains unpredictable and is erroneously predicted as viable pregnancies, with a high degree of confidence (with probability > 0.75). Overall, the predicted probabilities are not restricted to extreme predictions and spread on a large range of values which translates the good calibration performance of the model (Table 3.3).

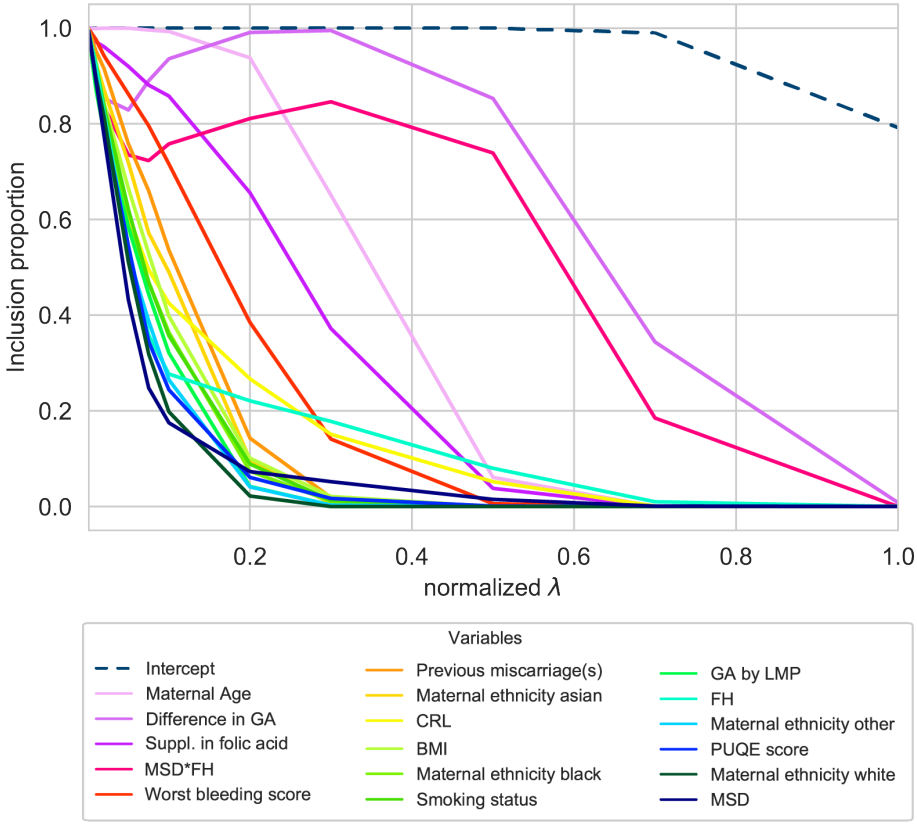


Figure 3.7: Stable feature selection with BI-SS.
Most of the variables are rapidly discarded (inclusion proportion dropping to zero) while four variables emerge clearly : *MSD*FH, maternal age, discrepancy in GA* and *folic acid supplementation*. A potential selection of *worst bleeding score* is more ambiguous.

slope was 0.969 (95% CI 0.799- 1.149) and the optimism-corrected ECI was 0.018 (95% CI -0.050 - 0.087) (Table 3.3). The non-parametric calibration curves are displayed on Figure 3.5. The distributions of the second model's predictions are displayed in Figure 3.8. Reducing the number of variables in the model did not drastically affect the distributions of the model predictions, as reflected by similar calibration curves, optimism-corrected calibration slopes and ECI for both models (Table 3.3). Reducing the number of variables did not impact the clinical usefulness of the model as demonstrated by the overlapping decision curves of both models in Figure 3.9.

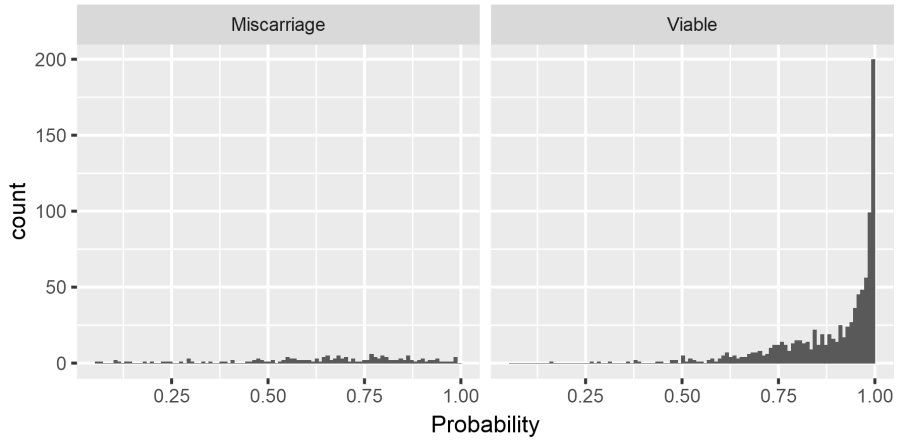


Figure 3.8: Distribution of the reduced model risks stratified per outcome. When compared to Figure 3.6, the distribution of the predicted risks is not drastically affected by the feature selection process.

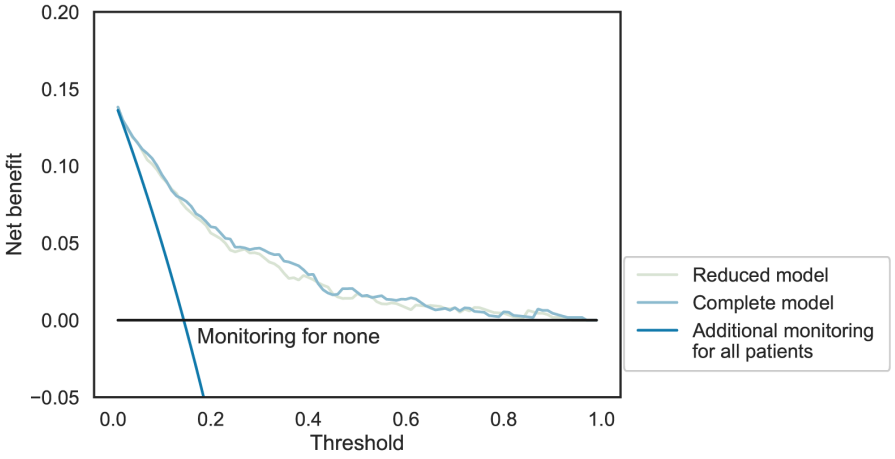


Figure 3.9: Decision curves of the complete and reduced clinical models. Both models demonstrated similar net benefit over the entier range of thresholds compared to generic policies.

Table 3.4: Predictive performance of clinical models trained on consecutive visits.
All metrics are corrected for optimism with enhanced bootstrap (B=2000). Aggregating information over the last two visits was not more informative than using the last visit only. The discrimination and the calibration performance of both models remain similar between models using only the second visit and models aggregating the first two visits.

	First visit	**Second visit**	**First and second visits**
Pre-specified models			
AUC (95% CI)	0.823 (0.784-0.862)	0.905 (0.876-0.934)	0.908 (0.877-0.938)
Calibration slope	0.821	0.878	0.869
Reduced models			
AUC (95% CI)	0.742 (0.692-0.792)	0.903 (0.870-0.936)	0.890 (0.849-0.930)
Calibration slope	0.981	1.070	0.995

Models trained on consecutive scans In the models using pre-specified variables, the corrected AUC was 0.823 (95% CI 0.784-0.862) using first scan data, 0.905 (95% CI 0.876-0.934) using second scan data and 0.908 (95% CI 0.877-0.938) using a combination of the two scans. After stability selection, the reduced model trained on second scan data included the following variables: *maternal age, the GA difference between LMP and MSD at the second scan, the GA by LMP at the second scan*, and *MSD*FH at the second scan.* The reduced model trained on aggregated features included: *maternal age, the maximum GA difference between LMP and MSD between first and second scan, MSD growth rate, CRL growth rate, FH at both scans summarized in a unique variable.* The corrected AUC was 0.903 (95% CI 0.870-0.936) using the second scan data, and 0.890 (95% CI 0.849-0.930) with a combination of variables from the two scans. A summary of the results is reported in Table 3.4 . There was no benefit in terms of predictive performance of using a combination of the first two scans compared to using the second scan only. Use of longitudinal data to account for growth of CRL and MSD does not appear to improve performance of the model. This is likely due to the fact that half of pregnancies that subsequently miscarry have a similar growth rate to viable pregnancies.

Conclusion

We have developed and internally validated two clinical models that enable accurate prediction of first trimester viability and is applicable to a large number of women seen within the early pregnancy unit (all confirmed intrauterine pregnancies, prior to the dating scan at 11-14 weeks). A parsimonious model based on 6 easily available predictors (*maternal age, folic acid supplementation, MSD, FH,* and *GA by LMP*) demonstrated excellent discrimination and calibration performance, even after correction for optimism. A more complex model based on an extensive set of predictors did not demonstrated clinical benefit over a this simpler approach. Our models use the updated criteria for miscarriage diagnosis which supersedes many pre-existing models. Use in the clinical setting may allow for better psychological preparedness when a poor outcome (miscarriage) is anticipated. It may also allow for improved resource allocation in units where a positive outcome is expected and follow-up not necessarily required.

3.3.2 Augmented clinical models with blood biomarkers

Blood biomarkers are costly to acquire and process and, as such, are not part of the first trimester clinical routine. However, they potentially carry a higher predictive power than traditional clinical and ultrasound features. These variables were only measured in a subset of randomly selected women in the EPOS database. Therefore, additional models trained on the appropriate subsets of data were developed.

Methodology

Data subset In order to maximize the use of the original dataset, model development was performed on a subset of data selected to include all scans containing at least 2 biomarker values. This strategy constituted a trade-off between complete case analysis that discards relevant information and an inefficient imputation due to a large proportion of missing values ($> 60\%$). Using this additional selection criteria, out of the 1105 pregnancies previously selected in Section 3.3 (160 miscarriages and 945 viable pregnancies), 428 pregnancies (116 miscarriages and 328 viable pregnancies) were included for the multivariable modelling with blood biomarkers (Figure 3.10). The list of preselected variables defined in (Table 3.2) was augmented by the following biomarkers : *progesterone, β-hCG, CA-125* and *α-FP*. The estradiol variable, although potentially relevant in the pregnancy context, had a too large proportion of missing values (>80%)

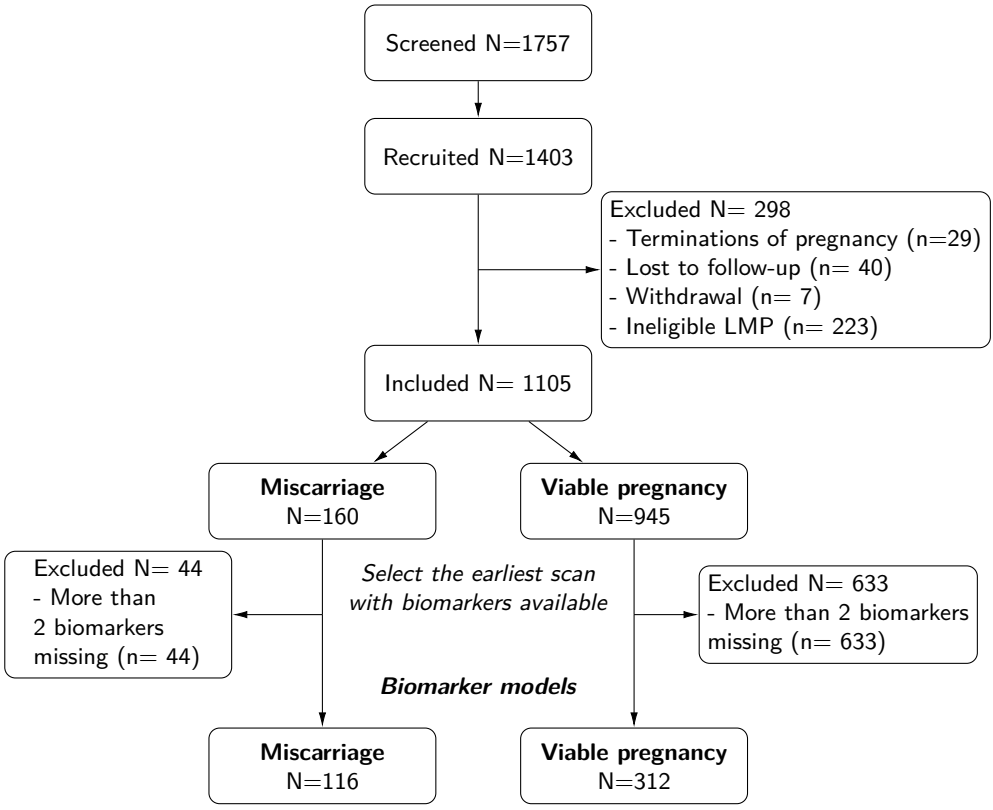


Figure 3.10: Flowchart of the inclusion process for the development of augmented clinical models.

was therefore omitted from further analysis. The descriptive statistics of this specific data subset for patient characteristics, first trimester symptoms, ultrasound variables and blood biomarkers are reported in Table 3.A.3.

Predictive performance of individual biomarkers Unlike traditional ultrasound measurements, which mostly reflect the degree of pregnancy advancement irrespective of maternal features, blood biomarkers are affected by additional maternal characteristics such as BMI or maternal age [154, 155]. To account for those confounding factors during the evaluation of each biomarker's predictive performance, we computed the predictive performance of each biomarker in a relative setting. A baseline model including relevant confounding variables was

first fitted the data. The same model was then augmented with the biomarker of interest. The baseline model was a multivariable logistic regression model composed of the following variables : *maternal age, maternal ethnicity, BMI* and *GA by LMP*. The assessment of each biomarker predictive power was reported with the AUC gain, resulting from the addition of the biomarker to the baseline model. Because AUC is a rank-based metric, AUC gain is a low-power procedure and does not sufficiently account for the degree of confidence in model' predictions. As a result, adding a new variable could improve a model's calibration without affecting the raking of instance [129, 156]. Therefore, we also reported the fraction of novel information (FNI) [157], which assesses the change in predictions variance between the two models, computed as :

$$FNI = 1 - \frac{var(P_{pre})}{var(P_{post})} \tag{3.5}$$

where $var(P_{pre})$ corresponds to the variance in risks obtained from a baseline model and $var(P_{post})$ is the variance in risks obtained from the same model augmented by a new marker.

Transformation of biomarkers Many biomarkers are not normally distributed (e.g. exponential distribution) and demonstrate a nonlinear relationship with outcome risks. For instance, the change in odds ratio occuring after an increase in biomarker value from 0 to 10, does not necessarily correspond to same change in risks resulting after an increase from 100 to 110. To overcome this non-linear behavior, some biomarkers have been transformed to make the relation to the risk of miscarriage more linear[4]. The need for biomarker transformation was based on the likelihood ratios (LR). Additionally, winsorization, i.e. the truncation of extreme values above the percentile 0.99, was performed to limit the influence of extreme values on the regression coefficients. The Figure 3.11 displays examples of non-linear transformations for the biomaker CA-125 with the corresponding LR.

Multivariable models Similarly to the aforementioned analysis of individual biomarkers, the added value of biomarkers in a multivariable model was assessed in a relative setting, i.e. we compared the model performance when it is trained with and without the biomarker variables. Therefore, two predefined sets of variables were used for model development : 1) the same set of variables as reported in Table 3.2 and, 2) a set of variables corresponding to 1), augmented by the following biomarkers: *progesterone, β-hCG , α-FP* and *CA-125*. Additional

[4]Note that this transformation is not performed to ensure normality, as is not required in this context.

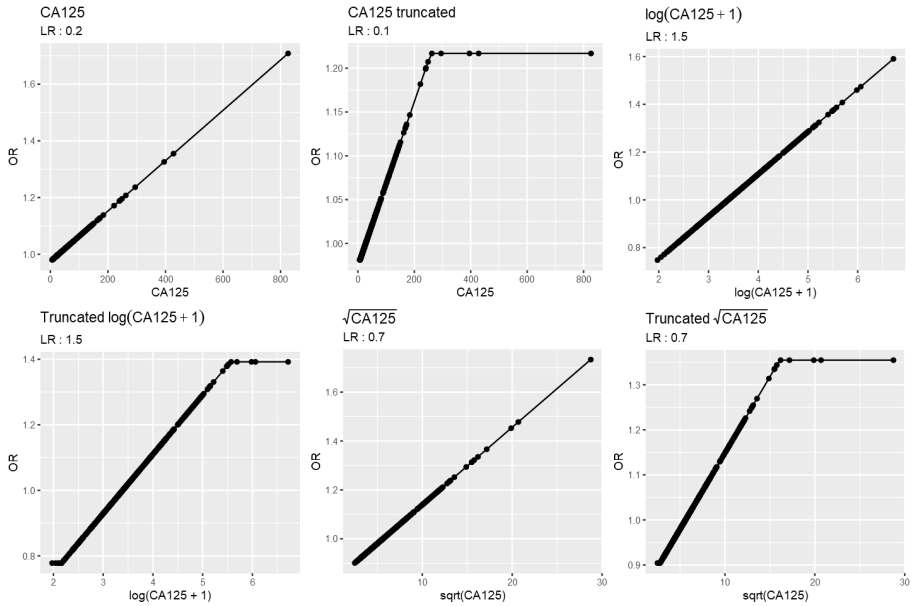


Figure 3.11: Example of various transformations applied to the biomarker CA-125. Truncation is performed is using winsorization at percentile $p99$. LR : likelihood ratio.

reduced models, with and without biomarkers, were developed using the feature selection approach based on BI-SS (see 3.3.1), with the following parameters : alpha=0.5, selection threshold =0.7, B= 2000. An overview of the methodology to assess the added value of blood biomarkers in multivariable models is displayed in Figure 3.12.

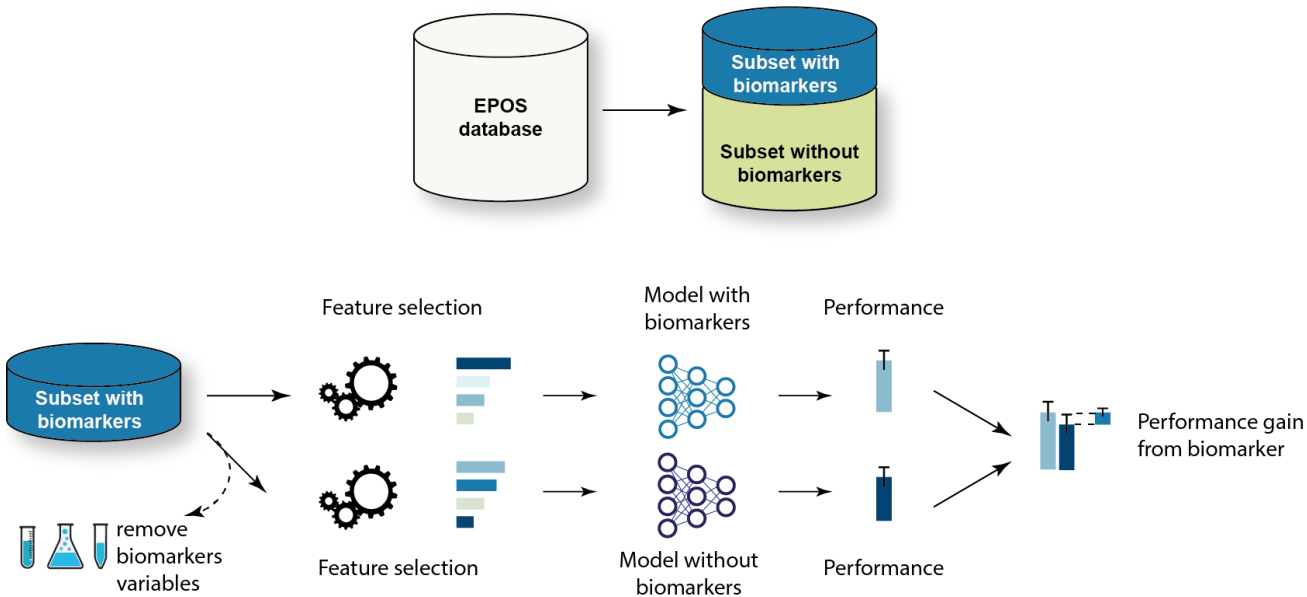


Figure 3.12: Overview of the methodology to assess the added value of blood biomarkers in multivariable models. The added value of blood biomarkers is assessed in a relative setting where the performance of models with and without biomarkers is compared. Based on the same subset of data with available biomarkers, two pipelines are developed and follow the exact same developmental steps. The only difference is that the first pipeline has access to the biomarkers as potential variables to be included by the automated feature selection method, while the second pipeline cannot make use of the biomarker variables. The two predictive models are developed in parallel and their discrimination and calibration performances are assessed. Ultimately, the gain in performance from adding the biomarkers is reported.

Internal validation and performance metrics The enhanced bootstrap for optimism-correction (with B=2000) was applied to internally validate the model's performance [129, 132]. The discrimination of the models was estimated with the optimism-corrected AUC (and 95% CI) and the optimism-corrected discrimination slope. The calibration performance was estimated with the optimism-corrected calibration slope. The AUC gain, the discrimination slope gain and the FNI were used to assess the usefulness of adding blood biomarkers to a multivariable model. The clinical usefulness of the models was compared with decision curves and net benefit.

All analyses have been performed with R 3.6.0 (R Core Team 2017) [144].

Results

Predictive performance of individual biomarkers The performance of the four individual biomarkers, as added value to a common baseline model, is reported in Table 3.5. Progesterone and β-HCG demonstrated a high predictive value, with a gain in AUC above 0.20 and a FNI above 0.80. α-FP and CA-125 did not demonstrate any added value to a baseline model in terms of discrimination. However, α-FP introduced a significant amount of new information, as demonstrated by its large FNI score (0.268).

Table 3.5: Predictive value of individual biomarkers measured in a relative setting. A baseline model including relevant confounding variables was first fitted on the corresponding data subset (the slight changes in its performance are explained by different subset of data for each biomarker). The same model was then augmented with a biomarker variable. The added value of each biomarker was measured as the gain in discriminative performance (AUC) and fraction of new information added to the model (FNI).

	Baseline	**Baseline + biomarker**	**Relative metrics**	
Biomarker	AUC (95% CI)	AUC (95% CI)	AUC gain (95% CI)	FNI (95% CI)
Progesterone†	0.589 (0.526-0.652)	0.805 (0.756-0.853)	0.216 (0.146-0.274)	0.842 (0.814 0.868)
β-hCG‡	0.599 (0.540-0.657)	0.812 (0.767-0.858)	0.213 (0.160-0.266)	0.824 (0.794-0.854)
α-FP	0.581 (0.519-0.643)	0.604 (0.542-0.666)	0.023 (-0.011-0.057)	0.268 (0.191-0.345)
CA-125¶	0.587 (0.524-0.649)	0.582 (0.519-0.645)	-0.005 (-0.007-0.003)	0.000 (-0.003-0.003)

†‡¶ The corresponding biomarkers were transformed as log(progesterone+1), β-hCG/1000, log(CA-125+1)

Multivariable models Compared to a complete model based only on clinical and ultrasound variables, introducing the progesterone and β-hCG variables resulted in a AUC gain of 0.045, a gain in discrimination slope of 0.108 and a FNI of 0.268. The change in individual probabilities after introducing biomarker predictors is displayed in Figure 3.13. The Figure 3.14 displays the predicted probabilities of both models in box plots, stratified on the outcome and the presence of FH. These plots constitute a visualization of the discrimination slope and reflect the predictions improvements in terms of discrimination after

Table 3.6: Impact of adding the biomarkers variables on the predictive performance of the clinical models
The discrimination performance is measured with AUC and the discrimination slope. The calibration is assessed with the logistic calibration slope and the ECI. All metrics are corrected for optimism with enhanced bootstrap (B=2000). Adding biomakers improves the discrimination performance of both complete and reduced models with barely no effect on the calibration. A significant proportion of novel information is introdued in both models, as demonstrated with the large FNI scores.

	Complete Models		**Reduced Models**	
Metric (95% CI)	Without BX	With BX	Without BX	With BX
AUC*	0.798 (0.756-0.851)	0.843 (0.799-0.887)	0.817 (0.773-0.861)	0.855 (0.814-0.890)
Discrimination slope*	0.255 (0.205-0.305)	0.362 (0.310-0.414)	0.267 (0.216-0.318)	0.361 (0.313-0.409)
Calibration slope*	0.797 (0.552-1.042)	0.782 (0.543-1.020)	0.930 (0.676-1.184)	0.946 (0.720-1.171)
ECI*	0.222 (-0.037-0.481)	0.135 (-0.023-0.295)	0.090 (-0.101-0.280)	0.069 (-0.100-0.239)
AUC gain	-	0.045 (0.017-0.073)	-	0.038 (0.019-0.074)
Discrimination slope gain	-	0.108 (0.081-0.125)	-	0.095 (0.060-0.122)
FNI	-	0.268 (0.189-0.347)	-	0.242 (0.215-0.269)
Calibration slope gain	-	-0.015 (-0.211-0.212)	-	0.016 (-0.240-0.241)
ECI gain	-	-0.087 (-0.210-0.211)	-	-0.020 (-0.183-0.185)

*corrected for optimism with enhanced bootstrap (B=2000).

adding biomarkers to a clinical model. The predicted risks increase in case of viable pregnancies while they decrease for miscarriages.

Mutlivariable models after feature selection After performing feature selection with BI-SS, the following variables were retained : *maternal age, difference in GA, folic acid supplementation, progesterone* and *β-hCG*. However, when the biomarkers were excluded from the complete set of variables, the feature selection process picked additional variables to compensate the absence of biomarkers variables. The reduced model without biomarkers contained the following variables : *maternal age, difference in GA, folic acid supplementation, worst bleeding score* and *MSD*FH*.

Compared to the complete models, both reduced models demonstrated better

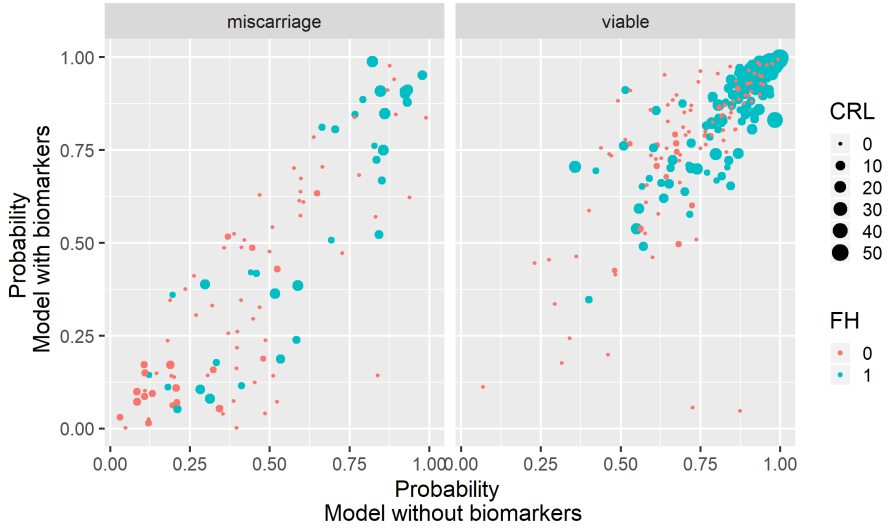


Figure 3.13: These plots visualize the changes in predicted probabilities after including the biomarkers variables in a complete model. Improvement in prediction is demonstrated by a probability increase for viable pregnancies and a probability decrease for miscarriages. The vast majority of viable pregnancies were already accurately predicted with a high confidence, even without biomarkers (most of the points are in the upper right quadrant). Introducing biomarkers resulted in significant probability changes for several miscarriage cases, as assessed by the presence of points in the lower right quadrant (pregnancies that were initially predicted as viable with a probability $p > 0.5$ but then reduced to $p < 0.5$ after introducing biomarkers information)

calibration performance (calibration slope: 0.934 vs 0.797 without biomarkers, calibration slope: 0.946 vs 0.792 with biomarkers) (Table 3.6). Similarly to the complete models, adding the biomarkers in the pool of potential variables resulted in a gain in discriminative performance (AUC gain: 0.038, discrimination slope gain: 0.095). The presence of biomarkers variables also added new information demonstrated by an FNI of 0.242. The impact on the calibration performance was minimal with a gain in discimination slope of 0.016 and an ECI gain of 0.020. The corresponding non-parametric calibration slopes are displayed in Figure 3.17. The changes in individual predicted probability between both reduced models, stratified on the first trimester outcome, are displayed in Figure 3.15.

Similarily to the complete models, the Figure 3.16 displays the predicted

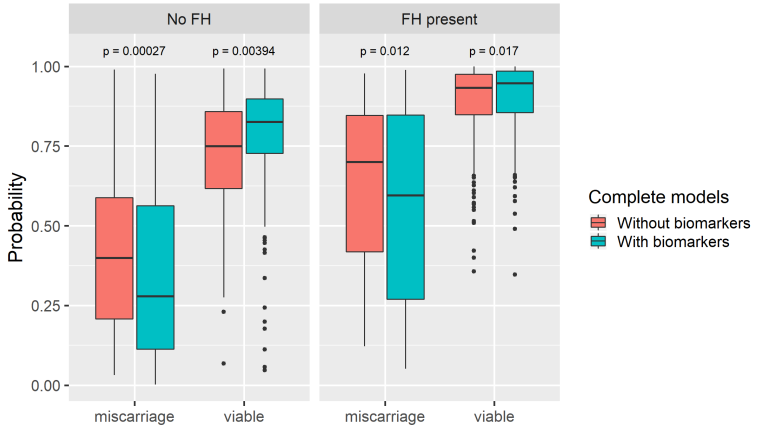


Figure 3.14: Changes in predicted probabilities after including the biomarkers variables in the complete model, stratified by the presence of fetal heartbeat. The improvement in discrimination is reflected with increased probabilities for viable pregnancies and decreased probabilities for miscarriage cases. Adding biomarkers information to the complete model significantly improves the predictions for both viable pregancies and miscarriages, irrespective of the presence or absence of heartbeat. The p-values correspond to Student t-tests.

probabilities of both reduced models in box plots, stratified on the outcome and the presence of FH. The gain of using biomarkers as model predictor is maintained for pregnancies of unknown viability (absence of FH). However, including biomarkers does not significantly modify the probability for pregnancies with observed FH. This absence of significant changes can be explained as follows: unlike the complete models, where the blood biomarkers constituted two additional variables to the same set of predictors, both reduced models were developed in parallel from two different sets of potential variables. As a result, the reduced models also differed in the clinical and sonographic variables included. In presence of FH, the information encoded by the blood biomarkers is likely to be represented, to some extent, by the MSD and FH variables, which are present in the first reduced model but are *not* in the reduced biomarkers model.

Decision curve analysis The decision curves for both complete and reduced models, with and without biomarkers are displayed in Figure 3.18. All models demonstrated improvement over a generic policy of subsequent monitoring

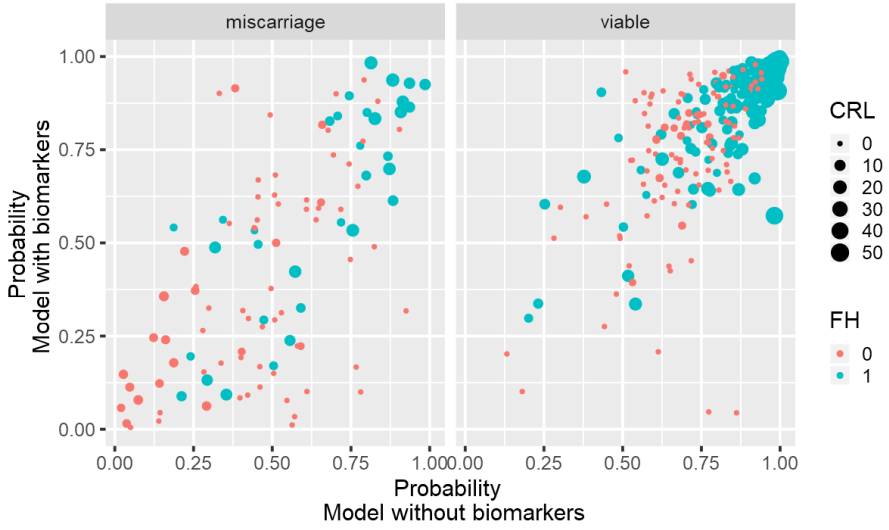


Figure 3.15: Changes in predicted probabilities between the two reduced models. These plots visualize the changes in predicted probabilities after including the biomarkers variables in a reduced model. Improvement in prediction is demonstrated by a probability increase for viable pregnancies and a probability decrease for miscarriages. The vast majority of viable pregnancies were already accurately predicted with a high confidence, even without biomarkers (most of the points are in the upper right quadrant). Introducing biomarkers resulted in significant probability changes for several miscarriage cases, as assessed by the presence of points in the lower right quadrant (pregnancies that were initially predicted as viable with a probability $p > 0.5$ but then reduced to $p < 0.5$ after introducing biomarkers information)

for all patients, for the whole range of decision thresholds. Complete models demonstrated a slight advantage in net benefit of over parsimonious models. Similarly, models with biomarkers had a slightly increased net benefit over similar models without biomarker information.

Conclusion Overall, adding biomarkers to multivariable models improved the discriminative performance of the clinical predictive models. They procured a net benefit compared to similar models without biomarkers (see Figure 3.18). Extensive models based on the complete list of variables, with or without biomarkers, did not improve the discriminative performance compared to simpler

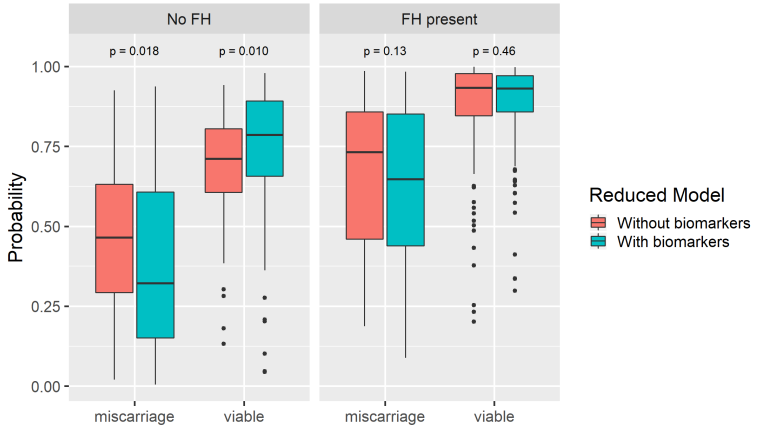


Figure 3.16: Changes in predicted probabilities after including the biomarkers variables in the reduced models, stratified by the presence of fetal heartbeat (FH).
The improvement in discrimination is reflected with increased probabilities for viable pregnancies and decreased probabilities for miscarriage cases. Unlike the complete model (Figure 3.14), adding biomarkers information to the reduced model only improves the predictions in the absence of heartbeat (no FH) but does not significantly affect the distribution of probabilities when the fetal heartbeat is observed (FH). In the presence of FH, the information encoded by the blood biomarkers is likely to be represented by the MSD and FH variables, which are *not* in the reduced biomarkers model. The p-values correspond to Student t-tests.

models. They also demonstrated worse calibration performance, potentially explained by the overfitting of the training datasets.

In view of these results, we decided to implement the following predictive modelling strategy in the DDSS. Whenever biomarkers are available, the reduced model with blood biomarkers should be applied (Model 2). Otherwise, the reduced model trained on clinical and ultrasound variables, as developed in Section 3.3, should be used (Model 1). Finally, we should note that even though extensive models do not demonstrate significant benefit over reduced models, collecting additional variables might still be beneficial to improve the imputation of potentially useful variables.

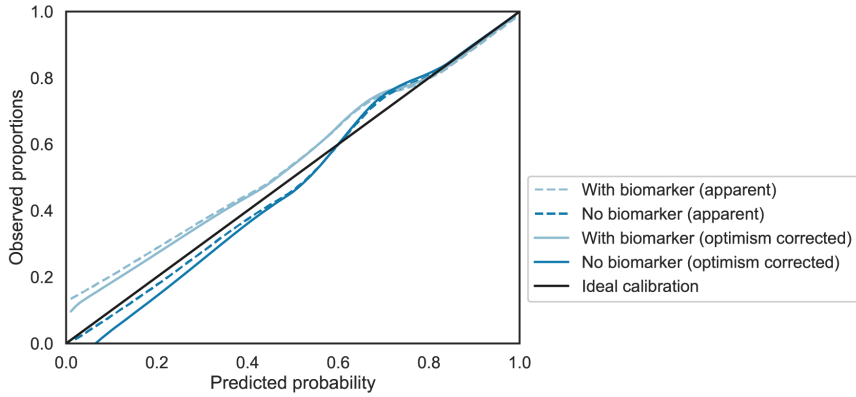


Figure 3.17: Apparent and optimism-corrected non-parametric calibration curves of the reduced clinical models augmented with biomarkers. Both models demonstrate excellent calibration performance, even after correction for optimism.

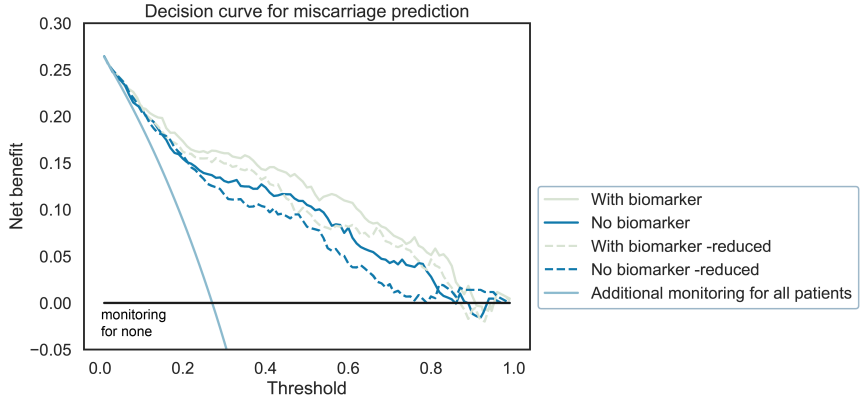


Figure 3.18: Decision curves of the predictive models with and without biomarkers. The models developed to predict first trimester viability offer a net benefit over a large range of potential decision thresholds compared to generic policies such as monitoring all pregnant patients or no monitoring. The complete model with blood biomarkers demonstrated the largest benefit although the four models shared similar profiles. As demonstrated for the predictive performance, adding blood biomarkers also improved the clinical utility of both reduced and complete models.

3.4 Query pre-processing

When deploying a model in a real setting, it is crucial to ensure that the same pre-processing steps are applied to each new data instance to guarantee the expected behavior of the predictive models. Additionally, because many predictive models cannot make predictions on incomplete data, appropriate strategies to mitigate the presence of missing values in new queries must be integrated in DDSS.

Objective

Implement a query pre-processing approach that performs online imputation of missing observation based on fixed chained equations.

In the present DDSS, the pre-processing component performs first a sanity check to assess if all the filled-in values correspond to the expected range and format of the variable used during model development. Subsequently, additional feature engineering steps create new variables from the existing set of recorded variables.

3.4.1 Variable transformations

Given the final set of variables included in the reduced models developed previously, each new data instance undergoes the following transformations :

- The MSD is truncated at the 99th percentile (53mm, derived from the training data)
- Progesterone is transformed into log(progesterone+1)
- The differential GA variable is created on the flight as GA by LMP – GA by MSD, where GA by MSD is obtained with the formula of $GA = MSD + 30$ [158].

Our multivariable logistic regression models require complete data to perform prediction. While the problem of missing values in model development is now well studied [75], incomplete data at prediction time is still regularly overlooked in the scientific literature. However, the proportion of missing values in clinical practice is certainly as important as during model development. Incomplete data can prevent a significant proportion of patients from benefitting of clinical predictive models. In the following section, we describe an online multiple imputation strategy to complete missing data at prediction time.

3.4.2 Online imputation

Background

A common practice to impute the new data consists in appending the incomplete data instance(s) to the training dataset and imputing the whole data simultaneously (a process known as *stacked imputation*). However, imputing a large dataset is time consuming, especially in the presence of hierarchical data structure, and impracticable in real-life. Alternative strategies consist in applying variable mean imputation or in developing an exhaustive collection of sub-models which contains all the different combinations of variables. In addition to the considerable computational cost of the later approach (for k variables, 2^k models must be constructed), these strategies can produce suboptimal results compared to multiple imputation [71, 159].

Imputation with fixed chained equation

Most of the current software do not provide the equations' parameters of the fully conditional models developed for the multiple imputation. Furthermore, knowing the fully conditional models would not be sufficient as fully conditional models require complete data, which is, by definition, never the case at test time as the outcome variable is always unknown. To overcome these problems, Hoogland et al. [70] leveraged fixed chained equations to easily impute data at prediction time. The procedure is summarized in the Algorithm 3. For each individual variable, this method extracts the equation of the corresponding fully conditional imputation model from the multiple imputed datasets. These parameters, along with the marginal means of the variables, constitute a summary of the original data needed to perform future imputation. Unlike stacked imputation, fixed chained equations do not necessitate large computational costs. Because of the fixed equations, the number of iterations needed is also reduced compared to the traditional MICE approach [70].

As pointed out by Hoogland et al. [70], this method, unlike the traditional chained equations approach, discards the stochastic component. Since our DDSS embeds a set of patient data for similar cases retrieval, we can extend the fixed chained equations approach by leveraging the internal database to perform predictive mean matching during the imputation process, and therefore re-introduce a stochastic component. The additional predictive matching step is summarized in Algorithm 4.

With predictive mean matching, the imputed value is guaranteed to remain within the observed values range [75]. For instance, Figure 3.19 displays a

Algorithm 3: Online Imputation with fixed chained equation.
$\boldsymbol{z}^i$ is the incomplete new data, FC is the set of fully conditional imputation models ($FC^{(j)}$ represents the model to impute the variable j), $(\overline{\boldsymbol{x}}, \overline{y})$ are the mean variables and outcome from the imputed data and γ is a threshold convergence parameter. $\boldsymbol{z}^c$ is the completed version of $\boldsymbol{z}^i$. Note that the binary variables are returned as continuous scores.

Input: $\boldsymbol{z}^i, FC, (\overline{\boldsymbol{x}}, \overline{y}), \gamma$
Output: $\boldsymbol{z}^c$

Set $\hat{y} = \overline{y}$;
Set M as the set of missing variables from $\boldsymbol{z}^i$;
Set $\boldsymbol{z}^c$=$\boldsymbol{z}^i$;
Set $y' = 2$;
for j *in* M **do**
  $\boldsymbol{z}^c_j = \overline{\boldsymbol{x}}_j$;
end
while $|y' - y| > \gamma$ **do**
  Update $y' = \hat{y}$;
  Update $\hat{y} = FC^{(y)}(\boldsymbol{z}^c)$;
  for j *in* M **do**
    $\boldsymbol{z}^c_j = FC^{(j)}(\boldsymbol{z}^c_{\setminus j}, y)$;
  end
end
return $\boldsymbol{z}^c_j$

Algorithm 4: Predictive mean matching.
$\boldsymbol{z}^c$ is the data instance completed by fixed chained equations, M is the initial set of missing variables of $\boldsymbol{z}^i$, $\mathbf{X}$ is the original dataset and $\mathbf{X}'$ is the predicted version of $\mathbf{X}$ by the fully conditional model FC

Input: $\boldsymbol{z}^c, M, \mathbf{X}, \mathbf{X}'$
Output: $\boldsymbol{z}^{pmm}$
Initialize $\boldsymbol{z}^{pmm} = \boldsymbol{z}^c$;
for j *in* M **do**
  Find k nearest donors based on $|\boldsymbol{z}^c_j - \mathbf{X}'_{\mathbf{j}}|$;
  Draw 1 donor i at random among the k nearest donors;
  Set $\boldsymbol{z}^{pmm}_j = \mathbf{X}_{\mathbf{i},\mathbf{j}}$;
end
return $\boldsymbol{z}^{pmm}$

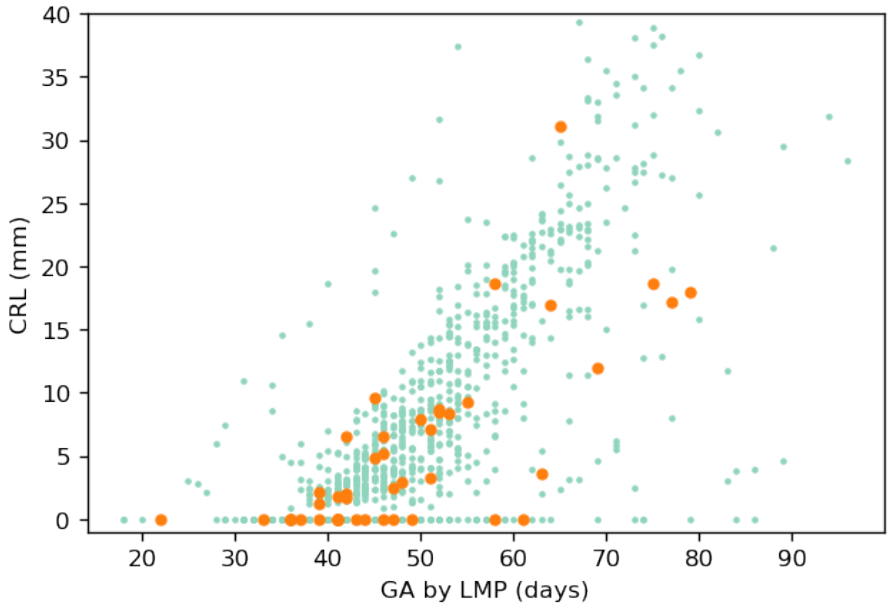


Figure 3.19: Example of online imputation with fixed chained equations and predictive mean matching. 40 incomplete CRL values were imputed. The imputed CRL values respect the truncated CRL distribution, reflecting the special CRL encoding (CRL = 0 when no embryo is visible). Additionally, no unrealistic combinations of CRL and GA by LMP are imputed.

single imputation of 40 incomplete CRL values along the GA by LMP variable. Figure 3.20 displays the distribution of imputed MSD values from 20 imputation models 20 using the fixed chained equations with predictive mean matching.

Assessment of the imputed values with fixed chained equations

Methodology We selected a random subset of 200 scans (n viable = 164 (82%) and n miscarriages = 36 (18%)) with complete values for the following predictors: *maternal age, folic acid supplementation, GA by LMP, MSD, β-hCG* and *progesterone*. These variables constitute the predictors of the reduced model from Section 3.3.2.

The number of possible missing variable combinations is too large to be tested and reported exhaustively. Therefore, a sampling approach was used to assess

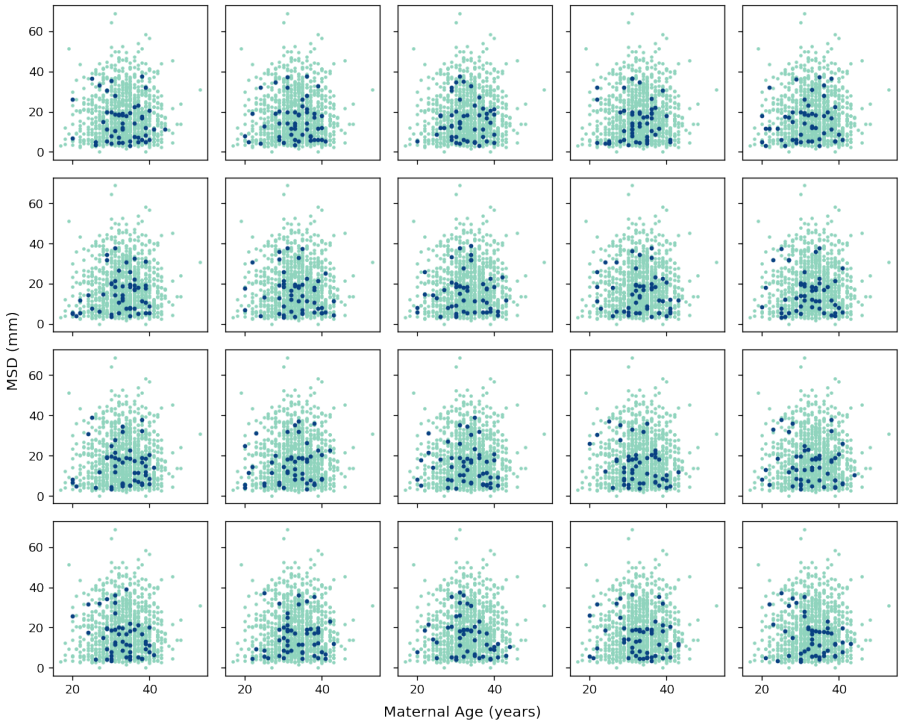


Figure 3.20: 20 runs of online imputation with fixed chained equations and predictive mean matching on missing MSD values.
Using different imputation models jointly with predictive mean matching introduces a stochastic effect during the imputation of missing values (blue dots) while maintaining the imputed values within a realistic range (complete data in grey)

the average impact of online imputation. For each scan, missing values were introduced in 1, 2, 3 and 4 randomly selected variables out of the 6 predictors used in the model. The incomplete data were then imputed following the procedure described in Section 3.4.2. For each artificially incomplete data point, twenty runs of the online multiple imputation approach were performed. The 20 imputed data were then used to predict the outcome with the predictive model. Finally, those predictions were pooled together into a unique predicted value according to Rubin's method [73]. An overview of the online imputation methodology is displayed in Figure 3.21.

To evaluate the efficiency of the online imputation scheme, we compared the performance of the reduced biomarker model (Model 2 from Section 3.3.2) on the complete subset with its performance on the various imputed datasets. For each completed dataset, the model discrimination was assessed with the AUC. The calibration was evaluated with non-parametric calibration curves and the ECI. The whole process was repeated 20 times, with variables randomly selected at each run. As overall estimates, the average AUC and ECI over those 20 runs were reported with standard deviation. We also reported the average execution time per single imputation for an individual data point[5].

Individual variables do not bear the same importance within a model. Therefore, when a value is missing in an important predictor, the final prediction will depend more heavily on the quality of the imputation method compared to missingness occurring in secondary predictors. Therefore, we also tested the impact of missing variables for a fixed selection of variables, i.e. the pre-selected variables were specified as *missing* for the whole subset of data.

Results The model performance on the complete data and the imputed datasets with various degrees of missingness is reported in Table 3.7. The performance measured by AUC and ECI over the 20 individual runs of online imputation is reported in Figure 3.22 and Figure 3.23, respectively. Increasing the number of missing variables slightly affects the discrimination performance. With 4 out of 6 variables missing, the average performance of the model based on imputed data decreased from 0.895 to 0.867, which remains satisfactory given the level of missingness. However, the level of variability between the different runs also increased with the number of missing variables (as depicted by the Figure 3.22). This renders the imputation with 4 missing predictors and above less reliable.

[5]Note that the original model was trained on the whole set of data, including this subset of 200 cases. Therefore, the performance metrics should not be regarded as an objective way to validate the model but rather as a means to estimate the impact of the variability introduced by the online data imputation method.

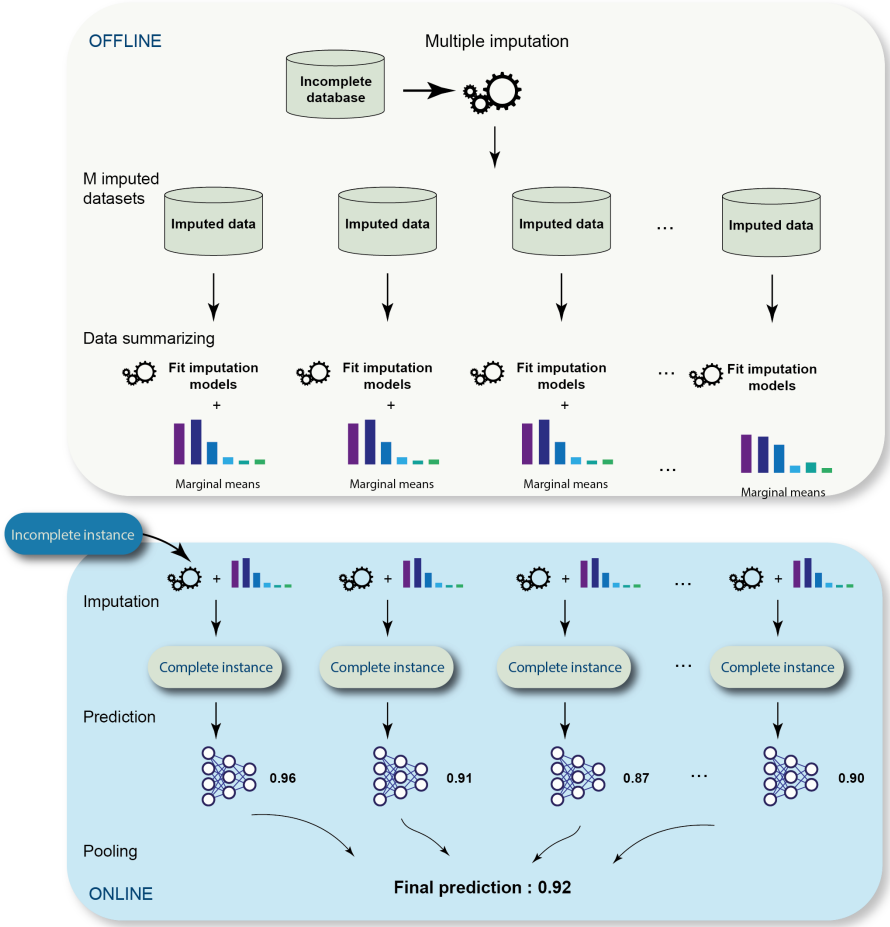


Figure 3.21: Overview of the online imputation process with fixed chained equations.

During the offline multiple imputation, a set of M fully conditional models and M corresponding vectors of variable means are produced and stored for future online deployment. Once a new incomplete data instance is entered, a set of M complete instances is produced with the online imputation procedure, as described in Algorithm 3. M predictions, one for each completed version of the new data point, are obtained from the model and then pooled together to produce the final prediction.

Table 3.7: Performance of online imputation with fixed-chained equations with increasing amount of missing variables.

	Complete subset	**Online imputation**			
		1 variable missing	2 variables missing	3 variables missing	4 variables missing
Mean AUC ± SD	0.895	0.883 ±0.0125	0.882 ±0.0137	0.870 ±0.0172	0.867 ±0.0302
Mean ECI ± SD	0.0018	0.00232 ±0.00073	0.00257 ±0.00103	0.00327 ±0.00178	0.00373 ±0.00205
Mean execution time (s) per imputation ± SD	-	0.036 ±0.069	0.065 ±0.102	0.078 ±0.122	0.091 ±0.135

Similarly, the model calibration was also affected. The variability introduced during the imputation process resulted in a higher degree of miscalibration. Although the ECI remained satisfactory based on the imputation of 1 or 2 missing variables, the calibration performance based on the imputation of 3 and 4 missing variables displayed lots of variability, as depicted by Figure 3.23 and the calibration curves of Figure 3.25. The imputation of more than 2 missing variables in Model 2 appears not suitable for a clinical usage.

The impact of predefined missing variables is reported in the Table 3.8. The corresponding calibration curves are displayed in Figure 3.24. While the imputation of most of the variables resulted in excellent model performance, the imputation of missing *progesterone* values significantly affected the model discrimination and calibration performance. A missing value in the *progesterone* variable will be more detrimental than a missing *maternal age* for instance. A potential explanation is that *progesterone* is not as strongly correlated with the other variables as the rest of the predictors. Therefore, it is more difficult for the fully conditional models to accurately impute *progesterone* values based on the rest of the variables.

An example of calibration curves for various predefined combinations of missing predictors is displayed in Figure 3.26. The corresponding discrimination and calibration metrics are reported in Table 3.9.

Finally, as expected, with a higher proportion of missing variables, the average execution time of the online imputation algorithm increased as it increases the number of iterations required before reaching the convergence criteria (Table 3.7). Still, even with a large amount of missing predictors, the execution time remains manageable for a single prediction. Furthermore, we should mention

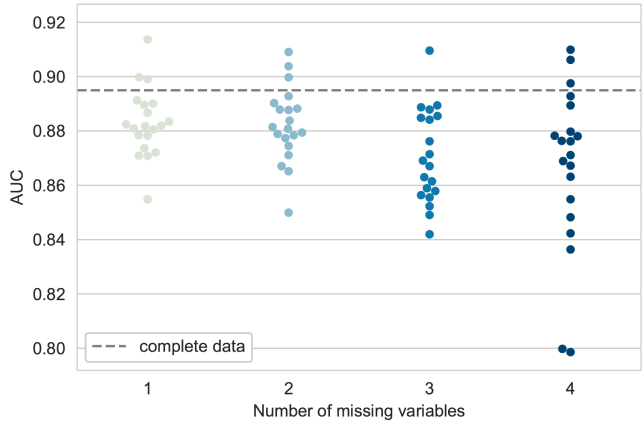


Figure 3.22: The discrimination performance of Model 2, measured by AUC, on a subset of imputed data with various amounts of missing variables. Each dot corresponds to a run of the online imputation method on the same subset of 200 patients but with a different random assignment of missing values. The dotted line represents the performance of the model on the original complete subset.

Table 3.8: Discrimination and calibration performance of Model 2 on a set of 200 patients after online imputation of each individual predictor. For most of the imputed variables, the model performance remains comparable to the complete data (AUC: 0.895, ECI: 0.0018). However, its performance on the imputed progesterone demonstrates a decrease of 0.048 in AUC and an increase of 0.0045 in ECI.

	Single fixed missing variable					
Metric	MSD	HCG	Maternal age	Proges-terone	Folic acid	GA by LMP
AUC	0.895	0.901	0.897	0.847	0.892	0.906
ECI	0.0018	0.0018	0.0025	0.0063	0.0054	0.0042

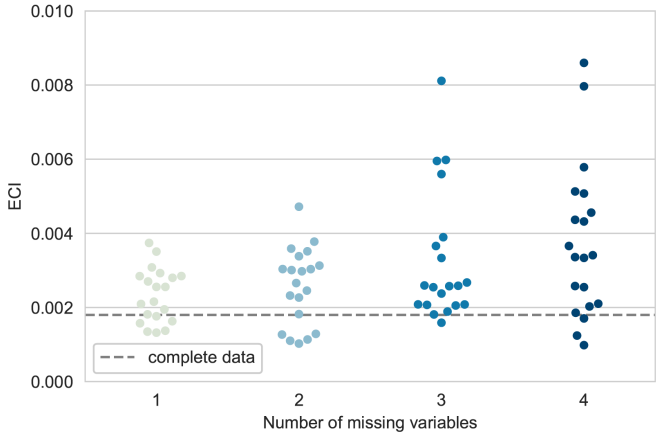


Figure 3.23: The calibration performance of Model 2, measured by ECI, on a subset of imputed data with various amounts of missing variables. Each dot corresponds to a run of the online imputation method on the same subset of 200 patients but with a different random assignment of missing values. The dotted line represents the performance of the model on the original complete subset. The lower the ECI, the better the calibration.

that the current implementation of the algorithm has not been optimized for time efficiency (e.g. parallelization of the computation for multiple imputation).

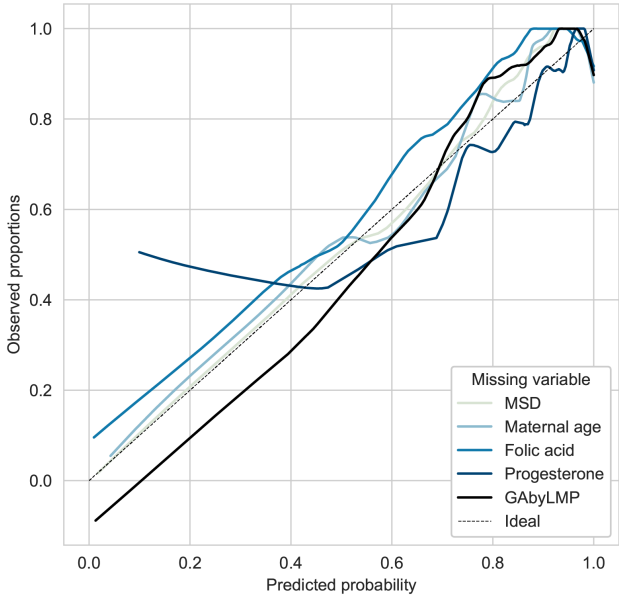


Figure 3.24: Calibration curves of Model 2 after online imputation of single missing variables. The model calibration on imputed progesterone demonstrates the largest departure from the ideal calibration line.

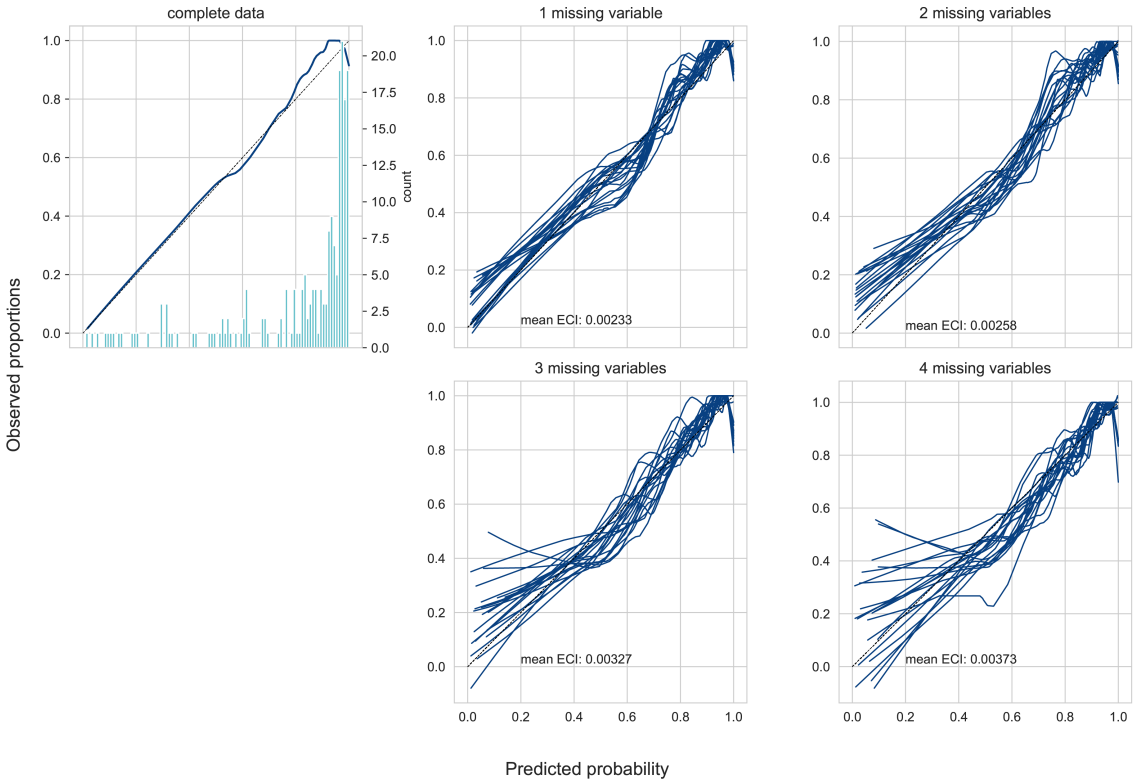


Figure 3.25: Calibration curves of Model 2 on the 20 subsets of imputed data with various amounts of missing variables. The corresponding mean ECI is also displayed. With an increasing amount of missing variables to impute, we observe a gradual departure from the ideal calibration diagonal line, which is also reflected by an increase in ECI.

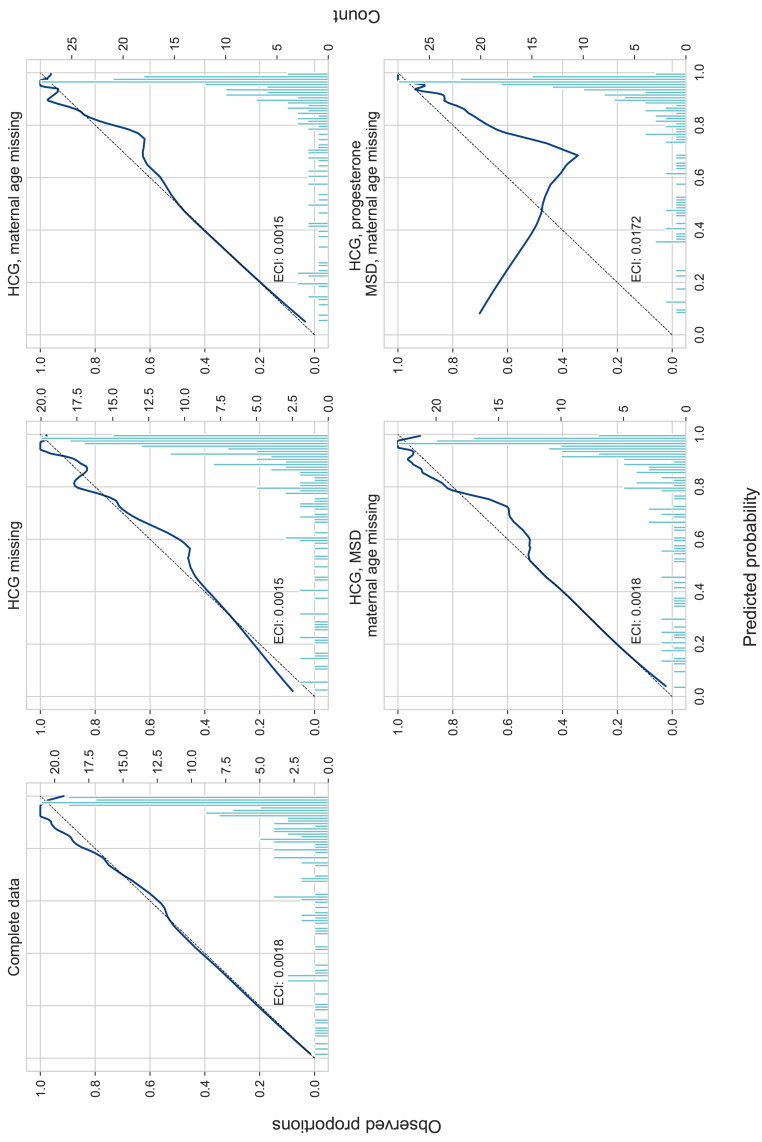


Figure 3.26: Calibration curves of Model 2 after online imputation of predefined combination of missing variables. The calibration curves remain close to the ideal calibration for the set of clinical and ultrasonographic predictors. However, adding progesterone to the set of missing variables results in poorer calibration performance.

Table 3.9: Performance of online imputation with fixed-chained equations with increasing amount of missing variables.

	Fixed combinations of missing variables		
Metric	HCG and Mater. age	HCG, MSD, Mater. age	HCG, MSD, Progesterone, Mater. age
AUC	0.906	0.907	0.838
ECI	0.0015	0.0015	0.0018

Although imputation at prediction time allows physicians to use predictive models that require complete data, the end-users should be encouraged to provide, whenever possible, complete data in the first place. The limitations of imputation approaches, e.g. the increase in prediction uncertainty, should be clearly notified to the end-users. To guarantee accurate predictions, we limit the presence of missing values to two predictors. Additional missing values would further drive the predicted value towards the average prediction determined by the marginal means.

Conclusion

These results demonstrated that online imputation based on fixed chained equations provides excellent and realistic imputed values when a limited number of predictors is missing. However, when more than 2 out of 6 predictors were missing, the online imputation approach could no longer deliver reliable imputed values. It results that missing values at prediction time should only be allowed for a restricted number of predictors, in this case no more than 2. The progesterone variable was the most difficult to impute as the performance of the model significantly dropped in presence of imputed progesterone values. In presence of missing progesterone values, it is advisable to switch over Model 1 which does not rely on biomarkers.

3.5 Patients search engine and case-based reasoning

Objective

Produce a patient search engine to retrieve similar cases from partial queries on the EPOS database.

As discussed in Chapter 2, physicians often reason about current cases by referring to their own previous experience [27]. Searching for meaningful examples of similar cases within an extensive database can enrich the clinical decision process with objective data. However, common EHR do not provide intelligent patient search as each medical condition demands a specific similarity metric (see Chapter 2). Additionally, missing values are omnipresent in real clinical databases. In order to exploit those incomplete datasets for case-based reasoning systems (CBR), specific metrics must be designed, as traditional similarity metrics cannot be computed on incomplete data. Furthermore, to improve the flexibility of a patient search engine, querying similar cases should not be restricted to a predefined list of mandatory variables. It might be of interest for the clinicians to retrieve similar patients based only on a few available variables of interest, referred here as *partial query*.

In the following section, we leveraged the findings of Chapter 2 to develop a supervised patient search engine for the EPOS database. We compared the performance of various pseudo-metrics derived from the proximity matrix of tree-based ensembles.

3.5.1 Methods

Dataset From the original EPOS dataset as described in Section 3.2, an extensive dataset of 3713 scans (n viable = 3355 (90.6%), n miscarriage=358 (9.4%)), from 1329 women (N viable =1139, N miscarriage =190) was used to populated the database of cases. All scans with miscarriages were excluded from the database. The list of the variables allowed in the query is reported in Table 3.10.

Similarity metrics For supervised CBR, we differentiate two types of databases: a database used to learn the supervised similarity metric (*training data*) and a database which acts as a repository to provide the similar cases (*queried database*). These two databases can be the same or different. The reason to use two different databases is that the original data can be constituted of different

Table 3.10: List of the variables allowed in the query to search for similar cases.

α-FP	Folic acid supplementation	Pain score at presentation
β-hCG	GA by CRL	Parity
Amnion	GA by LMP	Previous caesarean section
Bleeding days	GA by MSD	Previous miscarriage
Bleeding score at presentation	Gravidity	Progesterone
CA-125	Maternal age	PUQE score
Certainty of LMP	Maternal ethnicity	Smoking status
CRL	MSD	Worst bleeding score
Difference in GA	MYSD	Worst pain score
FH	Pain days	

well-defined subsets of data. Therefore, we might want to learn a similarity metric that is specific for each subset. Additionally, an important presence of missing values in the training database might negatively impact the learning of an efficient similarity metric. Therefore, the similarity metric might be learned on complete cases only, whereas similar cases can be retrieved from a much complete set of cases, including incomplete data.

As a baseline, the unsupervised standardized Euclidean distance was used. The distance was computed on the observed variables of the query (i.e. partial query). Therefore, only the complete cases of the queried database could be retrieved.

For the supervised similarity metrics, various approaches derived from the proximity matrix of gradient boosted trees were compared. The Light Gradient Boosted Machine (LGBM) was used to produce the proximity matrices. In the light of the results described in Chapter 2, the LBGM algorithm was not optimized but was set to the following parameters : `number_estimators`= 500 , `maximum_depth`= 6 , `learning_rate`= 0.1, `subsample`= 0.5.

The first pseudo-metric was a generic metric, trained on the whole training data, which also constituted the queried database. This approach considered the whole list of predefined variables (Table 3.10) and leveraged the LGBM default path for missing values. Because biomarkers were not measured in a large subset of patients, we also tested a stratified approach where two metrics were trained on two different subsets of patients: those who had at least one biomarker measured and those without any biomarkers available. During the retrieval of similar cases, this two-metrics approach was applied either on the full dataset or the corresponding subset (with or without biomarker). The appropriate metric was used depending on the presence of biomarkers in the query.

In the presence of a large amount of missingness in the query, the LGBM default path for missing values might dominate in the similarity score and might overshadow relevant non-missing variables. To overcome this situation, we implemented an alternative strategy more appropriate towards partial queries: individual similarity metrics were tailored to the subset of variables observed in the query. These metrics were learned either from the full training data or either from a restricted subset of complete cases. Similarly, the queried database was either restricted to the complete cases or extended to the full set of training data.

Performance metrics The relevance of the retrieved similar cases was assessed with k-NN classifiers for various values of $k \in \{1, 3, 5, 11, 15, 21, 31, 41, 51\}$. Odd numbers were chosen to avoid ties. The binary outcomes of the k nearest cases were averaged to compute a local risk of miscarriage. The discrimination performance was assessed with AUC and the calibration with ECI. Additionally, the calibration curves for various values of k were also reported.

Leave-one-out cross-validation The leave-one-out cross-validation process is summarized in Figure 3.27. Scans with miscarriage were excluded from both the training and test folds. For each individual patient, a supervised pseudo-metric was derived from the remaining set of patients using the unoptimized LGBM's proximity metric as described previously. This pseudo-metric was then used in a k-NN classifier to retrieve the k most similar patients of the training fold, which acted here as an external database of patients. By doing so, we prevented data-leakage between the training and the test folds in case a patient had several scans, i.e. a scan from a patient could never be considered similar to another scan of the same patient. The reported performance metrics were derived from the aggregated vector of individual predictions as depicted in Figure 3.27.

3.5.2 Results

As previously demonstrated in Chapter 2, the supervised metrics outperformed the unsupervised approach based on standardized Euclidean distance in terms of discriminative performance (Table 3.11). This difference was especially marked for smaller k.

Except for k equals 1 or 3 where the calibration performance was lower with the generic distance metric, the calibration of the k-NN classifiers was similar between unsupervised and supervised metrics (Table 3.12). Overall, all approaches resulted in similar calibration performance as demonstrated by the

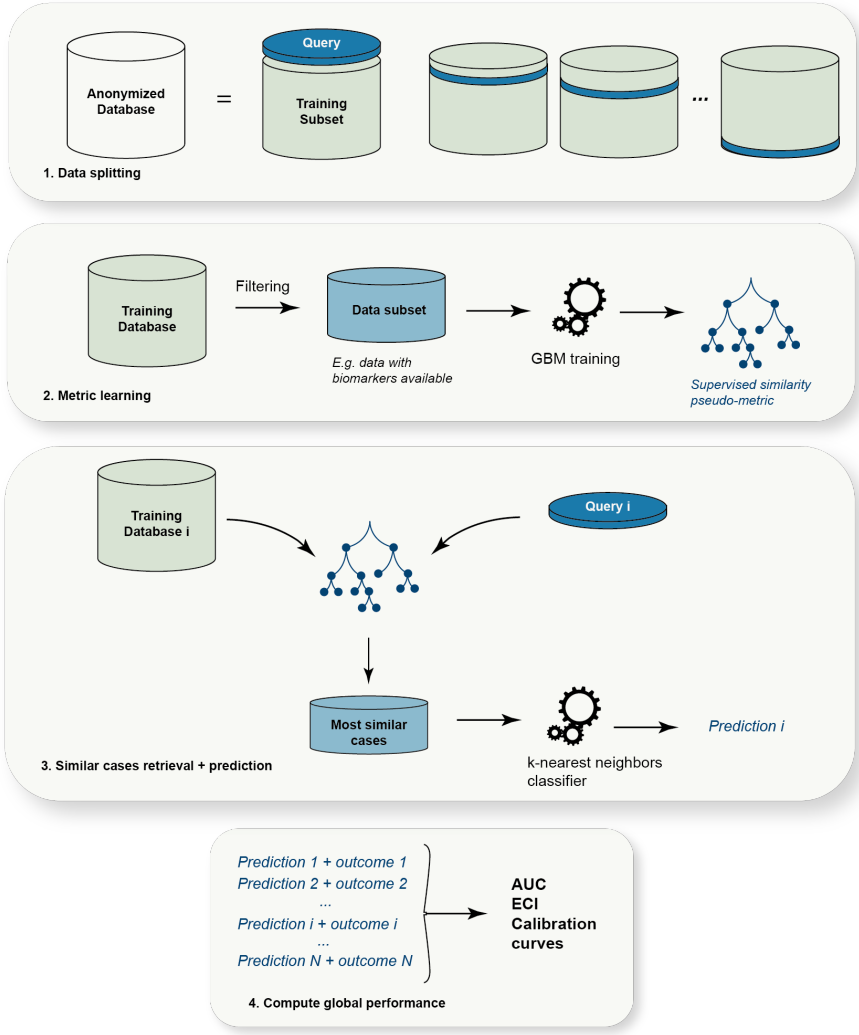


Figure 3.27: Overview of the leave-one-out cross-validation process.
1. At each iteration i of the LOOCV, the original dataset is split into an individual query at the patient-level and a corresponding training database. **2.** A tailored similarity metric is learned on the training data (or a subset of it) using gradient boosted machine (here, LightGBM). **3.** The k most similar patients are retrieved and a prediction is obtained from a k-NN classifier. **4.** Finally, the overall classification and discrimination performance is computed using the prediction and ground truth of each patient.

Table 3.11: Discriminative performance of k-NN classifiers measured by AUC for various supervised metrics and queried database subsets.
The generic metrics are trained on the full set of predictors, irrespective of their presence or not in the query, unlike the tailored metrics, which are adapted to the set of observed variables for each query. *Complete cases* subsets used only complete data to train the metric and/or retrieve similar cases. The *2 subsets* approach split the similarity learning into two parallel tasks depending on the biomarkers availability. The retrieval of similar cases was either performed on the whole data or also split into two subsets. The results are compared to k-NN classifiers trained on generic distance metrics (standardized Euclidean distance), tailored to the observed variables in each query. With an average AUC gain of 0.033, the k-NN classifiers with supervised similarity metrics constantly outperformed the classifiers based on the Euclidean distance. The difference was especially notable for small values of k.

	Std. Euclidean dist.	**Generic metrics**			**Metrics tailored for partial queries**		
Training data	Complete cases	Full data	2 subsets w/ & w/o biomarkers		Full data	Complete cases	Complete cases
Queried database	Complete cases	Full data	Full data	Subset	Full data	Full data	Complete cases
k=1	0.60	0.66	0.67	0.69	0.68	0.67	0.71
3	0.69	0.71	0.72	0.75	0.75	0.74	0.76
5	0.72	0.75	0.76	0.78	0.78	0.77	0.77
11	0.78	0.81	0.80	0.81	0.82	0.82	0.81
15	0.79	0.82	0.81	0.83	0.83	0.83	0.82
21	0.82	0.83	0.83	0.85	0.84	0.85	0.83
31	0.83	0.85	0.84	0.86	0.86	0.85	0.85
41	0.83	0.85	0.85	0.86	0.86	0.86	0.86
51	0.84	0.85	0.86	0.87	0.87	0.86	0.87

calibration curves in Figure 3.28. For most of the metrics, the best calibration occurred within the range of k equals 11 to 21, as depicted by the drop of ECI in Table 3.12.

Table 3.12: Calibration performance of k-NN classifiers measured by ECI for various supervised metrics and queried database subsets.
The generic metrics are trained on the full set of predictors, irrespective of their presence or not in the query instance to predict unlike the tailored metrics which are adapted to the set of observed variables for each query. *Complete cases* subsets used only complete data to train the metric and/or retrieve similar cases. The *2 subsets* approach split the similarity learning into two parallel tasks depending on the biomarkers availability. The retrieval of cases of similar cases was either performed on the whole data or also split into two subsets. The results are compared to k-NN classifiers trained on generic distance metrics (standardized Euclidean distance), tailored to the observed variables in each query. For most of the metrics, the best calibration occurred within the range of $k \in [11, 21]$.

	Std. Euclidean dist.	**Generic metrics**			**Metrics tailored for partial queries**		
Training data	Complete cases	Full data	2 subsets w/ & w/o biomarkers		Full data	Complete cases	Complete cases
Queried database	Complete cases	Full data	Full data	Subset	Full data	Full data	Complete cases
k=1	0.0455	0.0221	0.0282	0.0235	0.0278	0.0321	0.0313
3	0.0097	0.0058	0.0072	0.0066	0.0061	0.0084	0.0086
5	0.0044	0.0028	0.0034	0.0030	0.0029	0.0041	0.0059
11	0.0008	0.0014	0.0011	0.0012	0.0007	0.0009	0.0026
15	0.0007	0.0018	0.0007	0.0009	0.0005	0.0005	0.0018
21	0.0014	0.0018	0.0005	0.0008	0.0006	0.0004	0.0012
31	0.0014	0.0028	0.0006	0.0009	0.0010	0.0004	0.0007
41	0.0018	0.0033	0.0008	0.0009	0.0014	0.0006	0.0006
51	0.0024	0.0039	0.0010	0.0010	0.0017	0.0009	0.0009

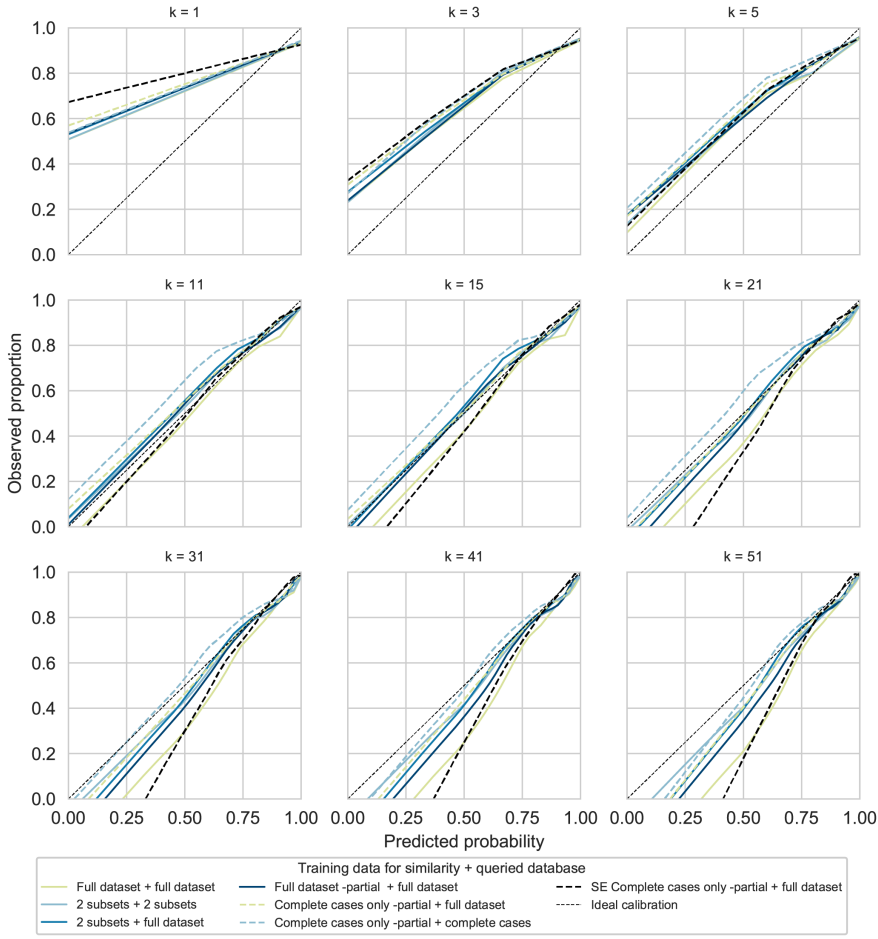


Figure 3.28: Calibration curves of the k-NN classifiers based on different task-specific distance metrics at various k.
As k increases, the granularity of the predictions improves, allowing for finer calibration. The classifiers on the second row with $k \in 11, 15, 21$ demonstrated the best calibration. SE= standardized Euclidean distance.

An example of similar cases retrieval is displayed in Figure 3.29. The similarity metric was trained on the full data using only the observed variables of the query. The main limitations of this approach is the need to compute a metric on the fly given the observed variables of the query. If the missingness occurs in some systematized ways, e.g. most of the incomplete variables are constituted

by the biomarkers variables, a good trade-off between a unique metric and individual metrics is a stratified approach were various metrics are trained on specific subsets of data (e.g. with and without biomarkers). However, such a method would limit the pertinence of partial queries, as metrics developed in a stratified approach exploit a list of predefined variables that might not be available in the query. Partial query search provide more explicit results than generic approaches, as only the variables included for the computation of the similarity are displayed to the end-users (Figure 3.29).

	# previous misc.	GA by MSD (days)	Worst pain score	Gravida	Certainty of LMP date	α-FP (ng/mL)	PUQE score	β-hCG (mIU/mL)	GA by LMP (days)	MSD (mm)	MYSD (mm)	Maternal Age (years)	Progesterone (ng\mL)	Difference in GA (days)	CRL (mm)	Similarity	Outcome
Query	0	41	0	3	10	3.09	10	11822	44	11	0	33	28.73	3	0		
2639	2	40	0	4	10	1.51	-	6417	44	10	0	34	30.8	4	0	0.518	Viable
3220	0	40.5	1	3	10	1.64	7	36701	44	10.5	2.6	37	29.99	3.5	2.9	0.4905	Viable
3552	0	37.3	0	2	10	3.72	6	10007	47	7.3	1.7	32	26.39	9.7	0	0.483	Viable
3658	1	38.8	0	4	10	0.968	3	13945	43	8.8	1	36	30.46	4.2	0	0.4785	Miscarriage
3357	0	35.9	0	4	2	1.82	3	5712	39	5.9	0.75	34	27.28	3.1	1.7	0.4775	Viable
827	1	38.4	0	2	10	-	6	-	43	8.4	0	34	-	4.6	0	0.4725	Viable
1046	0	42.5	0	3	9	1.52	-	16679	43	12.5	0	29	32.72	0.5	0	0.4715	Miscarriage
1687	1	38	0	2	10	1.28	3	4820	40	8	1.3	32	32.67	2	0	0.4685	Miscarriage
1972	0	41.3	0	6	10	3.01	3	16405	43	11.3	0	41	32.28	1.7	0	0.446	Miscarriage
3504	0	40.3	0	2	9	0.558	3	7970	44	10.3	-	35	27.14	3.7	0	0.4385	Viable
3800	0	37.8	0	2	10	1.26	6	9637	49	7.8	0	32	17.6	11.2	0	0.4305	Viable
3435	0	33.7	0	2	10	3.01	3	2252	32	3.7	0	30	27.38	-1.7	0	0.4075	Viable
1532	0	49.3	0	2	10	2.9	3	45158	52	19.3	2.9	33	42.55	2.7	6.6	0.406	Viable
2969	0	51.8	0	3	8	2.74	7	93604	54	21.8	2.7	32	21.02	2.2	12.3	0.404	Viable
2968	0	37.2	0	3	8	1.3	3	4684	40	7.2	2.5	32	23.73	2.8	0	0.402	Viable

Figure 3.29: Example of similar cases retrieval on a query randomly chosen from the original dataset. The similarity metric was trained on the whole training data using only the observed variables of the query (partial query search). The queried database was constituted of the full data. The color of each individual cell is determined by the level of disparity with the query for that specific variable.

3.5.3 Conclusion

Compared to predictive models, this CBR framework is not restricted to first scan data and does not require LMP values to produce predictions. In order to query the extensive but incomplete EPOS database, we implemented and compared various supervised similarity metrics derived from tree-based ensemble's proximity matrices. Unlike methods based on the Euclidean distance, those similarity metrics can retrieve incomplete cases, which are very common in real clinical databases. In addition, these methods outperformed the unsupervised standardized Euclidean distance by retrieving cases that produced more discriminant classifiers. Query-tailored metrics, trained on the subset of currently observed variables, provided more explicit results than generic approaches, as only the variables included for the computation of the similarity were displayed to the end-users. However, they require to be re-trained for each query while the generic supervised similarity metrics can be developed off-line and re-used without further adjustments.

In the light of these results, a supervised similarity metric tailored for the observed variables of the query, but trained on the whole dataset, constitutes an appropriate choice to retrieve similar cases from a database of a DDSS.

3.6 Interpretability and visualization

Objective

Deliver interpretable results of the predictive models through intuitive visualizations of individual explanations.

This last section is devoted to the interpretability of the DDSS. This component constitutes a crucial interface between non-expert end-users and the DDSS internal models. In what follows, we implement interpretability strategies to facilitate the understanding of the predictive models previously developed. We leverage the SHAP framework, a feature attribution method based on Shapley values from the collaborative game theory, to improve the models' transparency and to deliver individual interpretable decisions (model explainability).

3.6.1 Introduction

The prediction process of logistic regression models is fully transparent. It is self-contained within the variables' coefficients, which are directly available and easy to report for reproducibility purposes. Traditionally, logistic regression models are scrutinized through in-depth analysis of these coefficients. However, such analyses do not provide direct understanding of individual predictions. Additionally, inference on model's coefficients is subject to adequate model development (e.g. absence of multi-collinearity). Furthermore, although transparent, logistic regression models are often erroneously claimed to be interpretable and remains difficult for non-statistical experts to understand, especially in the presence of non-linear terms, multi-collinearity or a large number of variables [160]. It results that logistic regression offers poor global and individual understanding for non-expert clinical users.

To overcome some of these aspects, a common strategy is to develop scoring systems on the top of the model (e.g. [106]), although it might result in lack of precision. Historically, computing these scores was facilitated by charts and nomograms [161]. These approaches translate individual variables to appropriate scores and the final risk is obtained through the scaling of the total score. In the digital area, nomograms appear somehow obsolete [162] as it involves unnecessary manual computation steps such as individual scores summation that results in lack of efficiency.

For this DDSS, we leveraged the SHAP framework, a model agnostic feature attribution method based on the Shapley values (see Appendix 2.B from

Chapter 2 for a more detailed description), to facilitate the understanding of the predictive models developed previously. To briefly summarize, this method fairly decomposes each prediction among the model's variables. As a result, the final predicted value can be seen as additive contributions of individual variables. For non-expert end-users, such decomposition is intuitive, as it becomes directly apparent what variable(s) drove the final predicted outcome, in what direction and in what proportion.

3.6.2 Methodology

The SHAP interventional approach was used to compute the individual SHAP values. The interventional method accounts for inter-variables correlation, unlike the independent approach [87]. The resulting SHAP values share credit among correlated variables. This constitutes an additional benefit over traditional coefficients analysis, which assumes an absence of correlation between predictors (in practice, multi-collinearity is often present among variables).
Because the logistic regression models were developed under MICE to handle missing values, the pooled models were explained by SHAP. The 20 corresponding imputed datasets were used in parallel to perform the permutations required by the interventional SHAP method. The resulting SHAP values were then averaged over the imputed datasets to get the final SHAP values.

Model transparency

By aggregating many individual variable-wise decompositions over a representative dataset, the SHAP method also provides intuitive feature importance. Specifically, features importance can be represented as the average absolute SHAP value per variable, which corresponds to the average contribution of a variable to a set of predictions (e.g. the training data). A variable that greatly contributes to the final predictions, be it in a positive or a negative way, receives therefore a large average absolute SHAP value. The aggregated SHAP feature importance is reported as ordered barplots and beeswarm plots. The beeswarm plots contained additional information over the aggregation, as individual data-points are represented with the corresponding SHAP values.

Model explainability

The primary goal of a clinical predictive model is to provide individual predictions, at the patient-level. The individual predictions derived from

the SHAP feature attribution method can be visualized with decision plots. These plots display individual variable contributions expressed as scaled SHAP value, ordered by importance. Because this feature attribution method is contrastive, individual explanations are expressed relative to a baseline value, here represented by the viability prevalence at the end of the first trimester of pregnancy in the study population (0.85). For an individual prediction, the sum of each variable's SHAP value represents the contribution of the variables to explain the departure of the current prediction from the baseline value.

3.6.3 Results

The feature importance of both reduced models developed in Section 3.3 is represented on Figure 3.30. Unlike the raw feature importance reported on the bar plots, the beeswarm plots reveal the relation between an individual variable and its contribution to the outcome. For Model 1 (the clinical model), the interaction term MSD*FH had the largest feature importance with an average of 0.7 SHAP value. As expected, large values of MSD*FH were associated with a higher chance of viable pregnancy, demonstrated with positive SHAP values. The discrepancy between observed and expected GA was also attributed a large feature importance and corroborates the findings of previous studies [163, 164]. Positive differences, i.e. smaller observed GA than expected, were associated with a larger risk of miscarriage (more negative SHAP values). In both models, maternal age was correctly associated with a higher risk of miscarriage. This relationship between miscarriage and maternal age has been established by several large scale studies [98, 165].

Finally, both biomarkers, progesterone and β-hCG, were positively associated with an increased probability of viable pregnancy. They were ranked as the two most important variables of the biomarkers model. The predictive value of these two markers has been demonstrated in several studies (e.g. [117, 118]).

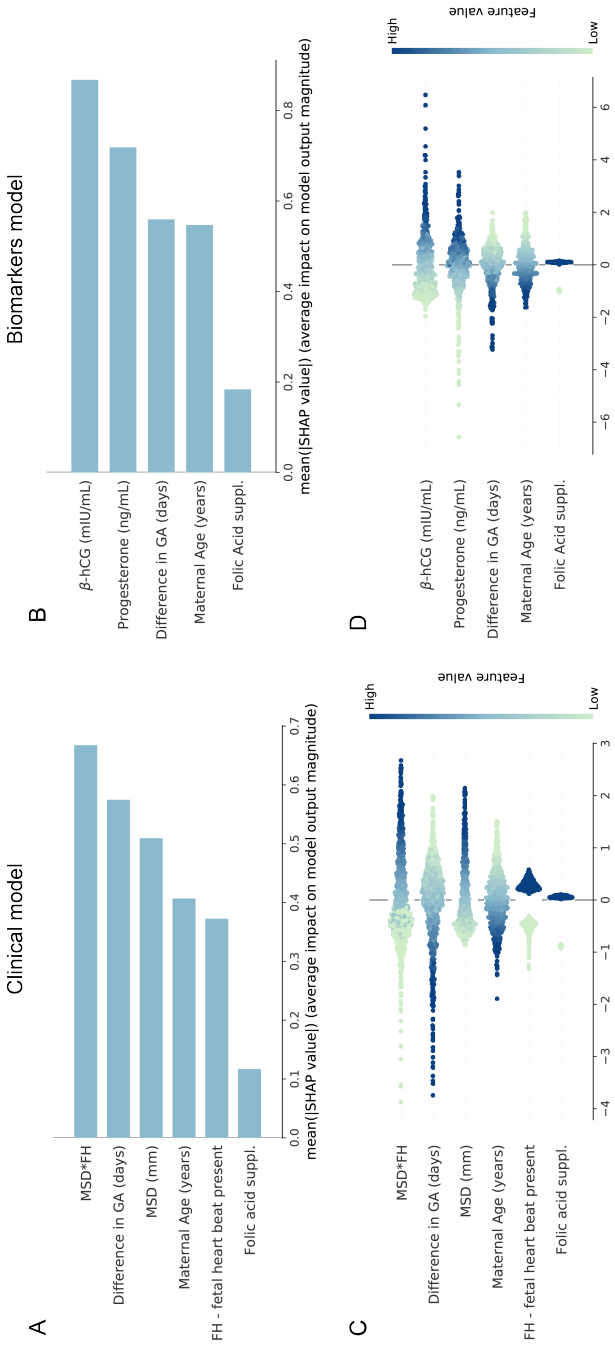


Figure 3.30: Feature importance of both reduced models developed in Section 3.3 measured as the mean(|SHAP value|) per variable. **A.** and **B.** display the simplistic feature importance as ordered bar plots for the clinical and the biomarker models, respectively. The beeswarm plots **C.** and **D.** demonstrate the distribution of SHAP values per variable, colored per feature value, for both models.

Examples of individual decision plots for both clinical and biomarkers models are reported in Figures 3.31 and 3.32, respectively.

Limitations The main limitation of the SHAP interventional approach is its computational cost. However, for simple models with restricted datasets (e.g. with N=1000), the average computational time is less than a second (average time: 0.71s, SD ±0.05s.). The computation of the transformation matrix to account for variable correlation can be performed offline, which further decreases the cost of computing the SHAP values for a single instance to 0.012s (SD ±0.003s). Finally, we should note that SHAP, although being model agnostic, is not model *independent*. Therefore, the relevancy of SHAP values are subject to appropriate model specification, training and calibration. A potential misuse of the SHAP framework is that the explanations provided become interpreted as *absolute* causal relations, obliterating the underlying model- and data-dependencies.

3.6.4 Conclusion

The SHAP method facilitates the understanding of logistic regression models and their predictions for first trimester viability from non-statistical experts. Individual explanations at the patient-level are delivered to the physicians as additive variable contributions towards the final predicted outcome. This additive process can be visualized on intuitive decision plots, which display the influence of each variable on the predicted risk of miscarriage. Additionally, while traditional inference on logistic regression's coefficients is subject to the absence of multi-collinearity, the interventional SHAP method provides fair explanations despites correlated variables. Finally, to improve the global understanding of the model, the variable-wise aggregation of individual SHAP values provides fair and intuitive feature importance measurements, irrespective of the variables unit.

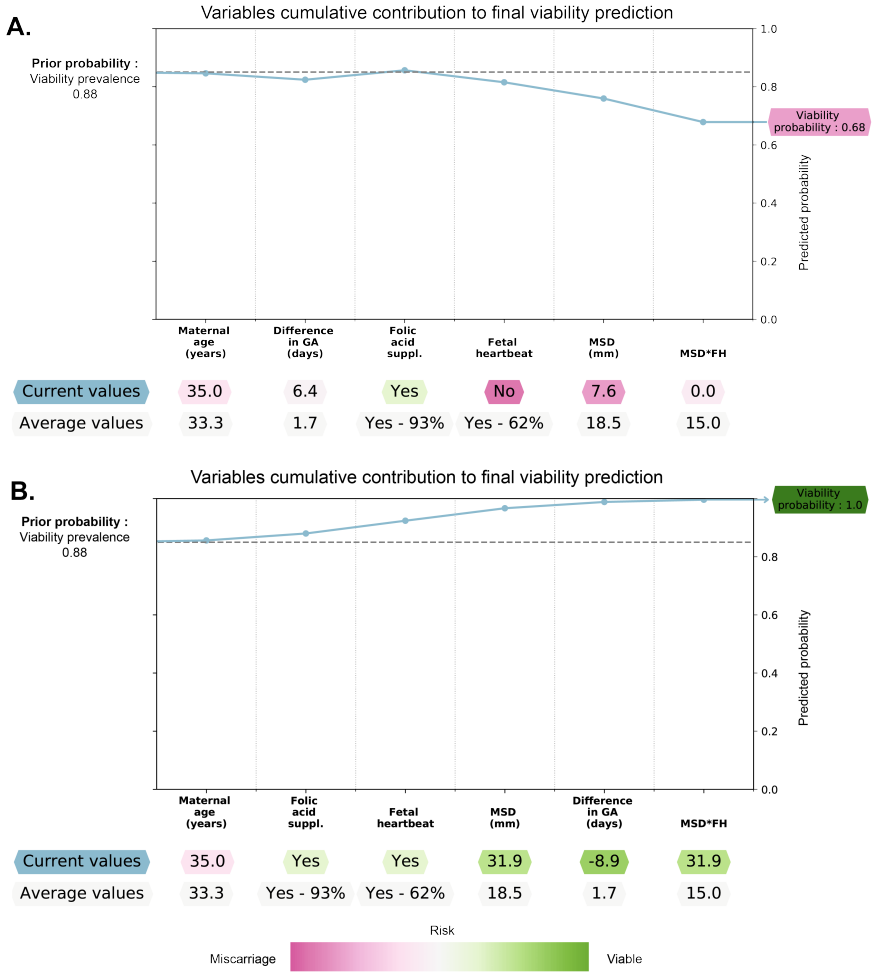


Figure 3.31: Examples of individual decision plots from the reduced Model 1 (clinical model) on two random patients with a different outcome. **A.** Prediction on first scan of patient E1104 (the first trimester outcome is a *miscarriage*). **B.** Prediction on first scan of patient E1041 (the first trimester outcome is a *viable* pregnancy). The variables are ordered by order of contribution to the final output.
Starting from the baseline value on the left (viability prevalence in the study population), the contribution of each individual variable is added to finally reach the predicted probability on the right. For each variable, the average value from the training data is reported underneath the current predicted instance for comparison purposes. A coloring scheme, based on the SHAP value, is added to facilitate the interpretation of each variable contribution towards a *viable* or *miscarriage* outcome.

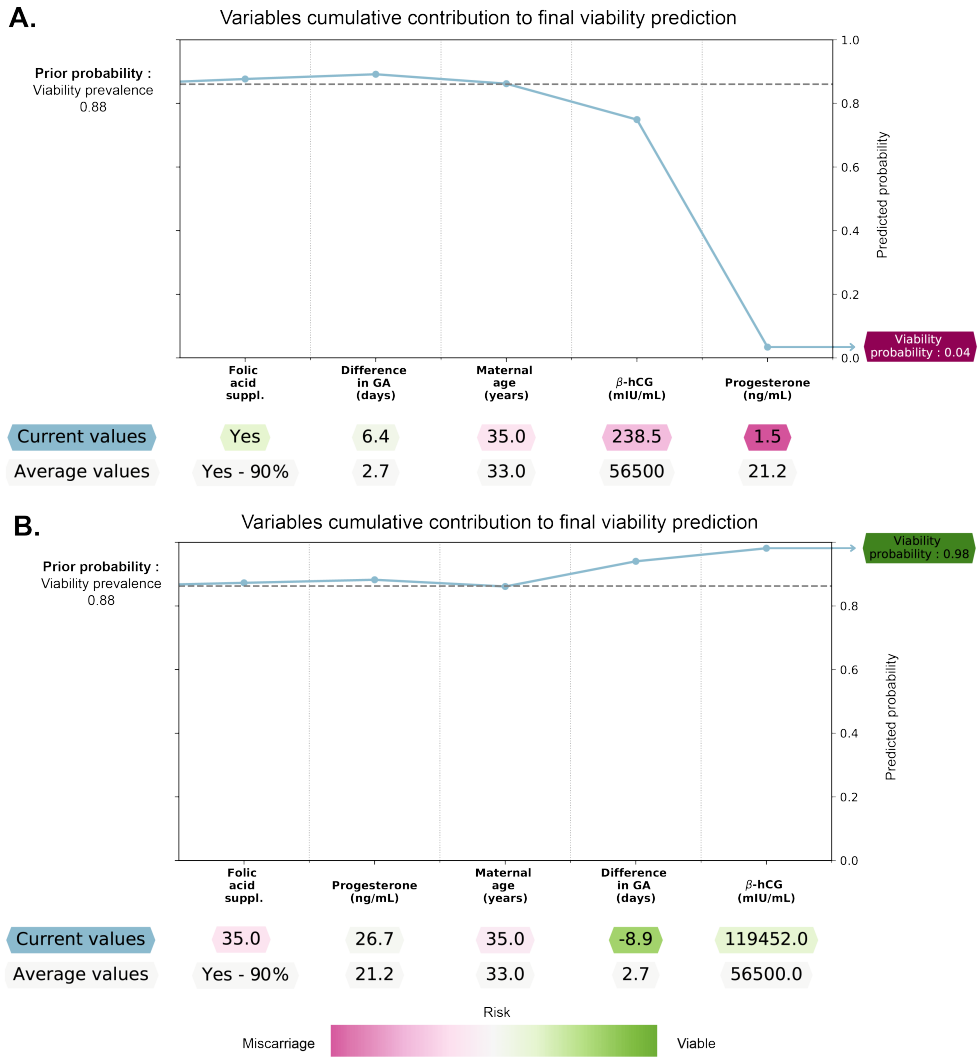


Figure 3.32: Examples of individual decision plots from the reduced Model 2 (biomarkers model) on the same two patients randomly selected for Figure 3.31. **A.** Prediction on first scan of patient E1104 (the first trimester outcome is a *miscarriage*). **B.** Prediction on first scan of patient E1041 (the first trimester outcome is a *viable* pregnancy). The variables are ordered by order of contribution to the final output.
Compared to Figure 3.31, the presence of the biomarker variables in Model 2 drastically improved the prediction of miscarriage in patient E1104 (**A**) while the prediction of the viability for patient E1041 (**B**) is maintained with a high degree of confidence.

3.7 Conclusion

In this project, we developed a complete DDSS framework for first trimester viability prediction that embeds interpretable predictive models, a case-based reasoning component and relevant visualization plots. Based on the extensive EPOS data resulting from a large prospective study, we developed and validated two clinically relevant models to predict the viability at the end of the first trimester of pregnancy, from early pregnancy scans. Given the availability of ultrasound (MSD, FH) and/or biomarker predictors (progesterone and β-hCG) at hand, the clinician can use the appropriate model to compute calibrated probabilities of viability. In addition, this DDSS includes an online imputation component that supports accurate predictions on moderately incomplete data.

By leveraging the EPOS dataset beyond model development, we also developed a case-based reasoning interface that facilitates the retrieving of previous similar cases. This patient search-engine supports partial queries and can also be exploited for retrospective data-analysis. This component relies on a supervised similarity metric derived from a tree-based ensemble proximity matrix. Therefore, the search of similar cases can be performed on incomplete databases and only considers variables relevant to the risk of miscarriage to compute the similarity.

Finally, we demonstrated that the SHAP framework, an additive feature attribution method, could facilitate the interpretation of logistic regression models for non-statistical experts by delivering individual explanations and improving the model's transparency with intuitive feature importance.

Appendices

3.A Supplementary tables

Table 3.A.1: Characteristics of the dataset used for clinical model development in Section 3.3.
The comparison of the cohort characteristics with regard to first trimester viability was assessed with the Student's t-test for continuous variables and the χ^2 test for binary or categorical variables.
BMI: body mass index, CRL: crown rump length, FH: fetal heart, GA: gestational age, LMP: last menstrual period, MSD: mean sac diameter, MYSD: mean yolk sac diameter, PUQE: pregnancy unique quantification of emesis score, SD: standard deviation.

	Miscarriage (N=160)	**Viable** (N=945)	**Total** (N=1105)	**p-value**
Maternal age (years)				< 0.001
Mean (sd)	34.8 (5.3)	33.0 (5.2)	33.3 (5.2)	
Range	22.0 - 46.0	17.0 - 53.0	17.0 - 53.0	
Maternal ethnicity				0.531
White	113 (70.6%)	642 (68.0%)	755 (68.4%)	
Black	15 (9.4%)	90 (9.5%)	105 (9.5%)	
Asian	15 (9.4%)	130 (13.8%)	145 (13.1%)	
Mixed	8 (5.0%)	32 (3.4%)	40 (3.6%)	
Other	9 (5.6%)	50 (5.3%)	59 (5.3%)	
Missing	0	1	1	
BMI (kg/m^2)				0.014
Mean (sd)	25.8 (5.7)	24.6 (5.0)	24.8 (5.1)	
Range	17.3 - 49.6	15.8 - 46.4	15.8 - 49.6	
Missing	32	15	47	
Previous miscarriages				0.274
0	74 (46.5%)	482 (51.0%)	556 (50.4%)	

1	49 (30.8%)	297 (31.4%)	346 (31.3%)	
2	21 (13.2%)	118 (12.5%)	139 (12.6%)	
3 +	15 (9.4%)	48 (5.1%)	63 (5.7%)	
Missing	1	0	1	
Certainty LMP (0-10)				0.694
Mean (sd)	9.8 (0.6)	9.8 (0.6)	9.8 (0.6)	
Range	7.0 - 10.0	7.0 - 10.0	7.0 - 10.0	
Smoking status				0.251
No	139 (88.5%)	858 (91.4%)	997 (91.0%)	
Yes	18 (11.5%)	81 (8.6%)	99 (9.0%)	
Missing	3	6	9	
Folic acid supplementation (yes/no)				< 0.001
0	21 (13.4%)	53 (5.6%)	74 (6.7%)	
1	136 (86.6%)	890 (94.4%)	1026 (93.3%)	
Missing	3	2	5	
Parity				0.566
0	77 (48.4%)	486 (51.4%)	563 (51.0%)	
1	60 (37.7%)	333 (35.2%)	393 (35.6%)	
2+	22 (13.8%)	126 (13.3%)	148 (13.4%)	
Missing	1	0	1	
Gravity				0.183
1	31 (19.5%)	225 (23.8%)	256 (23.2%)	
2+	128 (80.5%)	720 (76.2%)	848 (76.8%)	
Missing	1	0	1	
Worst bleeding score (0-4)				0.648
0	88 (55.3%)	525 (56.6%)	613 (56.4%)	
1	36 (22.6%)	237 (25.6%)	273 (25.1%)	
2	27 (17.0%)	122 (13.2%)	149 (13.7%)	
3	4 (2.5%)	27 (2.9%)	31 (2.9%)	
4	4 (2.5%)	16 (1.7%)	20 (1.8%)	
Missing	1	18	19	
PUQE score (3-15)				< 0.001
Mean (sd)	3.8 (1.7)	4.5 (2.3)	4.4 (2.2)	
Range	3.0 - 10.0	3.0 - 15.0	3.0 - 15.0	
Missing	9	63	72	
Fetal heart presence (yes/no)				< 0.001
No	108 (72.0%)	274 (29.7%)	382 (35.6%)	
Yes	42 (28.0%)	648 (70.3%)	690 (64.4%)	
Missing	10	23	33	
CRL (mm)				< 0.001
Mean (sd)	2.0 (3.7)	9.2 (11.3)	8.2 (10.8)	

Range	0.0 - 18.7	0.0 - 70.4	0.0 - 70.4	
Missing	2	17	19	
MSD (mm)				< 0.001
Mean (sd)	11.5 (7.8)	19.3 (11.6)	18.1 (11.4)	
Range	0.0 - 56.7	0.0 - 68.9	0.0 - 68.9	
Missing	3	51	54	
MYS (mm)				0.099
Mean (sd)	2.3 (3.2)	2.6 (1.4)	2.6 (1.8)	
Range	0.0 - 34.0	0.0 - 9.7	0.0 - 34.0	
Missing	26	119	145	
GA by LMP (days)				0.043
Mean (sd)	48.2 (11.7)	50.3 (12.2)	50.0 (12.2)	
Range	20.0 - 89.0	18.0 - 96.0	18.0 - 96.0	
GA difference (LMP-MSD) (days)				< 0.001
Mean (sd)	6.7 (9.8)	0.3 (7.9)	1.3 (8.5)	
Range	-96.2	-84.3	-101	
Missing	5	52	57	

Table 3.A.2: Characteristics of the data subset used to develop the models on *consecutive* scans in 3.3 Comparison of the cohort characteristics with regard to first trimester viability was assessed with the Student's t-test for continuous variables and the χ^2 test for binary or categorical variables. BMI: body mass index, CRL: crown rump length, FH: fetal heart, GA: gestational age, LMP: last menstrual period, MSD: mean sac diameter, MYSD: mean yolk sac diameter, PUQE: pregnancy unique quantification of emesis score, SD: standard deviation. The longitudinal aggregation of the binary variable FH was produced as follows : no FH at both scan = -2; FH a 1st scan but not 2nd =-1; FH at 2nd scan but not first =1 and FH at both scan =2.

	Miscarriage (N=94)	**Viable** (N=857)	**Total** (N=951)	**p-value**
Maternal Age				0.005
Mean (SD)	34.7 (5.4)	33.1 (5.2)	33.3 (5.2)	
Range	22.0 - 46.0	17.0 - 53.0	17.0 - 53.0	
Maternal Ethnicity				0.246
Black	10 (10.6%)	76 (8.9%)	86 (9.1%)	
Asian	9 (9.6%)	116 (13.6%)	125 (13.2%)	
White	61 (64.9%)	587 (68.6%)	648 (68.2%)	
Mixed	7 (7.4%)	30 (3.5%)	37 (3.9%)	
Other	7 (7.4%)	47 (5.5%)	54 (5.7%)	
Missing	0	1	1	
BMI (kg/m^2)				0.003
Mean (SD)	26.5 (6.0)	24.7 (5.0)	24.8 (5.1)	
Range	17.3 - 49.6	15.8 - 46.4	15.8 - 49.6	
Missing	19	15	34	
Number of previous miscarriages				0.0013
0	39 (41.9%)	431 (50.3%)	470 (49.5%)	
1	28 (30.1%)	271 (31.6%)	299 (31.5%)	
2	14 (15.1%)	113 (13.2%)	127 (13.4%)	
3+	12 (12.9%)	42 (4.9%)	54 (5.7%)	
Missing	1	0	1	
Certainty LMP				0.802
Mean (SD)	9.8 (0.6)	9.8 (0.6)	9.8 (0.6)	
Range	7.0 - 10.0	7.0 - 10.0	7.0 - 10.0	
Smoking Status				0.096
No	79 (85.9%)	776 (91.2%)	855 (90.7%)	
Yes	13 (14.1%)	75 (8.8%)	88 (9.3%)	
Missing	2	6	8	
Folic Acid				0.002
No	13 (14.1%)	48 (5.6%)	61 (6.4%)	
Yes	79 (85.9%)	807 (94.4%)	886 (93.6%)	
Missing	2	2	4	
Parity				0.62
0	44 (47.3%)	440 (51.3%)	484 (50.9%)	

1	34 (36.6%)	300 (35.0%)	334 (35.2%)	
2+	16 (16.1%)	117 (13.7%)	132 (13.9%)	
Missing	1	0	1	
Gravidity				0.107
1	15 (16.1%)	203 (24%)	218 (22.9%)	
2+	78 (84%)	654 (76%)	732(77%)	
Missing	1	0	1	
LONGITUDINAL				
Features				
GA by LMP 1st scan (days)				< 0.001
Mean (SD)	44.5 (8.2)	48.3 (9.9)	47.9 (9.8)	
Range	29.0 - 74.0	18.0 - 82.0	18.0 - 82.0	
GA by LMP 2nd scan (days)				< 0.001
Mean (SD)	54.9 (8.3)	61.1 (10.9)	60.4 (10.8)	
Range	36.0 - 88.0	27.0 - 97.0	27.0 - 97.0	
Worst Bleeding Score 1st scan				0.692
0	57 (61.3%)	483 (57.4%)	540 (57.8%)	
1	18 (19.4%)	215 (25.5%)	233 (24.9%)	
2	15 (16.1%)	110 (13.1%)	125 (13.4%)	
3	2 (2.2%)	22 (2.6%)	24 (2.6%)	
4	1 (1.1%)	12 (1.4%)	13 (1.4%)	
Missing	1	15	16	
Worst Bleeding Score 2nd scan				0.163
0	60 (64.5%)	605 (73.1%)	665 (72.2%)	
1	20 (21.5%)	164 (19.8%)	184 (20.0%)	
2	9 (9.7%)	42 (5.1%)	51 (5.5%)	
3	2 (2.2%)	11 (1.3%)	13 (1.4%)	
4	2 (2.2%)	6 (0.7%)	8 (0.9%)	
Missing	1	29	30	
PUQE score 1st scan				0.004
Mean (SD)	3.7 (1.7)	4.4 (2.2)	4.3 (2.2)	
Range	3.0 - 10.0	3.0 - 15.0	3.0 - 15.0	
Missing	8	56	64	
PUQE score 2nd scan				< 0.001
Mean (SD)	3.9 (1.6)	5.3 (2.3)	5.1 (2.3)	
Range	3.0 - 10.0	3.0 - 15.0	3.0 - 15.0	
Missing	8	80	88	
Fetal heartbeat presence 1st scan				< 0.001
No	70 (80.5%)	268 (32.1%)	338 (36.7%)	
Yes	17 (19.5%)	567 (67.9%)	584 (63.3%)	

Missing	7	22	29	
Fetal heartbeat presence 2nd scan				< 0.001
No	22 (24.4%)	13 (1.5%)	35 (3.7%)	
Yes	68 (75.6%)	839 (98.5%)	907 (96.3%)	
Missing	4	5	9	
CRL at 1st scan (mm)				< 0.001
Mean (SD)	1.1 (2.8)	7.1 (8.1)	6.5 (7.9)	
Range	0.0 - 17.7	0.0 - 44.0	0.0 - 44.0	
Missing	1	15	16	
CRL at 2nd scan (mm)				< 0.001
Mean (SD)	6.0 (5.3)	19.8 (12.4)	18.4 (12.6)	
Range	0.0 - 23.9	0.0 - 72.0	0.0 - 72.0	
Missing	1	21	22	
MSD 1st scan (mm)				< 0.001
Mean (SD)	10.0 (7.6)	17.9 (10.2)	17.1 (10.3)	
Range	0.0 - 56.7	0.0 - 52.6	0.0 - 56.7	
Missing	2	33	35	
MSD at 2nd scan (mm)				< 0.001
Mean (SD)	16.7 (7.2)	31.6 (11.2)	30.1 (11.7)	
Range	4.2 - 39.5	4.1 - 67.0	4.1 - 67.0	
Missing	4	53	57	
Outcome at 1st scan				< 0.001
IPUV	78 (83.0%)	277 (32.3%)	355 (37.3%)	
Viable pregnancy	16 (17.0%)	580 (67.7%)	596 (62.7%)	
Outcome at 2nd scan				< 0.001
IPUV	26 (27.7%)	12 (1.4%)	38 (4.0%)	
Viable pregnancy	68 (72.3%)	845 (98.6%)	913 (96.0%)	
AGGREGATED FEATURES				
PUQE score (mean)				< 0.001
Mean (SD)	3.7 (1.3)	4.9 (1.9)	4.8 (1.9)	
Range	3.0 - 8.5	3.0 - 12.5	3.0 - 12.5	
Missing	13	120	133	
Worst bleeding score				0.508
0	46 (49.5%)	442 (51.8%)	488 (51.5%)	
1	21 (22.6%)	235 (27.5%)	256 (27.0%)	
2	19 (20.4%)	130 (15.2%)	149 (15.7%)	
3	4 (4.3%)	32 (3.7%)	36 (3.8%)	
4	3 (3.2%)	15 (1.8%)	18 (1.9%)	
Missing	1	3	4	
CRL growth (mm)				< 0.001
Mean (SD)	0.4 (0.5)	1.0 (0.5)	0.9 (0.5)	
Range	-4.6	0.0 - 10.2	-13	
Missing	2	35	37	

FH_code*				< 0.001
-2	22 (26.2%)	11 (1.3%)	33 (3.6%)	
-1	0 (0.0%)	2 (0.2%)	2 (0.2%)	
1	45 (53.6%)	253 (30.5%)	298 (32.6%)	
2	17 (20.2%)	564 (68.0%)	581 (63.6%)	
Missing	10	27	37	
MSD growth (mm)				< 0.001
Mean (SD)	0.7 (0.7)	1.1 (0.5)	1.0 (0.5)	
Range	-5.7	-10.5	-10.8	
Missing	6	84	90	
Maximum discrepancy in GA (days)				< 0.001
Mean (SD)	8.3 (7.5)	1.2 (7.5)	1.9 (7.7)	
Range	-33.6	-46	-46	

Table 3.A.3: Characteristics of the data subset used to develop the biomarker models in 3.3
Comparison of the cohort characteristics with regard to first trimester viability was assessed with the Student's t-test for continuous variables and the χ^2 test for binary or categorical variables. BMI: body mass index, CRL: crown rump length, FH: fetal heart, GA: gestational age, LMP: last menstrual period, MSD: mean sac diameter, MYSD: mean yolk sac diameter, PUQE: pregnancy unique quantification of emesis score, SD: standard deviation.

	Miscarriage (N=116)	**Viable** (N=312)	**Total** (N=428)	**p-value**
Demographic and clinical variables				
Maternal Age (years)				< 0.001
Mean (SD)	34.9 (5.1)	32.7 (5.4)	33.3 (5.4)	
Range	22.0 - 46.0	18.0 - 46.0	18.0 - 46.0	
Maternal Ethnicity				0.777
Black	11 (9.5%)	37 (11.9%)	48 (11.2%)	
Asian	13 (11.2%)	43 (13.8%)	56 (13.1%)	
White	78 (67.2%)	204 (65.4%)	282 (65.9%)	
Mixed	7 (6.0%)	13 (4.2%)	20 (4.7%)	
Other	7 (6.0%)	15 (4.8%)	22 (5.1%)	
BMI (kg/m^2)				0.346
Mean (SD)	25.6 (5.9)	25.0 (5.1)	25.1 (5.3)	
Range	17.3 - 49.6	15.8 - 43.8	15.8 - 49.6	
Missing	24	5	29	
Previous Miscarriage				0.657
0	60 (51.7%)	152 (48.7%)	212 (49.5%)	
1	31 (26.7%)	96 (30.8%)	127 (29.7%)	
2	14 (12.1%)	45 (14.4%)	59 (13.8%)	
3	7 (6.0%)	13 (4.2%)	20 (4.7%)	
4+	4 (3.4%)	6 (1.9%)	10 (2.3%)	
Certainty LMP(0-10)				0.435
Mean (SD)	9.7 (0.7)	9.8 (0.6)	9.8 (0.6)	
Range	7.0 - 10.0	7.0 - 10.0	7.0 - 10.0	
GA by LMP(days)				0.844
Mean (SD)	50.6 (12.1)	50.3 (12.9)	50.4 (12.6)	
Range	20.0 - 89.0	18.0 - 96.0	18.0 - 96.0	
Smoking Status				0.483
No	102 (89.5%)	286 (91.7%)	388 (91.1%)	

Yes	12 (10.5%)	26 (8.3%)	38 (8.9%)	
Missing	2	0	2	
Folic acid supplementation				0.009
No	18 (15.8%)	23 (7.4%)	41 (9.6%)	
Yes	96 (84.2%)	289 (92.6%)	385 (90.4%)	
Missing	2	0	2	
Parity				0.885
0	57 (49.1%)	164 (52.6%)	221 (51.6%)	
1	41 (35.3%)	108 (34.6%)	149 (34.8%)	
2	11 (9.5%)	27 (8.7%)	38 (8.9%)	
3	3 (2.6%)	7 (2.2%)	10 (2.3%)	
4+	4 (3.4%)	6 (1.9%)	10 (2.3%)	
Gravidity				0.669
1	24 (20.7%)	72 (23.1%)	96 (22.4%)	
2	38 (32.8%)	91 (29.2%)	129 (30.1%)	
3	18 (15.5%)	61 (19.6%)	79 (18.5%)	
4+	36 (31.0%)	88 (28.2%)	124 (29.0%)	
Worst Bleeding Score (0-4)				< 0.001
0	66 (57.4%)	231 (75.5%)	297 (70.5%)	
1	26 (22.6%)	44 (14.4%)	70 (16.6%)	
2	18 (15.7%)	26 (8.5%)	44 (10.5%)	
3	1 (0.9%)	4 (1.3%)	5 (1.2%)	
4	4 (3.5%)	1 (0.3%)	5 (1.2%)	
Missing	1	6	7	
PUQE score (3-15)				0.004
Mean (SD)	3.9 (1.7)	4.6 (2.3)	4.4 (2.2)	
Range	3.0 - 10.0	3.0 - 14.0	3.0 - 14.0	
Missing	6	16	22	
Sonographic variables				
Fetal heartbeat (yes-no)				< 0.001
No	74 (68.5%)	112 (37.7%)	186 (45.9%)	
Yes	34 (31.5%)	185 (62.3%)	219 (54.1%)	
Missing	8	15	23	
CRL (mm)				< 0.001
Mean (SD)	3.0 (4.8)	8.6 (10.4)	7.1 (9.5)	
Range	0.0 - 18.7	0.0 - 55.3	0.0 - 55.3	
Missing	1	4	5	
MSD (mm)				< 0.001
Mean (SD)	13.3 (8.8)	19.5 (12.5)	17.8 (11.9)	
Range	2.5 - 56.7	2.4 - 55.9	2.4 - 56.7	

Missing	1	14	15	
MYSD (mm)				0.748
Mean (SD)	2.5 (3.7)	2.6 (1.5)	2.6 (2.2)	
Range	0.0 - 34.0	0.0 - 9.7	0.0 - 34.0	
Missing	20	34	54	
Difference in GA (LMP-MSD) (days)				< 0.001
Mean (SD)	7.3 (10.6)	0.5 (8.6)	2.4 (9.6)	
Range	-96.2	-84.3	-101	
Missing	2	14	16	
Biomarkers				
Progesterone (ng/mL)				< 0.001
Mean (SD)	15.9 (10.9)	26.3 (12.8)	23.5 (13.1)	
Range	0.0 - 86.8	0.0 - 131.3	0.0 - 131.3	
Missing	1	3	4	
β-hCG(mIU/mL)				< 0.001
Mean (SD)	17178.0 (21368.5)	45788.6 (40043.2)	38213.2 (38180.2)	
Range	42.9 - 110630.0	590.0 - 193318.0	42.9 - 193318.0	
Missing	4	1	5	
∝-FP (ng/mL)				0.101
Mean (SD)	2.5 (1.8)	2.9 (2.3)	2.8 (2.2)	
Range	0.0 - 13.8	0.0 - 25.6	0.0 - 25.6	
CA-125 (U/mL)				0.645
Mean (SD)	44.6 (54.9)	47.7 (64.5)	46.9 (62.0)	
Range	6.7 - 395.4	6.2 - 826.2	6.2 - 826.2	
Missing	3	1	4	

Chapter 4

Reclassification of kidney transplant biopsies

This chapter has been previously published as:
Vaulet, T., Divard, G., Thaunat, O., *et al.* Data-Driven Derivation and Validation of Novel Phenotypes for Acute Kidney Transplant Rejection Using Semi-supervised Clustering. *JASN.* ISSN: 1046-6673, 1533-3450 (Mar. 2021)

Most of the DDSS are developed based on established diagnostic criteria or accepted classifications. For instance, the previous chapter described the development of a DDSS for first trimester viability prediction based on well-defined miscarriage criteria and study end-points. In the following project, instead of providing additional guidance in the form of a diagnostic or a prognotic prediction, we first aimed to improve the diagnostic classification itself and then develop an intuitive web-based DDSS to allow physicians to classify their own data according to the new phenotypes.

Objectives

Develop and validate a new classification of acute kidney transplant rejection with semi-supervised clustering on extensive histological datasets and implement the new classification in an intuitive web-based application.

4.1 Introduction

Kidney transplant biopsies are crucial in the follow-up of patients after transplantation. Both at time of graft dysfunction (indication biopsies) and at time of stable graft function (protocol biopsies), the histological evaluation of these biopsies enables to distinguish rejection mechanisms from other injury processes and to orient the appropriate therapeutic interventions. Over the past decades, an international group of experts has developed a consensus classification of kidney transplant rejection phenotypes, known as the Banff classification [167–169].

The Banff classification relies on the histological evaluation of a set of well-defined lesions, further translated into semiquantitative, ordinal lesion scores [170]. The diagnostic classification process consists of a set of *if-then* rules that map conditional clauses based on lesion scores to a diagnosis category. Currently, the Banff classification encompasses 5 main categories [169] : *(1) Normal biopsy or non-specific changes; (2) Antibody-mediated changes (ABMR); (3) Borderline changes; (4) T-cell mediated rejection (TCMR);* and *(5) Polyomavirus nephropathy.* Several of these categories are further subdivided into subtypes. This classification was developed iteratively, based on studies that examine the associations between lesions and risk factors like donor-specific HLA antibodies, between lesions and graft failure, and among lesions themselves [171–173]. Banff diagnostic categories are not mutually exclusive and Banff lesions are not specific for disease processes, which leads to overlapping diagnoses and mixed rejection phenotypes. Although this reflects a histological reality, the clinical interpretation of this complex categorization process is difficult, leading to unstable clinical decisions.

Instead of this iterative consensus process for disease classification, data-driven mathematical modelling of the multidimensional histological data could be appropriate. Such approach could refine the thresholds for the diagnostic phenotypes, simplify the complex and discretionary *if-then* rules, avoid the issue of mixed phenotypes, and yield new phenotypes and disease reclassification. Categorizing data into groups without pre-existing labels is commonly referred as unsupervised clustering [81]. Although the resulting clusters (reclassified disease phenotypes) might be valid from a mathematical perspective, there is no guarantee that they will show relevant association with external outcome variables. To overcome this, introducing information on outcome in the clustering process could be of interest [42, 174, 175]. Whether such mathematical modelling approach would also be applicable to the classification of kidney transplant rejection, has not been evaluated yet.

On the basis of these considerations, we built and externally validated a model

for mathematical reclassification of acute kidney transplant rejection, based on the integration of the set of inflammatory lesions in kidney transplant biopsies, informed by graft failure, in a retrospective observational cohort study.

4.2 Method

4.2.1 Data

Patients and biopsies

For the training cohort, all consecutive adult recipients of a kidney transplant at the University Hospitals-Leuven between March 2004, the start of the protocol biopsy program, and February 2013 were eligible for this study (n=1137). A minimal of 5 years follow-up at time of data extraction (March 2018) was required. Recipients of combined transplantation (n=113) or kidney transplantation after another solid organ transplantation (n=24) were excluded. All transplants were performed with negative complement-dependent cytotoxicity crossmatches. The clinical data were collected during routine clinical follow-up in electronic medical records, which were used for clinical patient management and directly linked to the SAS database from which the research database was extracted. The standard immunosuppressive maintenance regimen consisted of tacrolimus, mycophenolate and corticosteroids [176]. The histological data consisted of all 3622 kidney transplant biopsies performed at the Leuven University Hospitals between April 2004 and February 2015 in 949 patients. Biopsies were performed upon medical indication (indication biopsies at time of graft dysfunction) or as part of an established follow-up protocol (protocol biopsies) [177]. Biopsies with missing lesion scores were excluded (n=112), due to missing HLA-DSA (n=73) and/or missing the score of C4d deposition in peritubular capillaries (n=40). 3510 biopsies from 936 recipients remained available for analysis. This study was approved by the Ethical Committee of the University Hospitals Leuven (S64006).

For the validation cohort, the electronic database of Lyon University Hospitals (registration #AC-2016-2706) and the Paris Transplant Group were screened with the same selection criteria as detailed above. Between January 2007 and December 2015 for the Lyon dataset, and between March 2009 and October 2019 for the Paris dataset, respectively 1356 (from 726 transplants) and 2479 biopsies (from 1304 transplants) were included as an independent validation set, performed either for indication or as part of the routine follow-up at 3 months and 12 months post transplantation). Only complete data were included. Clinical, histological and immunological data were extracted from

these databases, anonymized, and transmitted to Leuven to be used as an external independent validation cohort.

Histological scoring

In the training cohort, all post-transplant kidney allograft biopsies performed in this cohort, until the time of data extraction in December 2018, were included. One pathologist (EL) reviewed all biopsies, independent of clinical information to avoid bias. The severity of the histological lesions was semi-quantitatively scored according to the Banff categories with a small deviation for C4d thresholds [176]. The set of individual Banff lesions (N=14) represents either acute or chronic injury processes. We focused on the following 7 acute Banff lesions, with semiquantitative scores reflecting disease activity, in concordance with the Banff guidelines [170]: tubulitis (t; score 0 to 3), interstitial inflammation (i; score 0 to 3), glomerulitis (g; score 0 to 3), intimal arteritis (v; score 0 to 3), C4d deposition in peritubular capillaries (C4d; score 0 to 3), peritubular capillaritis (ptc; score 0 to 3), thrombotic microangiopathy (TMA; present vs. absent). We considered transplant glomerulopathy (cg; score 0 to 3), interstitial fibrosis (ci; score 0 to 3), tubular atrophy (ct; score 0 to 3), vascular intimal thickening (cv; score 0 to 3), mesangial matrix increase (mm; score 0 to 3), arteriolar hyalinosis (ah; score 0 to 3) and glomerulosclerosis (gs; score 0 to 3) as chronic lesions and did not take these lesions into account in the classification of acute rejection phenotypes. As the presence of HLA-DSA is a defining feature in the Banff diagnosis of ABMR, HLA-DSA was also considered in the clustering process (present vs. absent), as defined previously for this cohort [64].

The biopsies were classified into acute rejection categories based on the criteria as defined by the most recent Banff 2019 consensus [169]. Overall, each biopsy was assigned to one of the six following categories based on the Banff acute rejection phenotype: *(1) No rejection, (2) Borderline changes, (3) TCMR, (4) ABMR, (5) Mixed borderline rejection* and *(6) Mixed rejection.* Borderline changes were diagnosed as foci of tubulitis ($t > 0$) with minor interstitial inflammation (i1) or moderate-severe interstitial inflammation (i2 or i3) with mild (t1) tubulitis. Antibody-mediated rejection (ABMR) was diagnosed by the presence of the three Banff criteria for either acute or chronic active ABMR according to the Banff 2019 classification, but not taking into account potential non-HLA antibodies or gene expression changes. Due to lack of information on i-IFTA and total-i scores, chronic T-cell mediated rejection was not considered separately. We labeled biopsies presenting an overlap of ABMR and TCMR as Mixed rejection and the biopsies with an overlap of ABMR and Borderline changes as Mixed borderline rejection.

4.2.2 Data analysis

Semi-supervised clustering strategy

We scaled each histological lesion score (feature) into the unit interval. We adapted semi-supervised learning from [178], where additional information was used to facilitate the creation of clinically meaningful clusters. Specifically, Bair and Tibshirani used the Cox scores from univariate models to perform a feature selection prior to clustering [178], whereas we used the Cox score to weigh the features. We chose k-means as the core algorithm for the clustering process because of its straightforward implementation, its efficiency, its ability to accommodate the weighting of features and the possibility to classify new biopsies into non-overlapping clusters. The information from the death-censored kidney transplant survival outcome was introduced with a weighted Euclidean distance to provide additional guidance during the clustering process. Each feature was weighted with the normalized coefficient's z score of univariate Cox models, adjusted for clustered data, i.e. repeated biopsies from the same patients, using a *sandwich* variance estimate. Features with a higher weight contribute more heavily to the notion of dissimilarity between clusters than low weight features which will be less relevant to the definition of a cluster. Although guided by external survival information, the clustering task remains mostly unsupervised as the lesion scores patterns are the most influential driving force in the final clusters.

Consensus clustering

We used consensus clustering [179] based on 400 clustering partitions of the data with different random initializations of the k-means algorithm seed and a different subsampling (80%) of the original data, similar to the approach used by Monti [180]. For the clustering process, all biopsies were considered independent. We used the nearest centroid method to assign a cluster label to the remaining 20% of out-of-bag biopsies for each partition. The final consensus clustering was achieved through majority voting along the 400 partitions. A visual overview of the semi-supervised consensus clustering methodology is displayed in Figure 4.1. The pseudo-code for a semi-supervised consensus clustering is reported in Algorithm 5.

To avoid introducing biases in the clustering process by the overrepresentation of protocol biopsies, we adopted a scheme where indication biopsies and protocol biopsies were weighted based on the inverse of their total proportion in the dataset. Cluster profiles were reported using the normalized mean value of lesions, or for binary features the percentage of biopsies with the feature present.

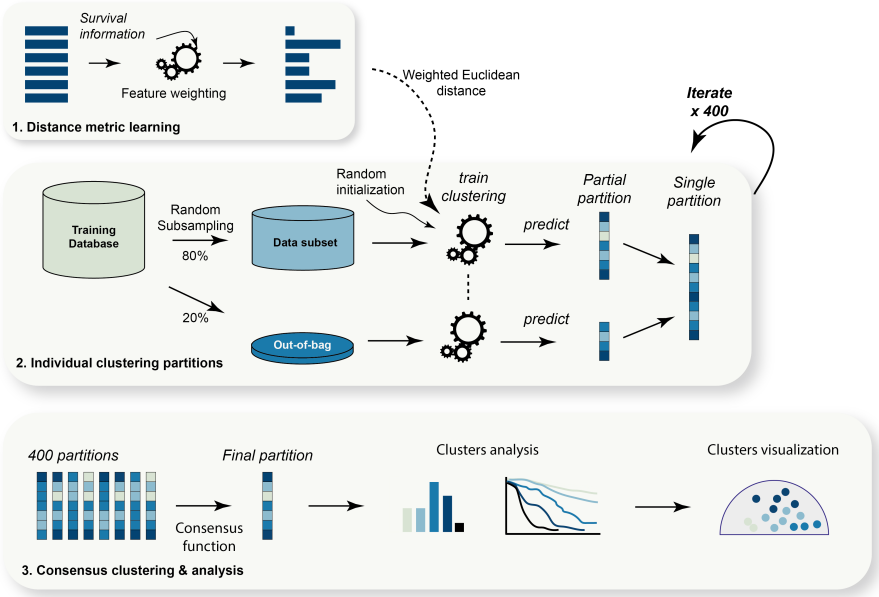


Figure 4.1: Overview of the semi-supervised consensus clustering for kidney transplant biopsies.
1. Individual variable weights are derived from the training data based on their overall association with graft survival. **2.** To create a single partition of the data, the k-means clustering algorithm is applied on a subsample (80%) of the training data with the weighted Euclidean distance from step 1. The remaining 20% of out-of-bag biopsies are classified with the nearest centroid method from the trained k-means algorithm. This process is repeated over different subsets and random initializations to create 400 different partitions. **3.** The final clustering partition is obtained with a consensus function (i.e. majority voting) over the 400 partitions. The whole data cohort is then analyzed and visualized on a cluster-basis in terms of lesion scores distribution and graft survival.

We also report the proportion of each original lesions score. Where appropriate, individual lesions scores were compared between a pair of clusters with a χ^2 test. The degree of similarity between two different partitions of the data was evaluated with the adjusted rand index (ARI). This index accounts for overlapping partitions due to chance. It varies from -1 to 1, an ARI of 0 meaning random partitioning. A decision tree was trained on the cluster-labelled data to mimic the internal clustering process. The tree was generated using the Gini criterion with a minimum of 10 biopsies per leaf.

Algorithm 5: Generic semi-supervised consensus clustering
$\mathbf{X}$ is the original data to cluster, $\boldsymbol{z}$ is an additional proxy variable to guide the clustering (e.g. outcome variable), P is the number of partitions to evaluate, σ is the subsample parameter, $Cluster$ is a disance-based clustering algorithm (e.g. k-means), Θ is a set of clustering configuration to be tested (e.g. various combinations k and initialization values) with $|\Theta| = m$ and f_{cons} is a consensus function (e.g. majority voting) to derive the final clustering parition $\boldsymbol{c}^*$ over M the whole set of computed partitions ($M = m \times P$).

Input: $\boldsymbol{X}, \boldsymbol{z}$
parameters: $P, \sigma, Cluster, \Theta, f_{cons}$
Output: $\boldsymbol{c}^*$

for *p= 1,...,P* **do**
  1. Subsample $\boldsymbol{X}_s^{(p)}$ from $\boldsymbol{X}$ based on σ
  2. Derive supervised distance metric $d^{(p)}$ from $(\boldsymbol{X}_s^{(p)}, \boldsymbol{z}^{(p)})$
  for $\theta_1, \theta_2, ..., \theta_m$ **do**
    3. Train $Cluster_{\theta_i}^{(p)}$ with $d^{(p)}$ on $\boldsymbol{X}_s^{(p)}$
    4. Predict $\boldsymbol{c}_{\theta_i}^{(p)} = Cluster_{\theta_i}^{(p)}(\boldsymbol{X})$
  end
end
Compute $\boldsymbol{c}^* = f_{cons}(M)$, where M is the set of all partitions $\{\boldsymbol{c}_{\theta_1}^{(1)}, \boldsymbol{c}_{\theta_1}^{(2)}, ..., \boldsymbol{c}_{\theta_i}^{(p)}, ..., \boldsymbol{c}_{\theta_m}^{(P)}\}$

Tuning of parameters

To define the optimal number of clusters, we used the proportion of ambiguous clustering (PAC) [181] to assess the stability of our results at different values of k, i.e. the number of clusters, with thresholds set at 10.0% and 90.0% of consensual clustering. Intuitively, PAC measures the proportion of all possible pairs of biopsies from the whole dataset which demonstrate inconsistent clusters attributions over the 400 partitions. The lower the PAC, the more stable the

clustering across different conditions. We discarded very low value of k, since they only create a restricted number of clusters (typically No rejection vs. any rejection with k=2), which is not helpful to describe different phenotypes.

Biopsy stability

In order to identify biopsies that are part of pairs with an unstable cluster assignment over the set of clustering partitions, we developed an empirical individual stability score based on the consensus matrix. The consensus matrix C is a$n \times n$ matrix, where n is the total number of biopsies and entries $C_{i,j}$ represent the proportion of times biopsies i and j are clustered in the same cluster over the whole set of different partitions. The stability score s_i for biopsy i was defined as $s_i = \frac{2}{N} \sum_{j=1}^{N} |C_{i,j} - 0.5|$ where $C_{i,j}$ represented the value from the consensus matrix at the i^{th} row and j^{th} column and N the total number of biopsies. Intuitively, a theoretical score of 1 reflects that the biopsy was consistently part of biopsy pairs that are either clustered together in the same cluster or clustered in different clusters. A theoretical score of 0 would reflect a biopsy that forms ambiguous pairs with any other biopsy.

Survival analysis

Graft survival times are reported as number of days until graft failure, calculated from each biopsy date. Patients were administratively censored at the of last follow-up date or at time of death. Survival curves are plotted with Kaplan-Meier estimators along with the 95% confidence interval. To avoid artificially increasing the incidence of transplant failure events due to repeated biopsies in a given individual, survival times from repeated biopsies in a given cluster were averaged for each patient. Pairwise comparison of survival curves was performed using Cox modelling and hazard ratio (HR) with 95% confidence interval. Because potentially, proportional hazard assumption violations might bias the HR, we also report the restricted mean survival times (RMST) [182] at 5 and 10 years and its confidence interval at 95%. This measure can be interpreted as the mean survival time without event within a pre-defined time range, representing the area under the survival curve up to a pre-defined time-point. We also report the differences in RMST (DRMST) with a baseline category, which estimates the difference in average event-free survival, in years, between a given category and the baseline group.

Visualization

Principal component analysis (PCA) was performed on the Cox-score weighted acute lesions scores and the first two components were used for 2D visualization purposes. To better visualize the heterogeneity in the acute lesion scores, we developed a two-dimensional plot using polar coordinates, with the radius calculated as the sum of re-weighted acute lesions scores, scaled to the unit interval (from 0 to 1), and the theta angle is a scaled version (for visual purposes) of the second component of the PCA, which is directly related to the main rejection phenotype. Because the sum of lesions is directly related to graft failure due to the individual weighting of lesions scores, this approach combines the severity and the phenotype trend in one single plot.

All analyses have been performed with Python 3.6 [79]. A web application where others can upload their own patient data and derive the clusters from the individual Banff lesion scores is available at https://rejectclass.pythonanywhere.com. The interface is reported on Figure 4.A.11.

4.3 Results

4.3.1 Patient and biopsy characteristics

Descriptive patient (N=936) and biopsy (N=3510) data of the training cohort are shown in Table 4.1. On average, 3.75 biopsies (range = 1 to 11) were performed per patient. Of the 773 indication biopsies, 644 (83.3%) were performed within the first year of transplantation (median at 22 days post-transplantation), and 129 (16.7%) after one year. HLA-DSA were present at the time of 468 (13.3%) biopsies.

4.3.2 Semi-supervised clustering of rejection phenotypes

Fully unsupervised clustering of our biopsy cohort (N=3510) yielded an optimum of 4 different clusters, based on the proportion of ambiguous clustering (Figure 4.A.1). Compared to cluster 1 (essentially normal biopsies), the three other clusters associated significantly with impaired graft survival. However, their histological and clinical relevance were less clear, as none of these three clusters were defined based on microcirculation inflammation and antibody activity (glomerulitis, peritubular capillaritis and C4d), suggesting that the number of clusters was insufficient to reflect the clinical reality and previous knowledge on the relevance of these lesions and ABMR. Increasing the number of clusters

Table 4.1: Demographic, clinical and histological characteristics of the patients and biopsies included

Cohort characteristics	**Total (N=936)**
Donor demographics	
Donor type	
Donation after brain death, N (%)	726 (77.6)
Donation after cardiac death, N (%)	153 (16.3)
Living donation, N (%)	57 (6.1)
Age (years), mean ± SD	47.7 ± 14.7
Male, N (%)	497 (53.1)
Diabetes, N (%)	24 (2.6)
Recipient demographics	
Age (years), mean ± SD	53.5 ± 13.3
Male, N (%)	572 (61.1)
Ethnicity	
Caucasian, N (%)	920 (92.3)
African, N (%)	12 (1.3)
Asian, N (%)	3 (0.3)
Hispanic, N (%)	1 (0.1)
BMI (kg/m^2), mean (range)	25.4 (4.5)
Pre-transplant donor-specific HLA antibodies, N (%)	408 (11.6%)
Repeat transplantation, N (%)	141 (15)
Cold ischemia time (hours), mean ± SD	14.2 ± 5.7
Total number of HLA A/B/DR mismatches, mean ± SD	2.8 (1.3)
Biopsy characteristics	**Total (N=3510)**
Banff 2019 diagnosis	
No rejection, N (%)	2671 (76.1)
Borderline changes, N (%)	333 (9.5)
TCMR, N (%)	314 (8.9)
ABMR, N (%)	110 (3.1)
Mixed rejection (ABMR + TCMR), N (%)	61 (1.7)
Mixed borderline rejection (ABMR + borderline changes), N (%)	21 (0.6)
Indication biopsies, N (%)	n=773 (22.0)
Days since transplantation, median (IQR)	22 (8-96)
eGFR at day of biopsy, median (IQR)	19.8 (10.9-29.0)
Protocol biopsies, N (%)	n=2737 (78.0)
3 months, N (%)	823 (30.1)
12 months, N (%)	759 (27.7)
24 months, N (%)	639 (23.3)
36 months, N (%)	205 (7.5)
48 months, N (%)	22 (0.8)
60 months, N (%)	289 (7.6)
Days since transplant, median (IQR)	377 (100-752)
eGFR at day of biopsy, median (IQR)	46.4 (36.5-57.8)

created clusters that were no longer associated with impaired graft survival compared to cluster 1 (Table 4.A.1).

To optimize the clinical significance of the clusters, we applied a semi-supervised clustering approach, weighing the histological features with survival information. The optimal number of clusters (k) was six, based on the proportion of ambiguous clustering (Table 4.A.1). We labeled the six identified clusters from 1 to 6, according to the overall association with graft failure (Figure 4.2; Table 4.A.2). Biopsies in cluster 1 were dominated by 0 scores for the lesions, in cluster 2 by high g scores, and in cluster 3 by t and i. Clusters 1 to 3 were HLA-DSA negative, while all biopsies included in clusters 4 to 6 were in patients with HLA-DSA. Biopsies in cluster 1 had no or very limited inflammation, and good outcome, and could be considered as No Rejection. In cluster 2, all cases had moderate to severe glomerulitis in the absence of HLA-DSA, sometimes accompanied by tubulo-interstitial inflammation and peritubular capillaritis. These cases of glomerulitis in the absence of HLA-DSA are currently not fully understood and not reflected in the current Banff classification, yet associate with impaired graft outcome. Cluster 3 is characterised by moderate to severe degrees of tubulo-interstitial inflammation, reminding of acute TCMR in the Banff classification. In cluster 4, no or only very limited inflammation is noted, sometimes C4d deposition in peritubular capillaries. All cases in cluster 4 were HLA-DSA positive, which appeared to be a risk factor for graft failure in comparison to cluster 1, even in the absence of extensive inflammation. Biopsies in cluster 5 were HLA-DSA positive with high g scores and could be considered to reflect active ABMR. Biopsies in cluster 6 were HLA-DSA positive with high t and i scores, often combined with g and ptc, and could suggest "Mixed rejection". Biopsies in cluster 1 and 4 were most often protocol biopsies, while clusters 2, 3 and 5 were similarly distributed between protocol and indication biopsies (Table 4.A.3). Cluster 6 was observed most often in indication biopsies and had worst eGFR and highest proteinuria.

Although we focused on acute histological lesions, these lesions often co-occurred with chronic lesions (Figure 4.A.2). With k smaller or greater than six, we observed larger PAC (Table 4.A.1). Increasing k did not drastically reshape previously found clusters, but rather added new clusters while conserving the similar centroids of the clusters derived at lower k. With k=7, we observed the separation of cluster 1 in two clusters based on the t lesion (and to a lesser extent i). However, the survival curves from those two sub-clusters were largely overlapping (log-rank test p-value 0.97), illustrating that also from clinical perspective, the optimal number of clusters was k=6. The ARI was similar between the various k values and the minimal distance between two centroids also decreased with greater k.

Based on the consensus matrix, the average stability scores per cluster were

0.98, 0.98, 0.99, 0.98 and 0.99 for clusters 1, 3, 4, 5 and 6, respectively. In comparison to the other clusters, cluster 2, characterized by glomerulitis in the absence of HLA-DSA, was less stable with an average stability score of 0.75. 44 biopsies had a low stability score (< 0.5): 30 biopsies from cluster 2 (29.7%) and 14 biopsies from cluster 1 (0.5%). Because k-mean is a distance-based algorithm, it is possible to compute relative distances to the closest clusters' boundary. If a biopsy is nearer to a cluster centroid than it is from the second closest centroid, the relative distance will be small. On the other hand, a biopsy that is near the clusters' boundary will get a relative distance approaching 1, translating an almost equidistant position (Figure 4.A.3). Biopsies with low stability scores were mostly found on the cluster edges.

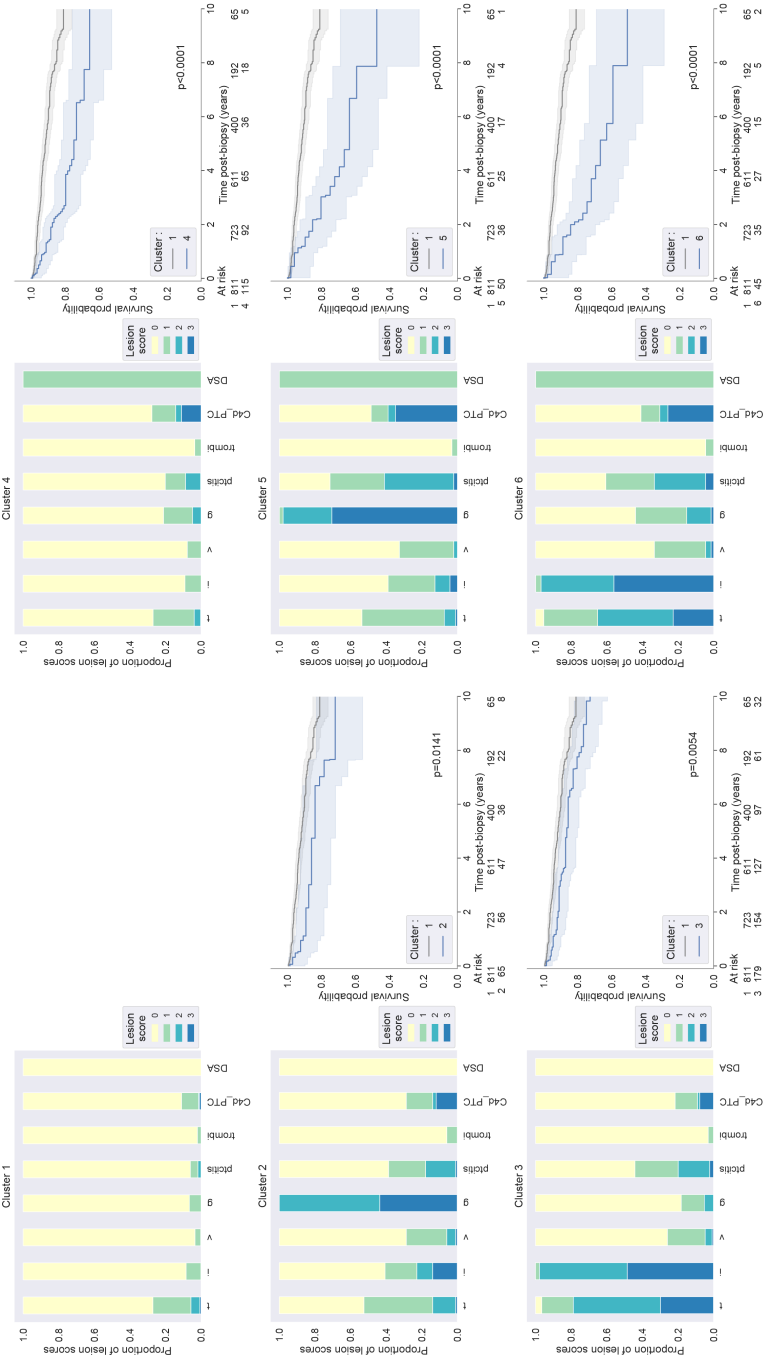


Figure 4.2: Distribution of the individual acute lesion scores in the different clusters, and post-biopsy Kaplan-Meier graft survival curves relative to cluster 1 of the derivation cohort (N=3510 biopsies).

Figure 4.2: Biopsies included in cluster 1 were dominated by 0 scores (more than 90 % have 0 score in all lesions, except for t (t0 in 72.9%) and C4d (C4d0 in 89.0%). High g scores drove cluster 2 (56.4% g2 and 43.6% g3; no biopsies with g0 or g1). Compared to cluster 1, biopsies in cluster 2 had a higher proportion of score 1 or 2 acute lesions other than g. High t and i scores dominated cluster 3 biopsies (48.9% t2, 29.9% t3; 49.4 % i2 and 48.5% i3). Biopsies in cluster 3 also had a higher proportion of score 1 and 2 acute lesion scores compared to cluster 1, but no g score 2 or higher. Cluster 4 was similar to cluster 1 and was dominated by low acute lesion scores. The main differences besides the presence of DSA was the higher proportion of g (g1 in 16.3% in cluster 4 vs. 6.6% in cluster 1; g2 4.9% in cluster 4 vs. 0.0% in cluster 1; p=<0.0001), a higher proportion of ptc (ptc1 in 11.4% in cluster 4 vs. 4.2% in cluster 1; ptc2 in 8.5% in cluster 4 vs. 1.2% in cluster 1; ptc3 in 0.3% in cluster 4 vs. 0.1% in cluster 1; p=<0.0001), and a higher proportion of C4d (C4d1 in 13.4% in cluster 4 vs. 9.5% in cluster 1; C4d2 in 3.3% in cluster 4 vs. 0.5% in cluster 1; C4d3 in 11.1% in cluster 4 vs. 1.0% in cluster 1, p=<0.0001). Cluster 5, similarly to cluster 2, was dominated by high g scores (27.4% g2 and 70.5% g3) and did not contain biopsies without g. As in cluster 2, we noted a higher proportion of score 1 and 2 acute lesions (ptc, t, I, v) compared to the cluster 1. Finally, in the presence of DSA, high t and i scores determined cluster 6 (42.4% t2, 22.7% t3; 40.9 % i2 and 56.1% i3), and frequent presence of g and ptc. t=tubulitis, i=interstitial inflammation, g= glomerulitis, v=intimal arteritis, C4d_ptc=C4d deposition in peritubular capillaries, ptc=peritubular capillaritis, thrombi=thrombotic microangiopathy, DSA= donor specific anti-HLA antibodies. P-values refer to HR from the Cox models.)

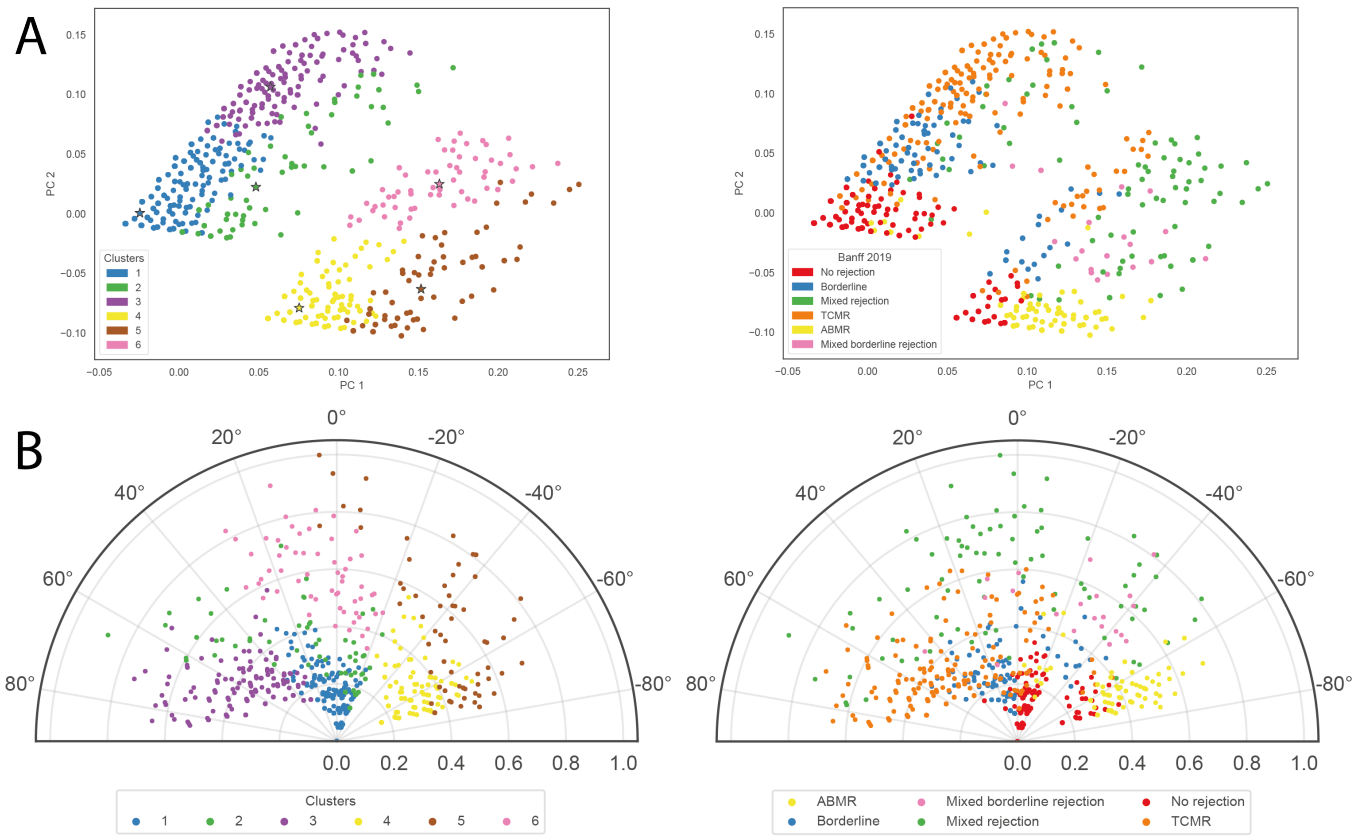


Figure 4.3: Visualization of the Banff classification and the 6 clusters on the whole set of kidney transplant biopsies

Figure 4.3: **A.** Principal component analysis of the 3510 derivation cohort biopsies calculated from the acute lesion scores and DSA status, overlaid with the 6 clusters obtained from the semi-supervised reclassification pipeline (left panel) and according to the Banff 2019 classification (right panel). Due to the distance-based approach of k-means, the clusters obtained have a visually better separation than the Banff classification on 2-dimensional plots. **B.** Polar plot of the 3510 biopsies, with the radius representing the sum of re-weighted acute lesions scores, scaled to the unit interval (from 0 to 1), and the theta angle being directly related to the phenotype using the second semi-supervised principal component, i.e. the second component of PCA after weighting the lesions scores, overlaid with the 6 clusters obtained from the semi-supervised reclassification pipeline (left panel) and according to the Banff 2019 classification (right panel).

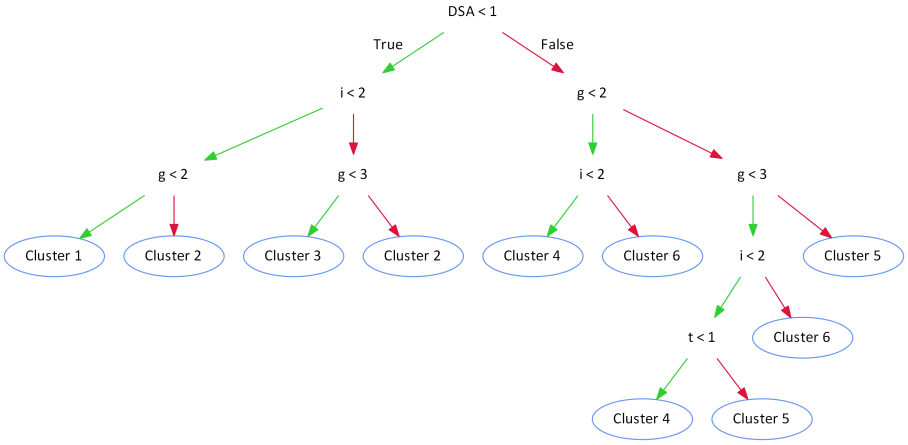


Figure 4.4: Decision tree of the clustering process
This decision tree was constructed with a minimum sample size in leaf nodes of 10 biopsies. It demonstrates that using only 4 parameters - DSA, i, t, g – is sufficient to be assigned in the correct cluster with 97% of balanced accuracy.

4.3.3 Comparison of disease clusters with Banff 2019 rules

There was important overlap between the clusters and the Banff categories with an ARI of 0.48 (Table 4.2; Supplemental Results 4.A). Due to its distance-based approach, the clustering algorithm led to better separation of the biopsies than the Banff 2019 classification, as also illustrated in the plots of PCA applied on the weighted acute lesions scores (Figure 4.3). Although all lesions were taken into account simultaneously to assign each biopsy to a cluster, a decision tree could be derived, based on the four main driving forces, g, HLA-DSA, i and t (Figure 4.4). This decision tree assigned the correct cluster with 97.0% of balanced accuracy, which confirmed the dominance of these 4 lesions in the phenotype reclassification. The 3% misclassified cases related to 24 biopsies.

4.3.4 Quantitative visual presentation of disease clusters

As expected, the superposition of the 6 disease clusters on the 2-D polar plot aligned better visually with the mathematical disease reclassification than with the different Banff 2019 phenotypes (Figure 4.3). Biopsies projected with a negative angle were mostly associated with Banff TCMR, whereas those with a positive angle represented Banff ABMR. Biopsies with mixed rejection

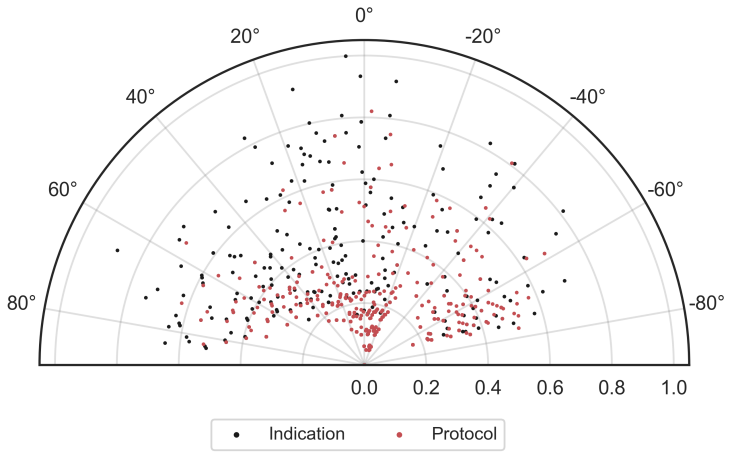


Figure 4.5: Comparison of indication vs. protocol biopsies, as superposed on the polar plot.
The radius from indication biopsies was on average higher compared to protocol biopsies (mean ± sd : 0.22 ± 0.23 vs. 0.08 ± 0.13 respectively, student t test p<0.0001), illustrating more inflamed biopsies at time of graft dysfunction than at time of stable graft function.

phenotypes were projected around 0 °. When plotting individual lesions, and also combinations of those (g+ptc = microcirculation inflammation) and (i+t = tubulo-interstitial inflammation)(Figure 4.A.4), the PCA and the theta values associated with these two major components (microcirculation inflammation vs. tubulo-interstitial inflammation) driving the disease reclassification. The radius on the plot was higher in indication biopsies compared to protocol biopsies (mean ± sd : 0.22 ± 0.23 vs. 0.08 ± 0.13 respectively, student t test p<0.0001), illustrating more inflamed biopsies at time of graft dysfunction than at time of stable graft function (Figure 4.5).

Table 4.2: Contingency tables comparing the Banff 2019 diagnosis and the 6 clusters derived from semi-supervised learning. Proportions represent the distribution in the clusters per Banff category (N=3510 biopsies).

Banff 2019 diagnosis	**N**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
No rejection	2659	2387 (89.8%)	53 (2.0%)	4 (0.2%)	215 (8.1%)	0 (0.0%)	0 (0.0%)
Borderline changes	327	261 (79.8%)	9 (2.8%)	26 (8.0%)	23 (7.0%)	0 (0.0%)	8 (2.4%)
TCMR	285	48 (16.8%)	25 (8.8%)	184 (64.6%)	5 (1.8%)	0 (0.0%)	23 (8.1%)
ABMR	122	8 (6.6%)	4 (3.3%)	0 (0.0%)	56 (45.9%)	53 (43.4%)	1 (0.8%)
Mixed borderline rejection	27	1 (3.7%)	3 (11.1%)	1 (3.7%)	3 (11.1%)	15 (55.6%)	4 (14.8%)
Mixed rejection	90	5 (5.6%)	7 (7.8%)	16 (17.8%)	5 (5.6%)	27 (30.0%)	30 (33.3%)
Total	**3510**	**2710 (77.2%)**	**101 (2.9%)**	**231 (6.6%)**	**307 (8.7%)**	**95 (2.7%)**	**66 (1.9%)**

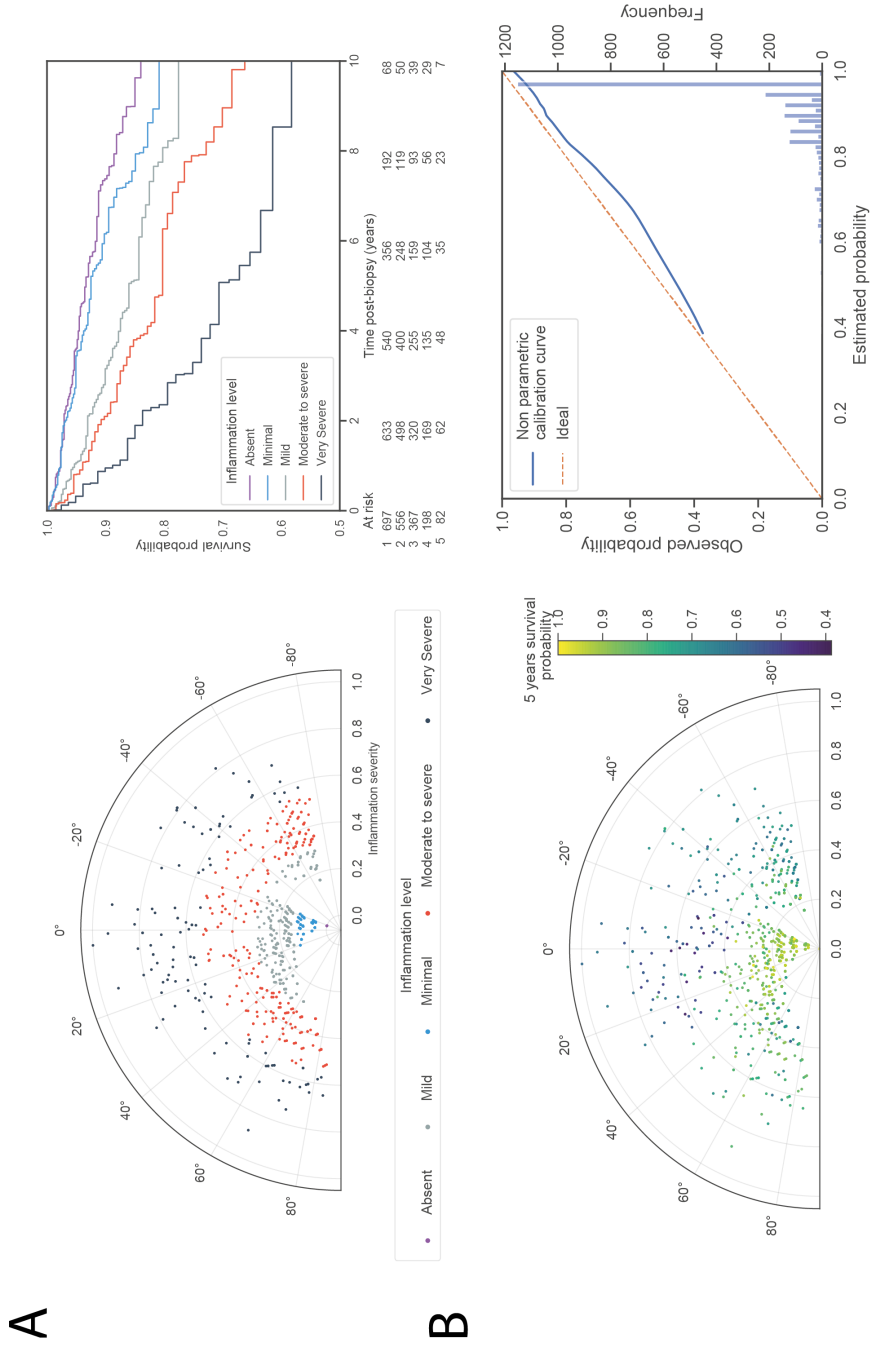


Figure 4.6: Association with graft survival in the polar plot visualization tool.

Figure 4.6: **A.** Association of the polar plot radius with graft survival in the derivation cohort. We stratified the 3510 biopsies along the radius axis in five strata and plotted the corresponding Kaplan-Meier survival curves. This demonstrates that the radius of the polar plot, which represents the extent of inflammation (the sum of the re-weighted acute lesions scores, scaled to the unit interval from 0 to 1), is positively associated with the risk of graft failure. The different levels of inflammation correspond to the following radius : *" No inflammation"* : radius 0.00 – 0.04; *"Minimal inflammation"*: radius 0.04 – 0.10 ; *" Mild inflammation"* : radius 0.10 – 0.24; *"Moderate to severe inflammation"* : radius 0.24-42; and *" "Very severe inflammation"*: radius 0.42 – 1.00 **B.** Estimated graft survival probability at 5 years post-biopsy, calculated from the nearest neighborhood with k=40 (left panel) with corresponding calibration curve (right panel).

4.3.5 Association between disease clusters and graft failure

During follow-up, 125 grafts failed, at a median of 3.67 years (1 day to 12 years) after transplantation. 9.1%, 22.4%, 25.0%, 30.0%, 37.7% and 50.0% of grafts failed within the first 5 years after the biopsy in respectively cluster 1 to 6. The disease clusters 2 to 6 all associated with an increased risk of graft failure in comparison the cluster 1 (Figrue 4.2 and Table 4.3). Although Banff rejection categories had significant association with graft failure, except for Borderline changes, the clusters weighted average in DRMST at 5 and 10 years were higher than the weighted average DRMST from the Banff classification (respectively 0.46 and 1.25 years for the clusters vs. 0.29 and 0.72 years for the Banff categories), illustrating an overall better discrimination in terms of graft failure (Table 4.3)). Furthermore, we observed an asymmetry between the first three and last three clusters, based on HLA-DSA status. Hazard ratios on the HLA-DSA negative/HLA-DSA positive pair of clusters reported the following values: cluster 1 vs cluster 4 : 2.84 (CI 95% 1.80-4.30; p<0.0001), cluster 2 vs. cluster 5 : 2.02 (1.00-4.12 ; p= 0.051), cluster 3 vs. cluster 6: 2.41 (CI 95% 1.35-4.30 ; p= 0.003) (Figure 4.A.5). The survival outcome of each cluster did not depend on the adjustment method for repeated biopsies per patient (Figure 4.A.6). The clustering of biopsies led to improved prediction of graft failure, compared to the Banff classification (Supplemental Results 4.A).

The radius on the polar plot of each biopsy associated independently with graft failure, with an AUROC for 2- and 5-year post-biopsy graft survival of respectively 0.70 (95% CI 0.66-0.73) and 0.69 (95% CI 0.67-0.72), respectively. Biopsies projected on the outer ranges of the radius had higher inflammatory lesion scores, and significantly worse survival compared to biopsies near the center of the polar plot (Figure 4.6). Similar association with graft failure was obtained when we predicted graft failure for each biopsy separately, from the information available on the nearest neighborhood, calculated using the weighted Euclidean distance. For example, Figure 4.6 displays the survival probability at 5 years post-biopsy, estimated from local Kaplan-Meier estimates based on 40 nearest neighbors. With this local approach, solely based on the lesion scores and HLA-DSA status and not taking into account graft functional data or post-transplant time, the AUROC of the probability for graft failure were 0.72 (95% CI 0.68-0.74) and 0.70 (95% CI 0.67-0.73), respectively at 2- and 5-years post-biopsy.

Table 4.3: Graft survival, restricted mean survival time (RMST) and difference in RMST (DMRST) at 5- and 10-years post-biopsy, according to each cluster and each Banff diagnostic category (N=3510).

Banff diagnosis	**% Graft survival at 5 years post-biopsy**	**% Graft survival at 10 years post-biopsy**	**RMST at 5 years post-biopsy (95% CI)**	**RMST at 10 years post-biopsy (95% CI)**	**HR vs. No rejection (95% CI)**	**HR p-value vs. No rejection**	**DRMST at 5 years vs. No rejection (95% CI)**	**DRMST at 10 years vs. No rejection (95% CI)**
No rejection	89.50%	51.00%	4.74 (4.63-4.85)	9.01 (8.56-9.46)	-	-	-	-
Borderline changes	83.10%	42.30%	4.66 (4.47-4.85)	8.87 (8.04-9.7)	1.27 (0.88-1.84)	0.201	0.08 (-0.06-0.22)	0.14 (-0.24-0.53)
TCMR	75.50%	41.20%	4.50 (4.27-4.74)	8.46 (7.7-9.22)	1.66 (1.17-2.36)	0.004	0.24 (0.05-0.42)	0.55 (0.09-1.01)
ABMR	70.20%	22.00%	4.26 (3.89-4.63)	7.63 (6.25-9.02)	2.63 (1.65-4.21)	<0.0001	0.48 (0.15-0.81)	1.38 (0.51-2.25)
Mixed borderline rejection	63.60%	8.30%	3.55 (4.06-4.57)	6.74 (4.06-7.72)	4.26 (2.29-7.94)	<0.0001	0.67 (0.06-1.28)	2.13 (0.65-3.61)
Mixed rejection	59.20%	20.50%	3.98 (3.48-4.48)	7.03 (5.74-8.32)	3.24 (2.08-5.05)	<0.0001	0.76 (0.34-1.18)	1.98 (1.00-2.96)
Average	-	-	-	-		-	0.45 (0.11-0.78)	1.24 (0.40-2.07)
Weighted average	-	-	-	-		-	0.29 (0.06-0.53)	0.72 (0.14-0.93)

Continued on the next page

Cluster	% Graft survival at 5 years post-biopsy	% Graft survival at 10 years post-biopsy	RMST at 5 years post-biopsy (95% CI)	RMST at 10 years post-biopsy (95% CI)	HR vs. cluster 1 (95% CI)	HR vs. cluster 1 p-value	DRMST at 5 years vs. cluster 1	DRMST at 10 years vs. cluster 1
Cluster 1	90.90%	54.60%	4.76 (4.65-4.87)	9.09 (8.63-9.56)	-	-	-	-
Cluster 2	77.60%	33.30%	4.42 (4.01-4.84)	8.28 (6.98-9.59)	1.98 (1.15-3.43)	0.014	0.34 (0.02-0.70)	0.81 (-0.03-1.65)
Cluster 3	75.00%	39.80%	4.54 (4.30-4.79)	8.56 (7.75-9.36)	1.72 (1.17-2.52)	0.005	0.22 (0.02-0.41)	0.53 (0.05-1.02)
Cluster 4	70.00%	28.60%	4.23 (3.87-4.58)	7.67 (6.36-8.98)	2.84 (1.88-4.30)	<0.0001	0.53 (0.24-0.82)	1.42 (0.69-2.15)
Cluster 5	62.30%	6.10%	4.06 (3.48-4.64)	6.84 (5.04-8.64)	4.17 (2.48-7.03)	<0.0001	0.70 (0.26-1.14)	2.25 (1.07-3.42)
Cluster 6	50.00%	6.20%	3.96 (3.43-4.48)	6.78 (4.96-8.59)	4.37 (2.59-7.35)	<0.0001	0.80 (0.31-1.29)	2.31 (1.11-3.52)
Average	-	-	-	-		-	0.52 (0.17-0.87)	1.46 (0.58-2.35)
Weighted average	-	-	-	-		-	0.46 (0.09-0.69)	1.25 (0.48-2.01)

4.3.6 External validation

Using the features weights and the cluster centroids obtained from the consensus clustering process, we are able to classify any new biopsy into one of the 6 previously described clusters. We applied this algorithm, starting from the lesion scores and HLA-DSA status only, without information on graft survival, to an external dataset of 3835 biopsies from Lyon University Hospital (N=1356) and the Paris Transplant Group (N=2479)(Table 4.A.4)). Note that this dataset did not include thrombi in its variables. We therefore imputed this feature from the mean value of our training data. A comparison of the final clusters proportions between the two centers is presented in Figure 4.A.7. Similar to the training set, biopsies from the external validation set were largely dominated by non-inflamed cluster 1. The main difference in cluster distribution was a higher proportion of cluster 4 biopsies in the external dataset compared to the Leuven dataset (26.0% vs 8.7%, $p<0.0001$), explained by a larger prevalence of HLA-DSA positive biopsies in the external data. Logically, the proportion of lesions within each clusters of the external validation set were very similar to the clusters obtained from the original data. There was also a similar association of the clusters with graft failure (Figure 4.A.9). A polar plot illustrates the full overlap in the histological presentations between the training and validation cohorts (Figure 4.A.8). Although the proportion of biopsies performed upon indication was notably higher in the validation cohort (22.0% vs 37.7%, χ^2 test $p<0.0001$), the overall distribution of inflammation, estimated using the radius on the polar plot, was comparable between the training and validation datasets (Figure 4.A.10). Comparing the clusters obtained on the validation dataset with the Banff categories, we obtained an ARI of 0.35. While maintaining a large overlap between the clustering method and the Banff classification (Table 4.A.5)), it demonstrates a higher reclassification rate in the validation dataset.

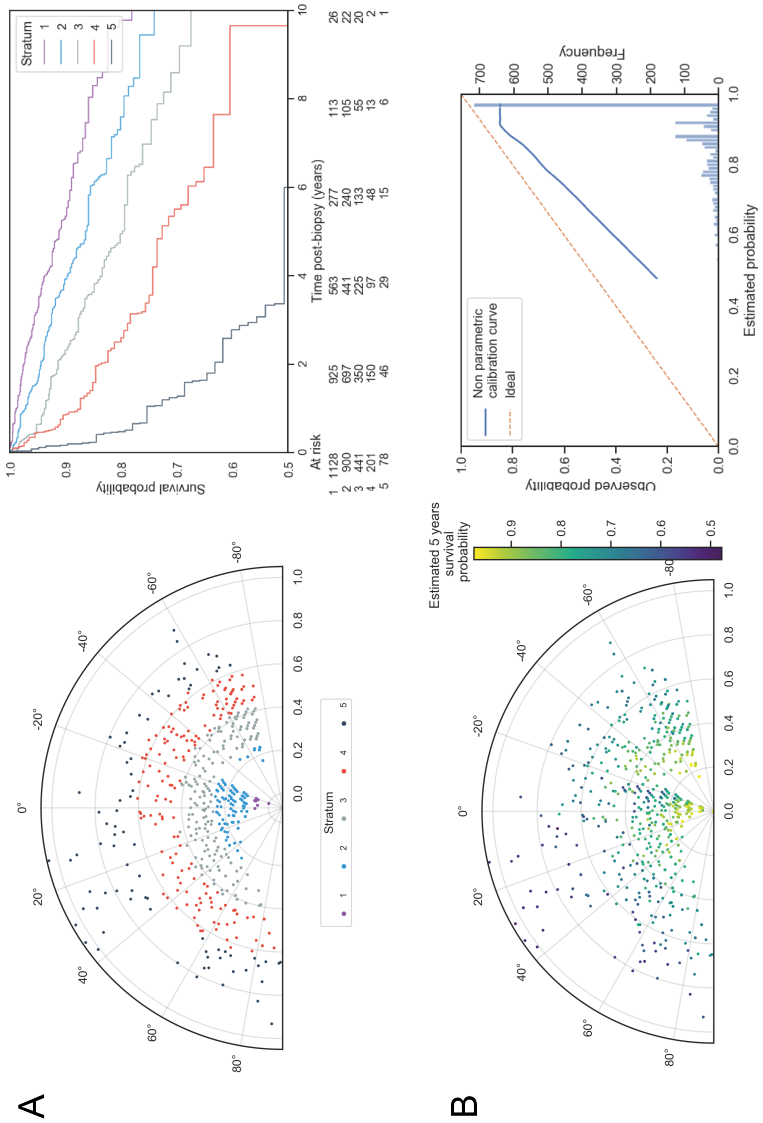


Figure 4.7: Association with graft survival in the polar plot visualization tool based on the validation data **A.** Association of the polar plot radius with graft survival in the external validation cohort. We stratified the 3835 biopsies along the radius axis in five subgroups and plotted the corresponding Kaplan-Meier survival curves. **B.** Estimated survival probability at 5 years post-biopsy, calculated from the nearest neighborhood with k=40.

4.4 Discussion

Using a semi-supervised and data-driven approach on 7345 post-transplant kidney biopsies with re-weighting of acute histological lesions, we derived and validated six distinct, clinically meaningful, phenotypic clusters. This mathematical clustering approach was fundamentally different from the iterative Banff classification process, which relies on a set of clinically derived *if-then* rules. Nevertheless, both in the training and the validation cohort, the novel phenotypes for kidney transplant rejection had a good degree of similarity with the Banff rejection categories, while redistributing intermediate and mixed phenotypes and maintaining the association with graft failure. The novel rejection phenotypes led to improved prediction of graft failure compared to the Banff classification. For integration of the novel phenotypic clustering with disease severity, and to move away from the black-white disease categorization, we proposed and validated a method for easily interpretable two-dimensional visual and quantitative presentation of the multidimensional histologic data.

Despite the similarity between the novel clusters and the Banff categories, we showed statistically improved prediction of graft failure with the clustering approach than when using the Banff categories, especially in ambiguous situations, like Borderline changes or mixed rejection phenotypes. The association between (non)-inflamed clusters and graft survival remained present even when the biopsies were stratified according to the rejection or non-rejection categories defined by Banff. An example of the clinical impact of this is e.g. the lack of cluster reflecting Banff Borderline changes. Borderline changes are not reflected in a separate cluster, but most often (79.8%) classified to non-inflamed cluster 1, with best post-transplant graft survival. Using this clustering approach therefore may solve the clinically difficult issue of how to deal with minimal tubulo-interstitial inflammation, below the current thresholds for TCMR. Also, the clustering algorithm proposed a novel phenotype, which is driven by glomerulitis in the absence of HLA-DSA. Although the causes of this phenotype are currently unknown, this resembles the phenotype described in recent publications on HLA-DSA negative microcirculation inflammation [176, 183–185]. This phenotype is currently not recognized in the Banff classification [172] and should be worked out in greater detail with respect to pathophysiology, risk factors and clinical presentation. Finally, cluster 6 represents cases with mixed rejection phenotypes (ABMR with TCMR or Borderline changes), which is not recognized as separate category in the Banff classification, but representing a clinical dilemma [186].

Our clustering method directly relies on distance computation and provides a clinically relevant similarity metric to compare biopsies without concomitant clinical data besides HLA-DSA. For instance, we demonstrated and validated

that local survival prediction based on the nearest biopsies exhibited prognostic value solely based on the histological lesions and HLA-DSA status, thus not taking into account graft functional parameters or demographic factors relevant for outcome [22]. Relying on this ad-hoc distance metric and to move beyond the black-white clustering approach, we developed an intuitive two-dimensional visualization tool, enabling to plot newly performed post-transplant biopsies and rapidly assess the disease severity along with the dominant phenotype of neighboring biopsies. Because the k-mean algorithm is a hard-clustering algorithm, biopsies near the clusters' boundaries get strictly allocated to one of the two neighbor clusters, preventing an overlap of diagnoses. This explains for instance that the mixed rejection biopsies are now split into one of the major clusters based on their dominant lesions. However, contrasting with the Banff categorization, our clustering system can provide some degree of certainty regarding the classification, as expressed in term of the relative distance to the closest cluster prototype (centroid). As a time-independent approach, our method is intended solely for reclassification of rejection (clustering algorithm and theta angle on the polar plot) and assessment of disease severity (radius on the polar plot). Our analysis on the accuracy of the local survival prediction needs to be seen as support for the clinical validity of the location of each sample on the polar plot and does not suggest clinical utility of the local survival prediction as a prognostic tool on its own. For prognostication, more granular tools are becoming available, such as the *iBox* prediction score [22], which also integrate time post-transplantation and graft functional parameters into the models. Finally, diagnosis of other relevant disease phenotypes like glomerulonephritis or polyomavirus nephropathy are based on other parameters that are currently not included in the algorithm. These diseases should not be evaluated with our system solely intended for reclassification of rejection phenotypes.

As chronic histological lesions in kidney transplant biopsies are non-specific [170], we focused solely on acute inflammatory lesions to derive the novel rejection phenotypes. The evolution from active/early stage disease, to chronic active, and finally chronic inactive forms of the same disease, was therefore not assessed and can be considered for future developments reclassification system. In addition, our approach fully depends on the quality of the histological assessment, which is pathologist-dependent and therefore not fully reproducible [187, 188]. More objective data, such as computerized imaging data or molecular expression data, or information on e.g. non-HLA antibodies and other immune risk factors, could further improve the reproducibility and accuracy of our system. Next, although the clusters described are sound biologically/clinically, whether treatment decisions based on clusters instead of based on Banff diagnosis will yield better outcome cannot be tested in this retrospective study. Similarly, data-driven algorithms do not assess pathophysiological mechanisms, hence

no causal relations can be deducted from any cluster. Besides these clinical aspects, also some technical limitations warrant discussion. In concordance to the method described earlier [178], we used the whole dataset to compute the lesion score weights. The more data available to compute the weights, the more precise their estimation will be. In this semi-supervised setting, weights overfitting is less detrimental than in a purely supervised approach. Despites its good performance, the k-mean algorithm remains simplistic. More elaborated core clustering algorithms, such as model-based or fuzzy clustering methods, could benefit the current approach and warrants additional studies.

Although we described a meaningful data-driven alternative to classify kidney transplant biopsies, and although our system has benefits over the current Banff categories, we do not suggest replacing the existing Banff classification with this algorithm but use it in addition to Banff categorization, especially in cases that are difficult to categorize according to Banff. The clinical or scientific utility of our approach needs to be shown in further studies that validate the improved clinical decision-making with regards to rejection treatment. Clinical implementation will depend on further external validation and detailed discussion at future Banff meetings and international consensus. The underlying risk factors and clinical presentations of each of the clusters still needs to be evaluated in greater depth, including information on HLA-DSA subtypes and profiles, non-HLA antibodies etc. Inference on treatment decisions could not be made on our cohort, given the fact that Banff-rejection cases were treated with high-dose corticosteroids, and that cases of ABMR were treated with antibody-targeted therapies only very rarely [176]. Nevertheless, the current study highlights the potential of using the full scale of lesion grades for classification of kidney transplant biopsies, rather than using discretionary cut-off values. In the era of increasing availability of morphometric [190] or molecular [168] data, advanced statistical analysis and machine learning, with many resources to handle high-dimensional continuous variables [191], the existing expert-based consensus of *if-then* rules could be further improved using our approaches.

4.5 Conclusion

We have developed and validated a semi-supervised clustering approach for the identification of clinically meaningful novel phenotypes for kidney transplant rejection, based on individual lesion scores. This approach has potential to offer a more quantitative evaluation of rejection subtypes and severity, especially in situations where the current histological categorization is ambiguous.

Chronic lesions

While this chapter describes the development and the validation of 6 novel phenotypes based on *acute* Banff lesion scores, we recently performed a similar research on the *chronic* lesions. We trained and validated a semi-supervised consensus clustering algorithm to cluster post-transplant kidney biopsies based on their chronic Banff lesions scores. Due to the overlap in the methodology and the redundancy of the analysis, we did not report the results of this additional research not to overload the current manuscript.

This new data-driven algorithm, trained on 3549 biopsies, produced 4 clusters of different degrees of chronicity. By demonstrating the discriminative power of the chronic lesions on graft failure, independently of inflammation, these recent results confirmed the clinical interest of exploiting a more extensive assessment of the chronicity of kidney transplant biopsies, which remains underappreciated in the current Banff classification. Although chronic lesions are associated with time, the clustering of the biopsies solely based on the time post-transplant variable produced clusters without discrimination in graft survival association. We further demonstrated that the pattern of the chronic lesions was more important than the timing of the biopsy in predicting graft survival. Additionally, in comparison to the existing IFTA grades[a] which relate to the chronic injuries of kidney transplants, the obtained chronic clusters were more discriminant towards graft failure. Similarly to Section 4.2.2, we developed a polar plot approach that provides a more granular visualization through a continuous assessment of the overall severity of the chronic injury. Finally, these results were validated on an extensive cohort of 4031 external biopsies.

This study will be submitted as Vaulet et al. *Data-driven chronic rejection phenotypes: a novel and validated complement for the complete histological assessment of kidney transplant biopsies* to the Journal of the American Society of Nephrology [189].

[a] Interstitial fibrosis and tubular atrophy

Appendices

4.A Supplementary results

4.A.1 Comparison of clustering vs. Banff classification for association with graft failure

The biopsies that were diagnosed as No rejection or Borderline changes according to Banff but reclassified to inflamed clusters 2 and 3 (N=92), had worse outcome compared to patients reclassified to non-inflamed cluster 1 (HR 1.98, CI 95% 1.18-3.34; p=0.01). Low numbers of reclassified cases with No rejection according to Banff in the HLA-DSA positive patients obviated to perform this analysis in this group. No rejection biopsies according to Banff had significantly worse graft survival when clustered to cluster 2, 3, 5, or 6 (taken together N=100) compared to non-inflamed cluster 1 or 4 (N=2886)(HR 1.99, 95% CI 1.03-3.84; p=0.04, adjusted for the presence of HLA-DSA). Restricted to the DSA-negative clusters, No rejection biopsies according to Banff had a significantly worse survival when classified in cluster 2 or 3 (n=57) vs. cluster 1 (N=2387) (HR: 1.96, 95% CI 1.01- 3.77; p= 0.04). An analogous trend was observed in the less numerous cases of Banff-defined Borderline changes, when classified in cluster 1 (n=261) vs. in clusters 2-3 (n=35) (HR = 2.052, CI 95% 0.93-4.55; p= 0.077). Similarly, rejection biopsies according to the Banff classification (ABMR, TCMR or mixed phenotypes; N=524) had significantly better survival when clustered to non-inflamed cluster 1 or 4 (N=131) compared to cluster 2, 3, 5, or 6 (N=393) (HR 0.583, 95% CI 0.385-0.881); p=0.01, adjusted for the presence of HLA-DSA). The Banff classification did not contribute to graft outcome prediction within the disease clusters. Non-inflamed biopsies (cluster 1 or 4) did not have different outcome when classified as either No rejection (N=2602) or rejection (N=135) according to the Banff classification (HR 1.11, 95% CI 0.58-2.13; p=0.75). Likewise, biopsies from clusters 2,3,5 or 6 did not have different outcome when classified as No rejection (N=57) vs. rejection (N=393) according to the Banff

classification (HR 0.83, 95% CI 0.41-1.69; p=0.60).

4.A.2 External validation

Using the features weights and the cluster centroids obtained from the consensus clustering process, we are able to classify any new biopsy into one of the 6 previously described clusters. We applied this algorithm, starting from the lesion scores and HLA-DSA status only, without information on graft survival, to an external dataset of 3835 biopsies from Lyon University Hospital (N=1356) and the Paris Transplant Group (N=2479)(Table 4.A.4). Note that this dataset did not include thrombi in its variables. We therefore imputed this feature from the mean value of our training data. A comparison of the final clusters proportions between the two centers is presented in Figure 4.A.7. Similar to the training set, biopsies from the external validation set were largely dominated by non-inflamed cluster 1. The main difference in cluster distribution was a higher proportion of cluster 4 biopsies in the external dataset compared to the Leuven dataset (26.0% vs 8.7%, p<0.0001), explained by a larger prevalence of HLA-DSA positive biopsies in the external data. Logically, the proportion of lesions within each clusters of the external validation set were very similar to the clusters obtained from the original data. There was also a similar association of the clusters with graft failure (Figure 4.A.9).

A polar plot illustrates the full overlap in the histological presentations between the training and validation cohorts (Figure 4.A.8). Although the proportion of biopsies performed upon indication was notably higher in the validation cohort (22.0% vs 37.7%, $\chi 2$ test p<0.0001), the overall distribution of inflammation, estimated using the radius on the polar plot, was comparable between the training and validation datasets (Figure 4.A.10). Comparing the clusters obtained on the validation dataset with the Banff categories, we obtained an ARI of 0.35. While maintaining a large overlap between the clustering method and the Banff classification (Table 4.A.5), it demonstrates a higher reclassification rate in the validation dataset. We observed that biopsies that were diagnosed as No rejection or Borderline changes according to Banff but reclassified to inflamed clusters 2 and 3 (N=165), had worse outcome compared to patients reclassified to non-inflamed cluster 1 (HR 2.55, 95% CI 1.76-3.67; p<0.001). Low numbers of reclassified cases with No rejection according to Banff in the HLA-DSA-positive patients obviated to perform this analysis in this group. No rejection biopsies according to Banff had significantly worse graft survival when clustered to cluster 2, 3, 5, or 6 (taken together N=57) compared to non-inflamed cluster 1 or 4 (taken together N=2826)(HR 4.84, 95% CI 2.90-8.07; p=0<0.0001), adjusted for the presence of HLA-DSA). Rejection biopsies according to the Banff classification (ABMR, TCMR or mixed phenotypes;

Table 4.A.1: Performance indices for a range of k values used to define the optimal number of clusters in the semi-supervised clustering algorithm (N=3510 biopsies of the derivation cohort).

Performance index	$\boldsymbol{k}$ =4	$\boldsymbol{k}$ =5	$\boldsymbol{k}$ =6	$\boldsymbol{k}$ =7	$\boldsymbol{k}$ =8
Unsupervised clustering (unweighted lesion scores)					
Proportion of Ambiguous Clustering (PAC)	0.115	0.168	0.154	0.141	0.129
Log-rank p-value between best and 2nd best survival cluster	0.034	0.102	0.155	0.245	0.98
Minimal distance between centroids*	0.942	0.941	0.981	0.954	0.837
Adjusted Rand Index (ARI)	0.49	0.478	0.486	0.495	0.473
Semi-supervised clustering (weighted lesion scores)					
Proportion of Ambiguous Clustering (PAC)	0.112	0.152	0.083	0.133	0.144
Log-rank p-value between best and 2nd best survival cluster	0.005	0.025	0.012	0.97	0.996
Minimal distance between centroids*	0.13	0.134	0.126	0.126	0.12
Adjusted Rand Index (ARI)	0.475	0.49	0.48	0.469	0.455

*Distances are specific to the space used and should not be compared between unsupervised and semi-supervised methods.

N=704), demonstrated better survival when clustered to non-inflamed cluster 1 or 4 (N=227) compared to cluster 2, 3, 5, or 6 (N=477) (HR 1.59, 95% CI 1.08-2.35; p=0.018, adjusted for the presence of HLA-DSA).

Table 4.A.2: Distribution of biopsies among clusters and association with clinical demographics, transplant and biopsy characteristics (N=3510 biopsies of the derivation cohort).

	Cluster					
	1	2	3	4	5	6
N (%)	2710 (77.2%)	101 (2.9%)	231 (6.6%)	307 (8.7%)	95 (2.7%)	66 (1.9%)
Donor demographics*						
Donor Type						
Donation after brain death, N (%)	626/811 (77.2%)	50/65 (76.9%)	141/179 (78.8%)	92/115 (80.0%)	44/50 (88.0%)	35/45 (77.8%)
Donation after cardiac death, N (%)	137/811 (16.9%)	12/65 (18.5%)	28/179 (15.6%)	15/115 (13.0%)	3/50 (6.00%)	8/45 (17.8%)
Living donation, N (%)	48/811 (5.9%)	3/65 (4.6%)	10/179 (5.6%)	8/115 (7.0%)	3/50 (6.00%)	2/45 (4.40%)
Age (years) Mean (SD)	47.6 (±14.5)	52.0 (±13.8)	48.5 (±14.1)	47.0 (±17.2)	47.3 (±16.9)	46.4 (±18.0)
Sex male n/N(%)	434/811 (53.5%)	33/65 (50.8%)	84/179 (46.9%)	57/115 (49.6%)	28/50 (56.0%)	22/45 (48.9%)
Diabetes n/N(%)	21/639 (3.3%)	1/55 (1.8%)	4/137 (2.9%)	2/83 (2.4%)	0/34 (0.0%)	2/31 (6.5%)
Recipient demographics*						
Rec age (years) Mean (SD)	53.4 (±13.2)	56 (±11.9)	52.2 (±14.8)	52.4 (±14.4)	53.4 (±13.6)	48 (±17.1)
Sex male n/N(%)	515/811 (63.5%)	42/65 (64.6%)	112/179 (63.1%)	52/115 (45.2%)	22/50 (44.0%)	24/45 (53.3%)
BMI rec. -Mean (SD)	25.3 (±4.39)	26.8 (±3.89)	25.3 (±4.34)	25.3 (±5.04)	25.6 (±5.16)	26.1 (±5.90)
Pretransplant-DSA n/N(%)	0/811 (0.0%)	0/65 (0.0%)	0/179 (0.0%)	90/115 (78.3%)	46/50 (92.0%)	33/45 (73.3%)
Graft *						
Repeat TX n/N(%)	81/811 (10.0%)	6/65 (9.2%)	14/179 (8.4%)	49/115 (42.6%)	26/50 (52.0%)	20/45 (44.4%)
Total Mismatch - Mean (SD)	2.72 (±1.33)	3.26 (±1.21)	2.83 (±1.42)	2.9 (±1.25)	2.68(±1.11)	2.91 (±1.12)
CIT (hours) -Mean (SD)	14.2 (±5.7)	14.9 (±6.1)	14.1 (±5.3)	14.3 (±5.8)	14.5 (±4.7)	14.3 (±4.6)
Biopsy *						
Protocol biopsy n (%)	2269 (83.7%)	57 (56.4%)	115 (49.8%)	218 (71.0%)	53 (55.8%)	25 (37.9%)
Indication biopsy n (%)	441 (16.3%)	44 (43.6%)	116 (50.2%)	89 (29.0%)	42 (44.2%)	41 (62.1%)
Proteinuria - Mean (SD)	0.26 (±0.54)	0.47 (±0.62)	0.41 (±0.65)	0.29 (±0.60)	0.44 (±0.96)	0.68 (±0.94)
eGFR - Median (IQR)	44.0 (32.2-55.6)	32.09 (17.1-50.8)	31.4 (19.86-44.8)	37.3 (24.5-52.9)	31.8 (19.2-48.0)	20.5 (14.7-36.3)
Days since transplantation - Median (IQR)	366 (92-744)	111 (14-733)	104 (13-729)	363 (89-740)	103 (17-666)	368 (17-790)

*when two or more biopsies from the same patient are classified in the same cluster, only the first biopsy is taken into account for graft-related statistics.

Table 4.A.3: Details of cluster composition according to the individual Banff lesions scores and donor-specific HLA antibodies (N=3510 biopsies of the derivation cohort)

Banff lesion	Lesion score	Clusters					
		1	**2**	**3**	**4**	**5**	**6**
		N=2710	N=307	N=231	N=101	N=95	N=66
Tubulitis (t)	0	72.90%	47.50%	3.50%	73.00%	46.30%	4.50%
	1	21.40%	38.60%	17.70%	23.10%	46.30%	30.30%
	2	5.00%	12.90%	48.90%	3.60%	6.30%	42.40%
	3	0.70%	1.00%	29.90%	0.30%	1.10%	22.70%
Interstitial inflammation (i)	0	91.50%	59.40%	0.00%	90.90%	61.10%	0.00%
	1	8.30%	17.80%	2.20%	9.10%	26.30%	3.00%
	2	0.20%	8.90%	49.40%	0.00%	8.40%	40.90%
	3	0.00%	13.90%	48.50%	0.00%	4.20%	56.10%
Intimal arteritis (v)	0	96.50%	71.30%	74.00%	92.20%	67.40%	66.70%
	1	3.20%	22.80%	21.20%	7.80%	30.50%	28.80%
	2	0.10%	5.00%	3.90%	0.00%	2.10%	3.00%
	3	0.10%	1.00%	0.90%	0.00%	0.00%	1.50%
Glomerulitis (g)	0	93.40%	0.00%	81.80%	78.80%	0.00%	56.10%
	1	6.60%	0.00%	13.00%	16.30%	2.10%	28.80%
	2	0.00%	56.40%	5.20%	4.90%	27.40%	13.60%
	3	0.00%	43.60%	0.00%	0.00%	70.50%	1.50%
Peritubular capillaritis (ptc)	0	93.90%	61.40%	55.80%	79.80%	28.40%	39.40%
	1	4.20%	20.80%	24.20%	11.40%	30.50%	27.30%
	2	1.70%	16.80%	17.70%	8.50%	38.90%	28.80%
	3	0.10%	1.00%	2.20%	0.30%	2.10%	4.50%
Thrombi	0	97.80%	94.10%	97.00%	96.40%	96.80%	95.50%
	1	2.20%	5.90%	3.00%	3.60%	3.20%	4.50%
C4d deposition in peritubular capillaries	0	89.00%	71.30%	78.40%	72.30%	51.60%	59.10%
	1	9.50%	14.90%	12.60%	13.40%	9.50%	10.60%
	2	0.50%	2.00%	1.30%	3.30%	4.20%	4.50%
	3	1.00%	11.90%	7.80%	11.10%	34.70%	25.80%
Donor-specific antibodies (DSA)	0	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%
	1	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%

Table 4.A.4: Demographic, clinical and histological characteristics of the patients and biopsies included in the validation dataset.

Cohort characteristics	**Total (N=1989)**
Donor demographics	
Donor type	
Donation after brain death, N (%)	1469 (73.8)
Donation after cardiac death, N (%)h	101 (21.1)
Living donation, N (%)	
Age (years), mean ± SD	53.3 (16.8)
Male, N (%)	1048 (52.7)
Recipient demographics	
Age (years), mean ± SD	51.0 (14.5)
Male, N (%)	1241 (62.4)
BMI (kg/m^2), mean (range)	24.4 (4.3)
Repeat transplantation, N (%)	271 (13.6)
Cold ischemia time (hours), mean ± SD	14.7 (10.5)
Total number of HLA A/B/DR mismatches, mean ± SD	3.4 (1.4)
Biopsy characteristics	**Total (N=3835)**
Banff 2019 diagnosis	
No rejection, N (%)	2883 (75.2)
Borderline changes, N (%)	248 (6.5)
TCMR, N (%)	194 (5.1)
ABMR, N (%)	416 (10.8)
Mixed rejection (ABMR + TCMR), N (%)	33 (0.9)
Mixed borderline rejection (ABMR + borderline changes), N (%)	61 (1.6 %)
Indication biopsies, N (%)	N=1442 (37.7)
Days since transplantation, median (IQR)	190 (45-655)
eGFR at day of biopsy, mean (std)	36.5 (18.4)
Protocol biopsies, N (%)	N=2393 (62.3)
Days since transplant, median (IQR)	180 (94-375)
eGFR at day of biopsy –mean (std)	54.7 (22.4)

Table 4.A.5: Contingency tables comparing the Banff 2019 diagnosis and the 6 clusters obtained on the external validation dataset.
Proportions represent the distribution in the clusters per Banff category.

		Clusters					
	N	1	2	3	4	5	6
No rejection	2883	2003 (69.5%)	54 (1.9%)	3 (0.1%)	823 (28.5%)	0 (0.0%)	0 (0.0%)
Borderline	248	106 (42.7%)	24 (9.7%)	84 (33.9%)	16 (6.5%)	0 (0.0%)	18 (7.3%)
TCMR	194	7 (3.6%)	10 (5.2%)	158 (81.4%)	2 (1.0%)	0 (0.0%)	17 (8.8%)
ABMR	416	63 (15.1%)	40 (9.6%)	0 (0.0%)	152 (36.5%)	158 (38.0%)	3 (0.7%)
Mixed borderline rejection	33	0 (0.0%)	3 (9.1%)	10 (30.3%)	2 (6.1%)	9 (27.3%)	9 (27.3%)
Mixed rejection	61	0 (0.0%)	8 (13.1%)	20 (32.8%)	1 (1.6%)	6 (9.8%)	26 (42.6%
Total	**3835**	**2179 (56.8%)**	**139 (3.6%)**	**275 (7.2%)**	**996 (26.0%)**	**173 (4.5%)**	**73 (1.9%)**

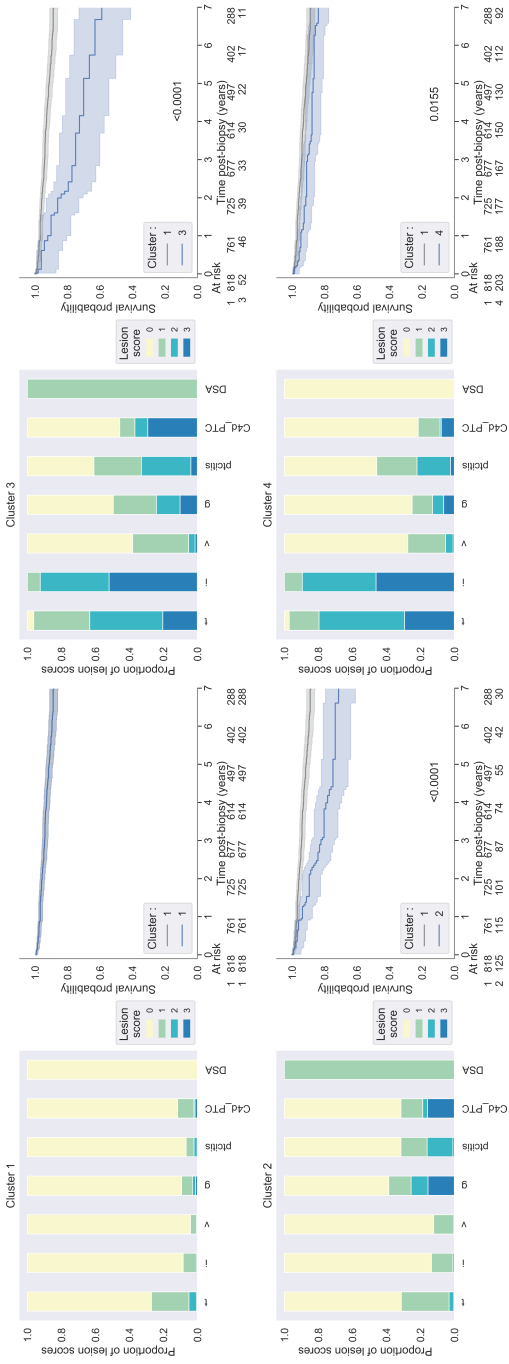


Figure 4.A.1: Distribution of the individual acute lesion scores in the clusters using an unweighted approach, and post-biopsy Kaplan-Meier graft survival curves relative to cluster 1 of the derivation cohort (N=3510 biopsies). Compared to cluster 1 (essentially normal biopsies), the three other clusters associated significantly with impaired graft survival. However, their histological and clinical relevance was more ambiguous. P-values refer to HR from the Cox models.

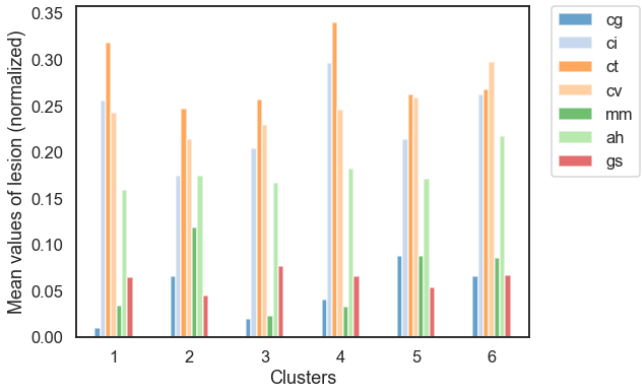


Figure 4.A.2: Distribution of chronic lesions in the 6 acute lesion clusters (N=3510 biopsies).
Chronic lesions shared similar profiles across the 6 clusters based on acute lesion score, translating presumably independent mechanisms. cg=transplant glomerulopathy, ci=interstitial fibrosis, ct=tubular atrophy, cv=vascular intimal thickening, mm=mesangial matrix increase, ah=arteriolar hyalinosis, gs=glomerulosclerosis.

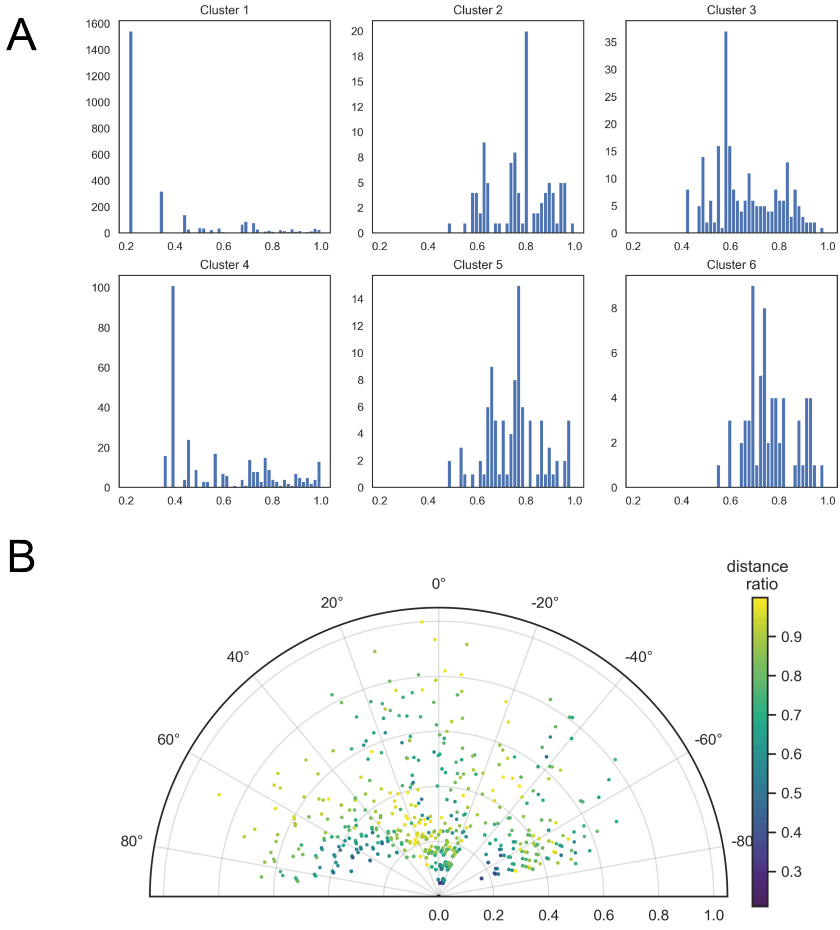


Figure 4.A.3: Relative distances to the closest clusters' boundary.
A. Distribution of the relative distance from the centroids of the biopsies included in each cluster. **B.** Visualization of the relative distances to the cluster boundaries, plotted on the polar plot. A relative distance close to one indicates an almost equidistant position from the two closest cluster centroids. Conversely, a biopsy that is much closer to a centroid than it is from the second closest centroid has a lower relative distance.

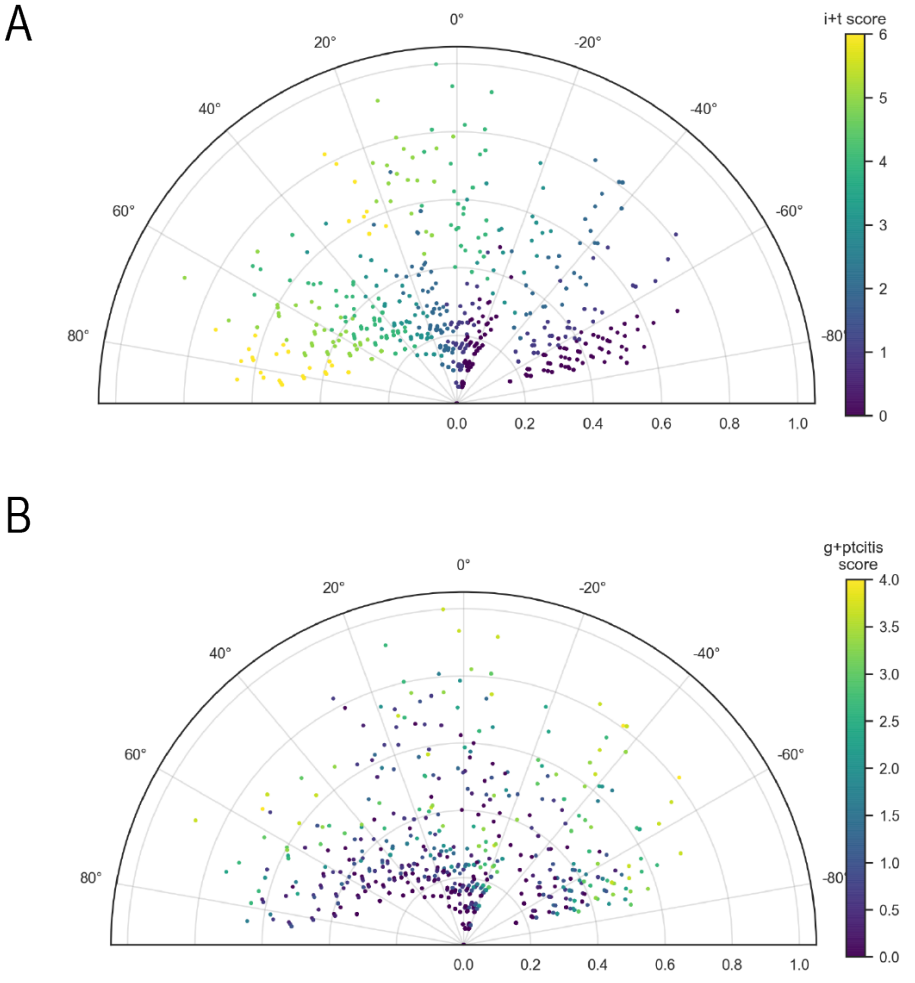


Figure 4.A.4: Various combinations of lesions scores displayed on the polar plots.
The sum of lesions scores is represented by the color of the points. **A.** Biopsies with a high combined score of t and i are predominant on the extreme left (*"Pure TCMR"*) and on middle of the polar plot (mixed rejection). **B.** Biopsies with a high combined score of g and ptc are predominant on the right side of the plot (*"pure ABMR"*) and on the middle of the plot (*"mixed rejection"*).

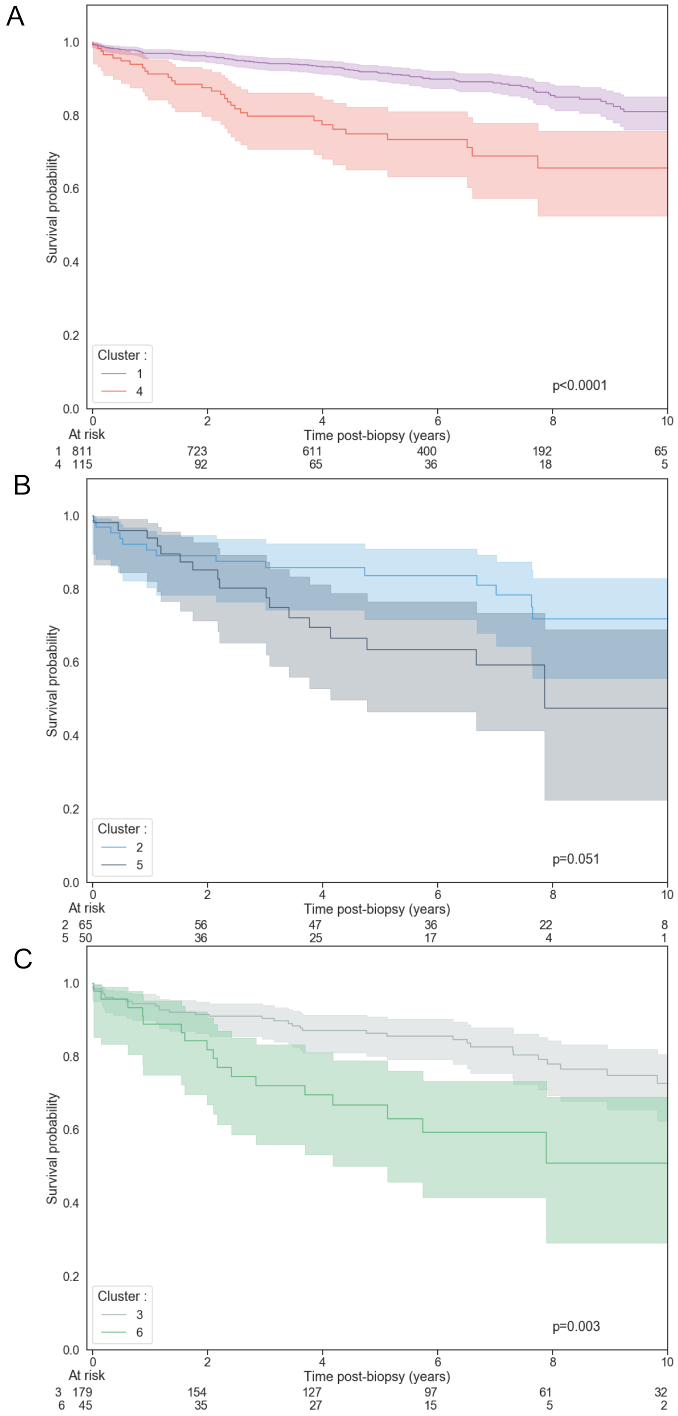


Figure 4.A.5: Post-biopsy graft survival in the three DSA-/DSA+ pair of clusters.
A. Cluster 1 vs. cluster 4; **B.** Cluster 2 vs. cluster 5; **C.** Cluster 3 vs. cluster 6). These plots demonstrate the negative impact of the presence of DSA on graft survival. P-values refer to HR from Cox models.

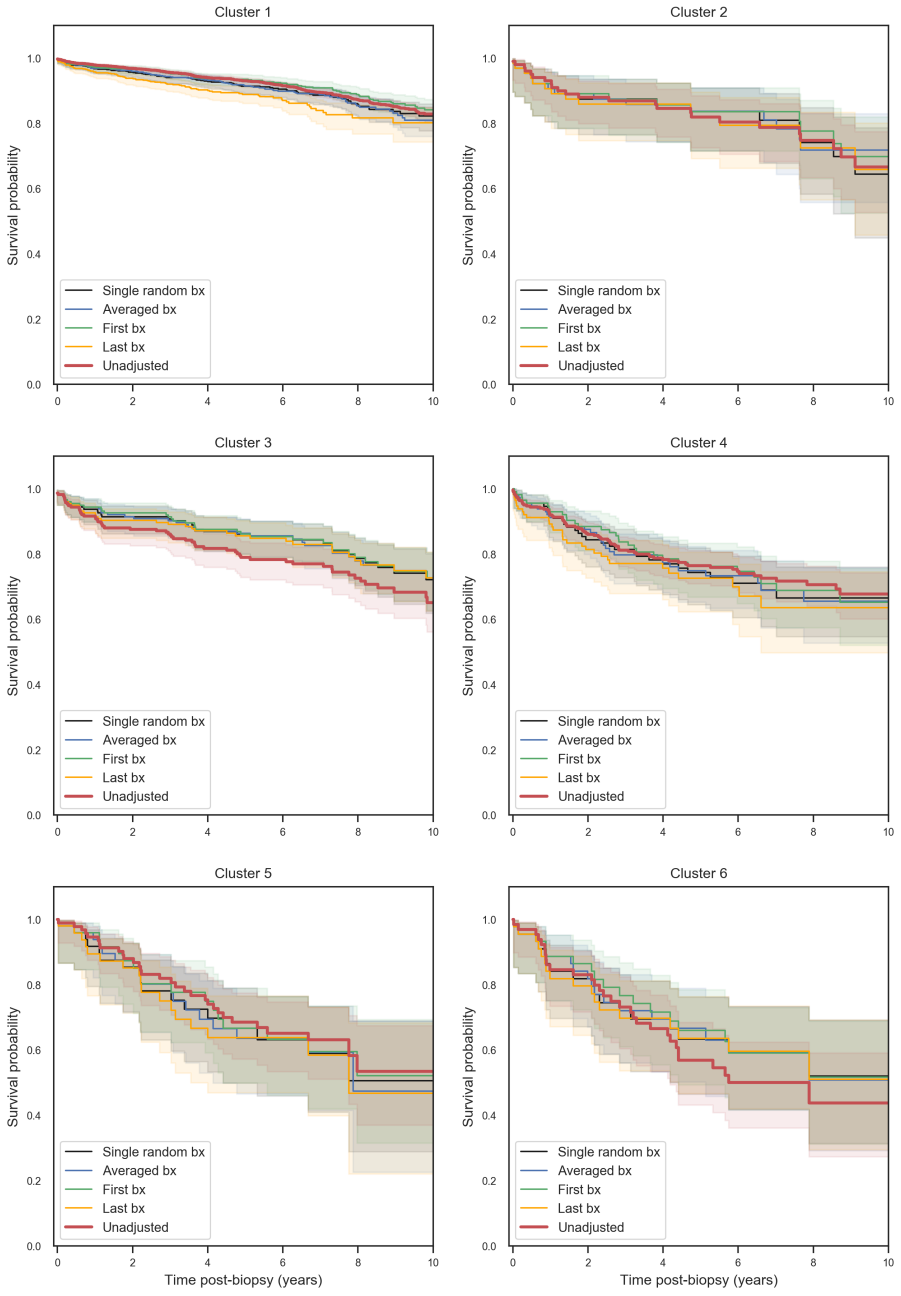


Figure 4.A.6: Post-biopsy graft survival in the six clusters, according to the adjustment method for repeated biopsies per patient.
The following Kaplan-Meyer curves demonstrate various adjustment method for repeated biopsies per patient in Survival times from repeated biopsies were either averaged for each patient, based on a single random biopsy, using the first or last biopsy per patient. bx=biopsy.

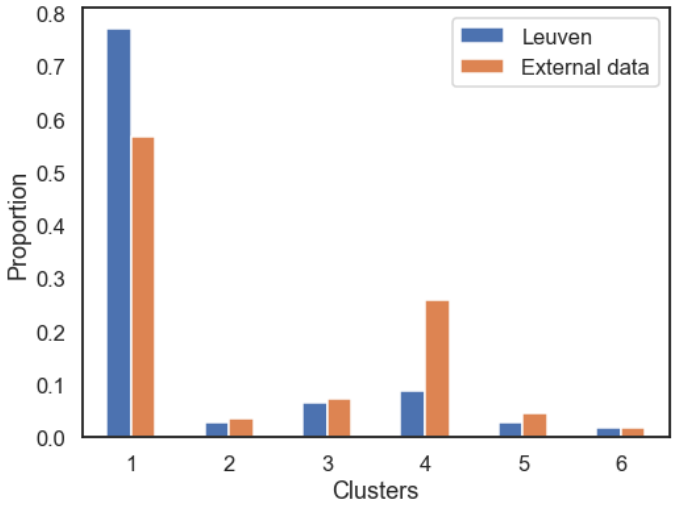


Figure 4.A.7: Comparison of cluster proportion per center
N=3510 biopsies from the derivation cohort (Leuven) and 3835 biopsies from the external validation cohort. Non inflamed clusters 1 and 4 were the dominant clusters in both datasets. The larger prevalence of HLA-DSA positive biopsies in the external data is reflected by a higher proportion of cluster 4 biopsies in the external dataset compared to the Leuven dataset (26.0% vs 8.7%, $p<0.0001$). Other clusters have similar proportion in both datasets.

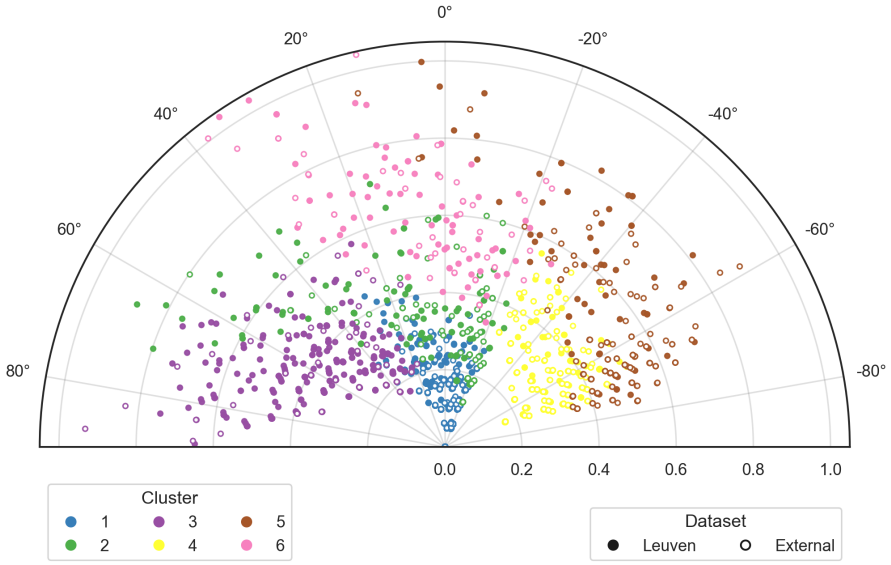


Figure 4.A.8: Overlay of the data from Leuven and the external dataset in the polar plot, according to the six clusters identified in the derivation cohort. This plot demonstrates the full overlap in the histological presentations between the derivation and validation cohorts.

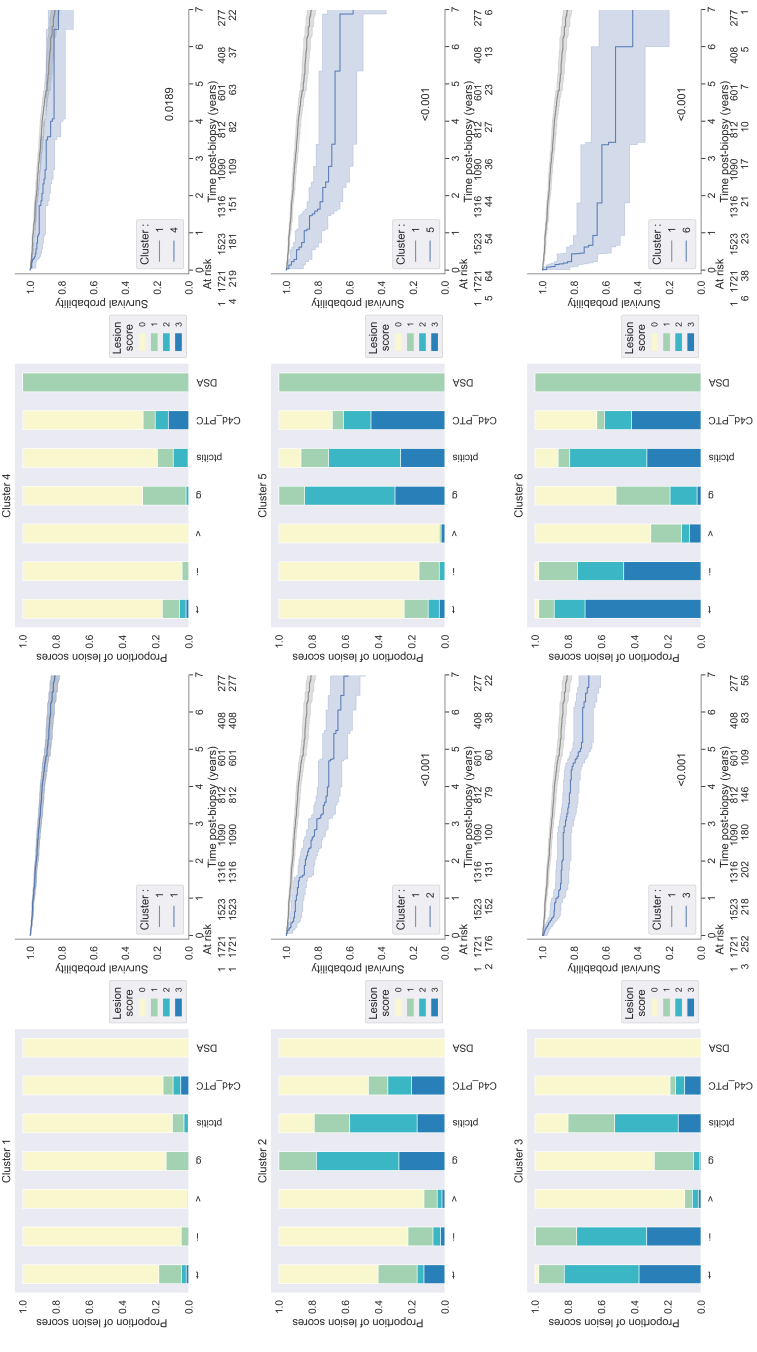


Figure 4.A.9: Distribution of the individual acute lesion scores in the different clusters, and post-biopsy Kaplan-Meier graft survival curves relative to cluster 1 of the external validation cohort (N=3835 biopsies).

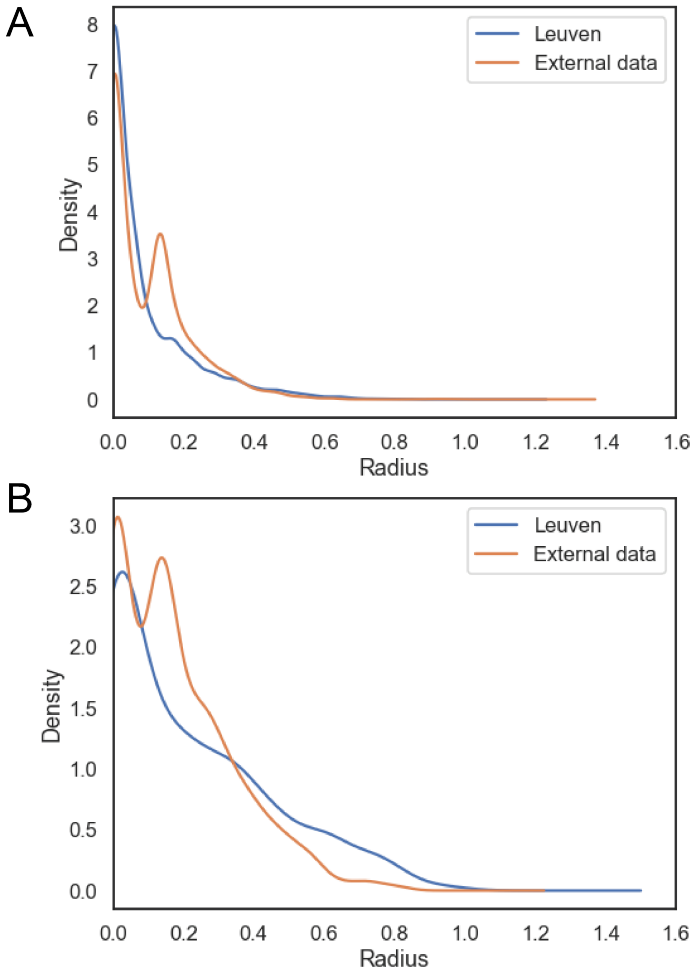


Figure 4.A.10: Distribution of the radius from the polar plot
A. Distribution of the radius on the polar plot for protocol biopsies, according to center (N=2737 biopsies from Leuven; N=2393 biopsies from the external validation cohort). **B.** Distribution of the radius on the polar plot for indication biopsies, according to center (N=773 biopsies from Leuven; N=1442 biopsies from the external validation cohort). Note the extra peak in the external data corresponding to the larger proportion of DSA positive biopsies without inflammation.

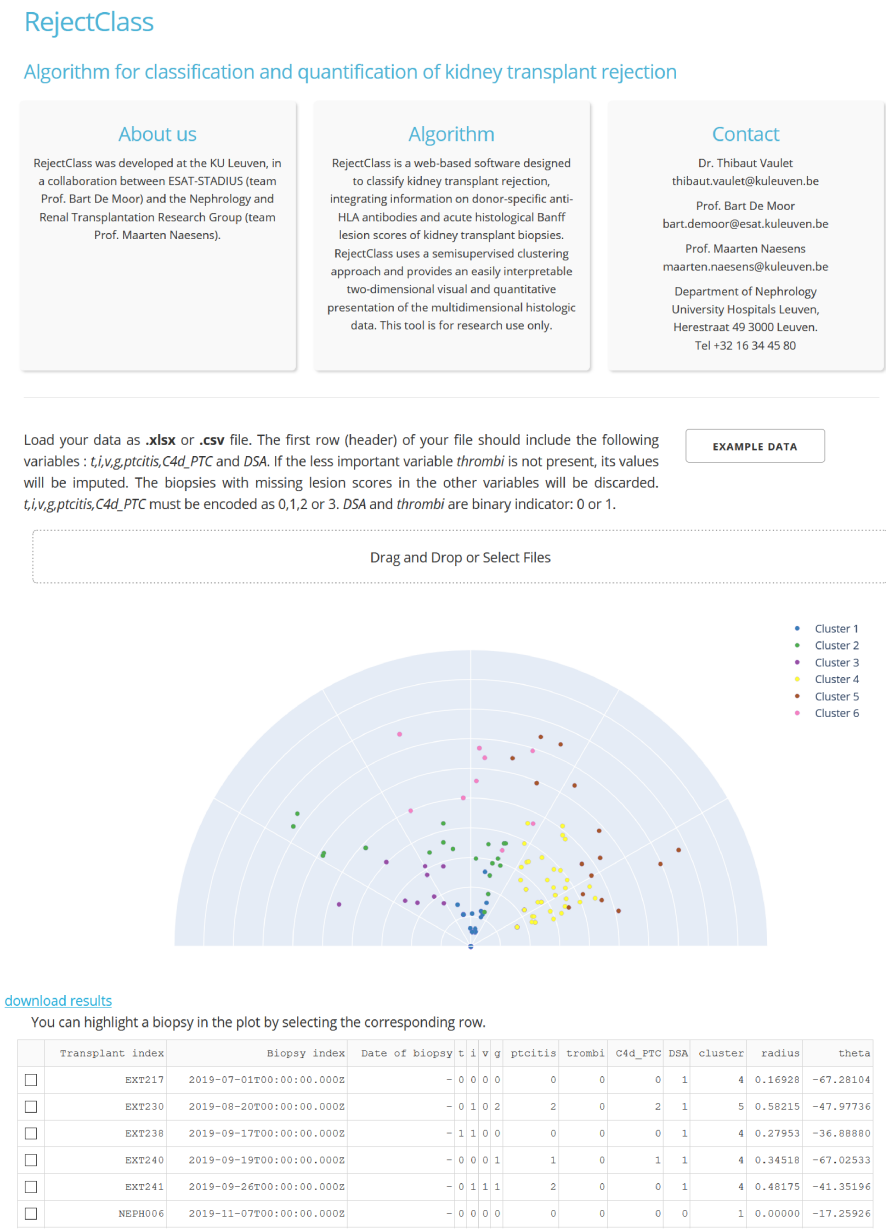


Figure 4.A.11: Interface of the free web-based application available at https://rejectclass.pythonanywhere.com
The users can upload their own database as an Excel or a .csv file, following a predefined structure. The algorithm computes the classification of the biopsies on the fly and returns the corresponding clusters. The whole set of biopsies is displayed on the polar plot following the approach developped in 4.2.2. Individual biopsies can be highlighted on the plot by ticking the boxes on the left. Additional information on the radius and the theta angle are readily available for the users. Finally, all the results can be easily downloaded as an Excel sheet for further analysis.

Chapter 5

A generic DDSS framework

Objective

Outline the structure of a generic data-driven diagnostic decision support system.

From the original research developed in this thesis, we can elaborate a generic structure for DDSS that provide diagnostic and prognostic risk estimates based on individual patient specifications. These data-driven DDSS integrate calibrated predictive models with interpretable explanations, along with an intelligent patients search engine to retrieve similar previous cases. As such, DDSS can be decomposed into five key components : *the local database, the query pre-processing engine, the patient similarity metric, the internal knowledge* and *the interpretability component*. This chapter discusses those five elements with additional considerations from a theoretical perspective. An overview of the proposed generic DDSS structure is displayed in Figure 5.1.

5.1 The databases

With the democratization of computational resources, it is now cheaper than ever to collect, store and process data, leading to an explosion of datasets in every industry. In this current data-driven environment, data constitute the mandatory fuel for machine-learning and data-driven algorithms. However, the results of those algorithms directly depend on the quality, the quantity and the

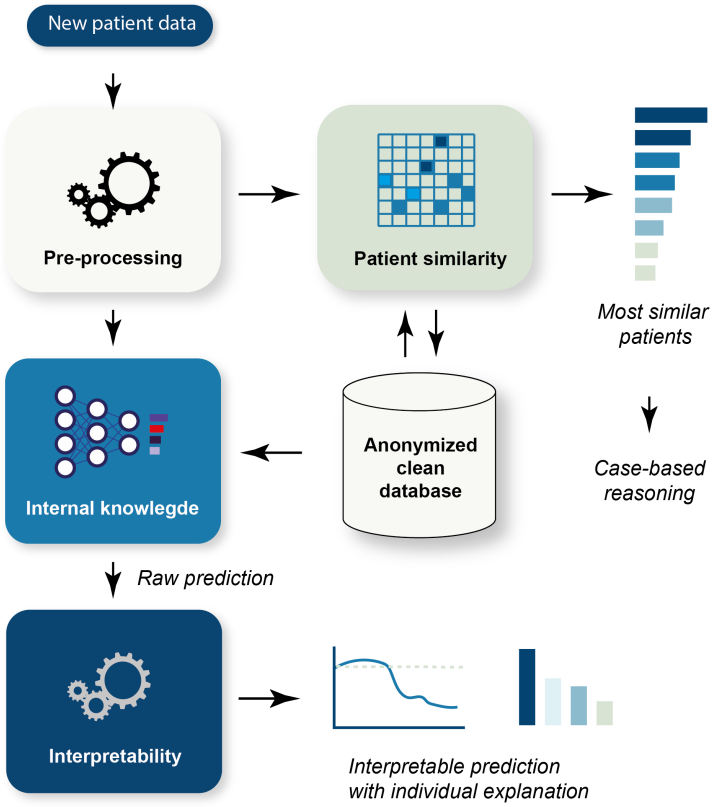


Figure 5.1: Overview of the generic DDSS structure with 5 key components. To ensure the appropriate behavior of the trained predictive models, new patient data must first be *pre-processed* following the same processing steps as performed during model development. Once pre-processed, the data instance triggers a raw prediction from the *internal knowledge*, abstracted here as a predictive model. Finally, the prediction is decomposed into explainable outputs with the *interpretability* interface, a mandatory element to translate complex diagnostic predictions into intuitive and understandable knowledge for the end-users. The local *database* plays a crucial role in data-driven DDSS, as it provides the raw material from which the *internal medical knowledge* is extracted. In addition, the database constitutes a valuable resource for the retrieval of similar patients to enable case-based reasoning. To that end, the design of tailored *patient similarity metrics* is essential.

processing of the data available during the training phase. It naturally results that data occupy a center piece in data-driven DDSS.

While the digitalization of healthcare has facilitated the creation of extensive clinical databases, the majority of health-related data remains unstructured and non-curated to be directly exploitable for data-driven DDSS development. The secondary use of local structured clinical databases constitute a candidate of choice for an efficient development of DDSS.

5.1.1 Structured and unstructured data

A common way to categorize health-related data is to differentiate two categories : *structured* vs *unstructured* data. Structured data are characterized as *"organized into specific fields as part of a schema, with each field having a defined purpose"* [192]. Therefore, structured data are much more straightforward to analyze and to process than unstructured data. However, they require the existence of a pre-defined data model, with pre-defined variable descriptors. Structured data are often stored using well-established frameworks, typically based on relational databases such as SQL. A prototypical example of structured data is the tabular *Excel* datasheet, where a column encodes a unique and predefined feature, while the rows can represent individual patients or different visits.

On the other end of the spectrum, unstructured data are data that are not organized in a clear, standardized and defined framework. A significant part of useful medical information is currently stored as unstructured data. A common estimation of the ratio between structured and unstructured data revolves around 20% vs. 80%, i.e. the vast majority of data in the healthcare industry remains unstructured [193]. A typical example of unstructured data consists in physician notes, i.e. the textual information collected during a medical consultation or after a daily patient examination. These notes often contain valuable information about symptoms, dates, diagnoses, medications,... but the core medical knowledge is not directly accessible for machine learning models. With regards to those observations about unstructured medical data, one of the biggest challenges to unlock the full potential of machine learning in healthcare facilities will be to improve the development of structured and standardized databases. The major differences between structured and unstructured data are reported in Table 5.1.

We can summarize the range of clinical data available to build data-driven DDSS in a continuum, as shown in Figure 5.2. While there is a large amount of unstructured and non-curated datasets collected on a daily basis within healthcare facilities, these data are not directly exploitable for DDSS development. On the other hand, clean and structured datasets, often resulting

Table 5.1: A summary of the major differences between unstructured and structured health-related data.

	Structured data	**Unstructured data**
Prior structuration	Pre-defined data models	No pre-defined structure
Proportion in EHR	∼20 %	∼80 %
Collect	Requires active collect	Often automated
Knowledge extraction	Easy to query	Difficult
Pre-processing	Limited	Intensive
Internal representation	Tabular data, Relational databases	NoSQL databases
Healthcare examples	*Lab results, billing data,...*	*Physician notes, MRI scans,...*

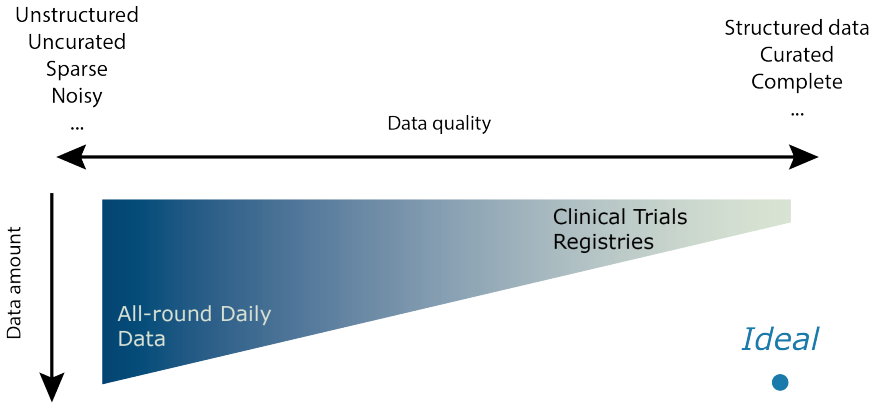


Figure 5.2: Theoretical representation of the type of clinical data collected in healthcare.
Most of the clinical data remain unstructured and non-curated (the left side of the spectrum) while the ideal situation to develop accurate DDSS consists in having large curated and structured datasets (the right side of the spectrum). This type of data is often limited to clinical trial registries, which enforce strict data collection protocols.

from strict clinical trials protocols, are rather scarce but constitute candidates of choice to efficiently develop DDSS. Naturally, many datasets will fall somewhere in-between of those two extremes.

5.1.2 The main characteristics of clinical datasets

Although it is impossible to depict a general assessment applicable to all medical datasets, most of the structured medical datasets will share the following characteristics : *noisy, sparse, longitudinal, biased* and *incomplete*.

Noisy

Medical data, as resulting from biological variables measurements, will always present some random variations. In addition, other types of noise can perturb the data. These include measurement noise, encoding errors and signal filtration and discretization. Fraudulent data manipulation can also introduce artificial data behavior, apparent to noise. The presence of noise needs to be accounted for, else models will produce unreliable estimates and/or unstable results. Collecting enough data is often sufficient to mitigate the inherent random variations of biological signals. However, large datasets will not prevent the introduction of systematized measurement errors.

Longitudinal

The temporal dimension is omnipresent in healthcare-related data as life and diseases are both dynamic processes. In longitudinal data, or panel data, some variables might be considered static, i.e. time-independent (e.g. date of birth, address of residency), while others are time-dependent (e.g. glycemia, blood pressure...). Depending on the problem at hand, taking this time-dependency into account during the modelling process might be imperative (or not) to accurately depict the clinical reality. Time-series data can be considered as a special case of longitudinal data. In practice, data points are generally equally spaced in time-series, whereas longitudinal clinical data are recorded irregularly at various time intervals. This poses an additional challenge as different patients will have data available for different time points. Some medical datasets are constituted of cross-sectional data, where the temporal dimension is omitted. These data are collected at a specific time point or across a pre-defined period of time, which facilitates subsequent modelling but bears the risk of oversimplifying the reality.

Biased

Biased datasets contain data that do not accurately depict reality. In turns, algorithms trained on biased data produce biased predictions. For instances,

race or gender discrimination is now well documented in several machine learning algorithms [194]. As some elements are more heavily represented than others, the databases skew in favour of certain demographic trends. Although such biases are often unintentional and simply reflect the existing distribution of minorities in a general population, some subpopulations become hampered by the algorithm predictions, leading to poor accuracy and poor generalization in these population strata.

Biased results can also appear from a mismatch between the training data population and the population on which the model is subsequently applied. Because randomized clinical trials follow strict exclusion and inclusion criteria, these biases are common when generalizing their result to a population outside of the pre-defined criteria. Using in-house data to develop DDSS is a convenient approach to mitigate this generalization problem but is subject to in-house data quality control. Finally, because many datasets are populated by human beings, they implicitly inherit some cognitive biases of the human mind. A common cognitive bias encountered in medical databases is due to the recall bias [40]. In this measurement bias, patients fail to accurately recall past events. It follows that the recalled memories can be significantly different from the observed reality. Therefore, the databases populated with recalled data should be analyzed carefully when used within DDSS.

Sparse

Patients attend the hospitals for different reasons, meet different physicians who prescribe different exams and treatments. Therefore, the available features of a given patient are typically only observed for a tiny fraction of the thousands of possible variables measured and registered in an EHR. In structured databases, the other variables remain empty leading to sparse data representations. Sparsity introduces an additional challenge for data analysis, as it becomes more difficult to assemble homogenous datasets of patients. Algorithms based on pre-defined variables do not accommodate sparse representation. A preliminary feature selection step is therefore necessary to constitute datasets with a common set of pre-defined variables. In opposition, the data collected during clinical research are often highly structured and not sparse, as these studies follow rigorous and pre-defined protocols. As a result, the same set of features is registered for every patient.

Incomplete

The notion of incomplete datasets relies on the concept of *missing data*, which implies that a value was expected but has not been provided. This concept is important, as regarding the sparsity of medical datasets, any empty variable cannot be considered as properly *missing* in medical repositories [1]. In consequence, a pre-defined framework is necessary to designate what variables should be complete or not.

In practice, missing values result from various causes. Encoding errors can produce unrealistic values forcing the scientist to discard the data and create a missing entry. Patients can refuse to answer certain questions (e.g. sensitive questions about sexuality or drug consumption) or unintentionally skip an item while filling in a form. Physicians can forget to order a test or encode a value. Additionally, patients can drop out of a clinical study. Finally, technical problems such as storage failure or altered samples can lead to losing data or systematized missingness. Because many algorithms cannot deal natively with missing values, missing data is an important problem in clinical modelling. Strategies to mitigate the presence of missing values must be included during DDSS development.

5.2 Pre-processing

The main characteristics of medical data listed above introduce several challenges for subsequent data analysis. In order to be efficiently processed by an algorithm, a certain number of preliminary steps, referred as data *pre-processing*, needs to be performed on the original data. In this section, we briefly describe the main pre-processing steps that should be included within a data-driven DDSS framework.

5.2.1 Data cleaning

Clinical predictive models expect data to follow preset specifications. The aim of *data cleaning* is to produce datasets that are error-free, homogeneous and

[1] The EHR data from two different patients will very often not share the exact same set of variables, unless the data acquisition is part of a pre-defined protocol. Therefore, if the variable $\boldsymbol{x}$ is measured in patient A but not in patient B, the variable $\boldsymbol{x}$ is not strictly speaking missing in patient B but was simply not collected. Similar difficulty occurs with the irregular sampling of longitudinal data as it introduces a *"time sparsity"* in the dataset.

comply with existing standards. One essential step of data cleaning is *outlier analysis*, sometimes referred as anomaly detection[2].

While some outliers are simply real extreme values, others can result from encoding or measurement errors. Erroneous data might produce unrealistic measurements, e.g. negative age or impossible height. Occasionally, while individual values seem within the normal variable range, the combination of two or more variables is unlikely (e.g. gender = women, diagnosis = prostate cancer). Such outliers are more difficult to detect. These errors can partially be prevented by introducing sanity check and quality control upstream, i.e. during the data encoding.

Although encoding errors should always be corrected, some experts advocate not to perform any outliers removal, as raw data including outliers are actually more representative of the data encountered in real-life conditions [199]. Arbitrarily removing some instances from the original dataset also questions the scientific validity of outliers analysis. Ultimately, the definition used to label a data point as an outlier and the choice to discard it or not, is highly dependent on the problem and the data at hand.

Data cleaning also encompasses data formatting, i.e. revise the data to a standardized format (e.g. standardized data/time format, international units). Data formatting is a necessary step to ensure the data-points are comparable and compatible with the algorithm on which they will be used. Overall, the process of data cleaning is often time intensive but it is a necessary step toward uniform and error-free datasets.

5.2.2 Feature selection

Like many of modern datasets, structured medical datasets tend to have a high dimensionality d, i.e. they contain a very large number of variables. Within a healthcare facility, raw datasets cover the whole range of available variables. However, only a subset of variables is often relevant to solve a specific medical problem.

The selection of a subset of meaningful variables is known as *feature selection* and often constitutes a necessary step prior to model development. As a dimensionality reduction method, one of the aims of feature selection is to reduce the risk of overfitting [129], i.e. the risk of developing models that fit

[2]Outliers are data points that significantly differ from the rest of the observations. While there is no standard definition of what constitutes an outlier, different definitions and testing procedures exist in the literature (e.g. 3 times higher or lower than the interquartile range [195], Cook's distance [196], Chauvenet's criterion [197], Grubbs's test [198],...)

too accurately the training dataset and lead to poor generalization ability on external data. Another benefit of feature selection is to limit the distance concentration phenomenon which appears in high-dimensional spaces.

In practice, feature selection can be automated or manual based on prior knowledge. Automated approaches are data-driven and can be performed in a univariate or multivariate fashion[3]. While automated feature selection is widespread in the scientific literature, solely and blindly relying on automated approaches to perform variable selection is often criticized as it frequently leads to unstable results, spurious selection or arbitrary cut-offs choices [200]. In clinical model development, a manual selection is often more appropriate and consists in selecting a pre-defined set of potential variables based on prior similar studies and/or expert knowledge [129] as we did in Chapter 3 prior to model development.

5.2.3 Feature transformation

Additional feature transformations are sometimes needed to make a dataset suitable for subsequent analysis. For instance, standardization and normalization are often performed prior to Euclidean distance measurements to ensure that the variables expressed in different units have the same impact on the final distance. Other examples of feature transformations include the *min-max* scaling of continuous variables $x' = \frac{x - min(x)}{max(x) - min(x)}$, which scales the variable to the unit interval [0,1] and *Winsorizing*, sometimes erroneously referred as *feature truncation* [4], which is performed to limit the impact of outliers measurements by replacing extreme values with truncated version of themselves, based on pre-defined thresholds [135]. See Chapter 3 for an example of Winsorizing.

[3]Common univariate strategies rely on p-values, correlation coefficients, chi-squares, variances analysis,... along with pre-defined thresholds to select a subset of predictive variables. Multivariate approaches assess the variables importance jointly, often using a multivariate model. L1-regularization methods, such as LASSO, and stepwise methods (backward elimination, forward selection) are common candidates to perform multivariate feature selection [81]. These multivariate approaches are sometimes referred as *wrapper* selection methods (when a search of the feature space occurs jointly with model optimization, e.g. backward selection) and *embedded* selection methods (when the algorithms embed a feature selection mechanism internally, e.g. LASSO.). Conversely, methods where the selection is performed independently of a model, are referred as *filter-based* selection.

[4]The proper usage of truncation is the exclusion of values above or below a given threshold.

5.2.4 Feature extraction

Feature extraction is another common pre-processing step beside feature selection and feature transformation. It refers to the construction of novel features based on meaningful combinations of original features. Such combinations can be obtained with linear (e.g. PCA) or non-linear transformations (e.g. auto-encoder). Feature extraction fulfils a triple goal: reducing redundant information, reducing the original dimensionality and facilitating the data understanding (e.g. derive high-level feature representation instead of individual pixels). Feature extraction is often performed on unstructured data such as images where individual pixels do not bear meaningful value. Unsupervised clustering of clinical variables is a form of feature extraction performed to reduce the impact of multi-collinearity in regression analysis [129].

5.2.5 Missing values

The problem of missing values is regularly overlooked in many machine learning publications, where clean data is often assumed from the beginning. However, for practical DDSS whose finality is to be deployed in a real clinical environment, assuming neat and complete data is unrealistic. Therefore, strategies to overcome the omnipresence of missing values in medical datasets must be considered during the development of DDSS.

The main strategies to mitigate the presence of missing values can be classified into three main categories: *deletion methods, imputation methods* and *the native handling of missing data*. Naturally, if possible, the first approach to consider is to retrieve the missing data from the source. However, this is often difficult in practice, as some missing values are irrevocably lost or absent.

Deletion methods are represented by *listwise* and *pairwise* deletions. Complete cases analysis or listwise deletion simply discards a case if it contains a missing value, i.e. the whole instance is discarded from analyses, even if it contains only one missing value. This method often introduces severe biases and decreases the overall statistical power by discarding a significant portion of information. A slightly modified approach is called pairwise deletion. In this setting, cases are discarded only if the subsequent analysis requires a variable that contains missing data. As long as the cases are complete for the subset of variables included in the analysis, the cases remain. Pairwise deletion is also at risk of introducing biases. Furthermore, as different analyses are now based on different subpopulations of the original sample, any sound comparison becomes extremely difficult, if not meaningless.

Imputation approaches substitute a new value where the data is missing. The main advantage is that imputed datasets can be plugged into any traditional algorithms requiring complete data without loss of statistical power. However, the choice of the imputation technique is primordial to guarantee accurate results. Imputation methods can be classified as single or multiple imputations.

Finally, some methods can handle incomplete data natively. Unlike imputation approaches, those approaches do not provide a completed dataset. They use incomplete data as such and maximize the information available without the need of imputing missing values. A common representative of these methods is the Full Information Maximum likelihood [201]. Several machine learning models also allow missing data. However, they often rely on specific data assumptions. As we have seen in Chapter 2, some implementations of gradient boosting tree algorithms also tackle the problem of missing values. During the training phase, the algorithm learns default and optimized paths in presence of missing values. At prediction time, the variables with missing values follow those default paths and allow the algorithm to produce a prediction even with incomplete data.

5.3 Patient similarity metric

As we have seen in Chapter 2, the notion of patient similarity, defining to what degree two cases are considered similar or not, is highly important for personalized approaches and case-based reasoning. Patient similarity has been applied in a wide range of applications (See [202–204] for recent and extensive reviews on the topic.)

Although the requirements of a patient similarity metrics are often dictated by the problem at hand, many studies used default and general-purpose metrics such as the Euclidean distance or the cosine similarity metric. While this choice might be relevant in some settings, the use of those generic metrics is often suboptimal regarding the variety of challenges and constraints imposed by real clinical databases. Based on the specificities of medical datasets as described previously in Section 5.1.2, a patient similarity metric, in the context of DDSS, should validate the four following properties : *task-specific, robust to incomplete data, adapted to heterogeneous data types* and *interpretable*.

5.3.1 Task-specific

As already pointed out by Ng et al. [28], in a medical context, different doctors face different clinical scenarios. Therefore, the notion of similarity should often be expressed regarding the risk of an external outcome [28]. Such outcomes typically consist in the risk of developing a disease or the estimated disease-free survival [42]. For instance, two patients A and B might be considered similar regarding their risk of prostate cancer but might present a very different risk profile regarding cardio-vascular diseases. Therefore, while A and B are considered similar by urologists, they will likely be considered as very different by cardiologists. For that reason, general-purpose metrics, e.g. Euclidean distance, often underperformed compared to task-specific approaches informed by additional outcomes [43–45, 80]. In Chapter 2, we demonstrated that metrics derived from tree-based ensembles also outperformed the generic standardized Euclidean distance. As a result, the patient similarity (or distance metric) should be tailored for the problem at hand.

5.3.2 Robust to incomplete data

As previously mentioned, missing values are omnipresent in clinical databases. In the context of similar patients retrieval, incomplete databases pose an additional challenge for metric learning. The vast majority of existing patient similarity metrics do not accommodate the presence of missing values. In the literature on patient similarity metrics, most of the studies simply discard incomplete cases [205–210]. Others simply impute the missing values with the variables mean/median values [211]. Although simple to implement, these naive approaches usually discard significant portion of valuable information and/or introduce severe biases in subsequent analyses [48–50]. In addition, missingness can also appear at the query-level. Traditional approaches, even based on complete cases, do not explicitly allow incomplete query. It is therefore fundamental to develop patient similarity approaches that accommodate missing values. The pseudo-distance derived from tree-based ensembles developed in Chapter 2 mitigates the issue of incomplete data.

5.3.3 Adapted to heterogeneous data types

Clinical data repositories contain various sources of data which are frequently heterogeneous in terms of variable types and formats. While traditional metrics often target a unique data modality (e.g. Euclidean distance for continuous data),

generic patient similarity metrics should account for the large heterogeneity of clinical variables encountered in practice.

5.3.4 Interpretable

In a patient search engine, the quantification of patients similarity should be interpretable for the physicians, as it should be clear why two patients are considered similar or not. Although no general guidelines exist to develop interpretable metrics, we propose the following characteristics as necessary features for interpretable similarity metrics: *low dimensionality*, *additive* and *related to the original input space*, i.e. the medical units.

While advanced machine learning algorithms easily handle hundreds of different variables, the cognitive faculty of the human mind imposes constraints on the maximum number of items one can reason about in a short period of time. In addition, the number of relevant variables to establish a diagnosis or a prognosis, is often limited in practice. Not to mention that, due to the sparsity of medical datasets, the set of common variables that are clinically useful for a problem at hand is often reduced. A lower dimensionality will also prevent the phenomenon of *distance concentration* [212, 213] occurring in high dimensional spaces[5]

Similarly to linear regression, the additivity property facilitates the comprehension of the overall distance values. This property makes it straightforward to decompose the overall distance as individual variable contributions. It also results that complex feature interactions impede the overall interpretability of distance metrics. Finally, although the internal distance can be computed on scaled variables to account for the variables heterogeneity, it should be evident how such scaling affects the overall distance or similarity computation. Therefore, the results of a similarity search should be delivered after proper re-conversion into meaningful medical units. This can be achieved by appropriate visualization of the different variables' weights or by using post-hoc interpretability strategies such as feature attribution methods.

5.4 Clinical predictive models and inference

On a daily basis, the two most difficult questions a physician has to answer are : *what is the right diagnosis for this patient ?* and *what is the expected*

[5]As the dimensionality d increases, the relative distance between data instances decreases, translating an almost equidistant set of data points in high dimensional spaces. The direct implication of distance concentration is that for large d, a slight change in the similarity threshold will result in large changes in the nearest neighborhood composition.

prognosis ? The answers to those questions are crucial as, with additional information on the patient specificities, they will govern all future investigations and/or therapeutic interventions. As a result, the main objective of a DDSS is to provide information to better answer these key questions and therefore, to perform enhanced medical decisions. DDSS can provide predictions of the diagnosis itself or provide additional pieces of information that will be integrated in the diagnostic process by the physician. Beyond fancy user-interfaces, the core of any DDSS is the embedded medical knowledge that is actuated to support the medical decisions.

While traditional DDSS, sometimes referred as *knowledge-based* systems, encode explicit rules to guide decisions, in data-driven DDSS, medical knowledge is abstracted as a predictive model or as a decisional algorithm trained on the appropriate data.

In machine learning, predicting the right diagnosis is depicted as a classification problem, where given an instance $\boldsymbol{X}_i$ one aims to predict its correct class k_i. In this regard, it is necessary to differentiate models that produce hard classification, i.e. where the outcome is the class itself, from soft classification methods that assign a class membership score or probability[6].This distinction, often overlooked in machine learning, might appear trivial, but is highly relevant for (clinical) decision-making. *Hard classification*, by producing crisp outcomes, can be considered as a premature decision, as it actually combines prediction with decision-making [216]. On the other hand, *soft classification* estimates patients' risks but do not produce a decision. It is therefore possible for the physician to further optimize a decision based on patient-tailored risk thresholds and additional clinical context (e.g. the cost of misdiagnosis, specificities of the patient, practical constraints,...) unavailable to the model. In fact, many of the medical decisions are based on probabilistic bets. Very often, physicians have only access to partial information to make a diagnosis, as costly analysis would be required to achieve diagnoses with an unnecessary degree of certainty (e.g. sequencing the genome of a virus to diagnose an influenza infection). Because DDSS aims to support a decision, and not to provide a decision itself, the use of hard classifiers in clinical practice should be limited.

Historically, predictive modelling in the medical field hinged on regression models. With the advances of machine learning, new types of models are now

[6]For instance, support vector machine (SVM) [214], a margin-based machine learning algorithm, typically produces hard classification by attributing classes based on the side of the optimal hyperplane (the decision boundary) where the instances are located. (Note however that some modifications involving Platt scaling extend SVM to a probabilistic approach [215]). Contrarily, the logistic regression model, a soft linear classifier, produces continuous scores (and probabilities) instead of hard class predictions. The final hard classification can be achieved under the *argmax* operator, i.e. the class with the highest probability or membership score.

Table 5.2: Main regression models based on the type of outcome

Type of outcome	**Example of regression model**
Continuous	Linear
Binary	Logistic
Categorical	Polytomous logistic
Ordinal	Proportional odd logistic
Survival	Cox proportional hazards

frequently encountered in the medical literature. In the following section, we briefly develop the models that are of particular interest for clinical predictive modelling in DDSS.

5.4.1 Regression-based methods

Generally speaking, *regression* analysis estimates the association between a *dependent* variable and one or several *independent* variable(s). Those models rely on the linear combination of variables $\boldsymbol{X}_i$ with the learned coefficients $\boldsymbol{\beta}$. The additive constraint of multivariable regression restricts the model capacity but facilitates the understanding of the prediction process. To overcome the lack of flexibility from the linear combination of variables, additional complexity can be introduced with interactions and/or non-linear transformations of the independent variables. Unlike, data-driven machine learning alternatives. those model specifications must often be manually specified upfront by the statistician, although some (semi)-automatic procedures exist (e.g. multivariable fractional polynomial [136])

The choice of the regression model is mostly dictated by the type of outcome to predict (Table 5.2). A common requirement to all regression models is that data instances $\boldsymbol{X}_i$ are assumed to be independent of each other. In addition, because regression models are based on a linear combination of the variables, linearity and variable additivity are assumed to some extent.

5.4.2 Machine learning approaches

Beside the (semi)-parametric regression-based models, non-parametric alternatives based on machine learning have emerged from the computer science and the

engineering field [7]. In comparison with traditional regression-based methods, machine learning approaches implicate less human decisions in the model specification [223] as the process is mostly driven by the data.

Machine learning methods mostly focus on building a mapping from $\boldsymbol{X} \longrightarrow \boldsymbol{y}$ that can generalize well on new and unknown data, irrespective of the underlying mechanisms. In contrast to regression-based models, which are founded on a linear combination of variables, machine learning models are often intrinsically non-linear and non-additive. Therefore, they offer a greater flexibility and a greater capacity, i.e. they can approximate a wider range of functions. In addition, they tend to focus less on matter knowledge unlike traditional methods which leverage prior and expert knowledge during the model specification [224]. For those reasons, they often require larger datasets to be trained.

Decision trees and tree-based ensembles are particularly attractive for structured clinical datasets.

Decision-tree

A decision tree is a tree-like representation of a decisional process often represented as an upside-down tree. A tree is composed of *nodes* connected via *branches*. Each node performs a test on a given variable (e.g. age ≤ 50 years?) resulting in a *yes/no* answer and producing two children nodes. The root node is the *top* node, i.e. the first decision made by the tree. Intermediate nodes are called *internal* nodes. The terminal nodes at the bottom of a decision tree are called *leaves* and contain a constant value (e.g. a class or a numerical value). The path from the root node to a leaf describe the classification rules. A decision tree is therefore composed of a combination of *if-then-else* rules.

Decision trees offer several advantages for clinical modelling as decision trees of reasonable size are easily interpretable, even for non-expert end-users. Furthermore, they can easily accommodate mixed data type, e.g. continuous and categorical variables. Decision tree learning algorithms also integrate a feature selection mechanism such that less relevant variables are less likely to be selected

[7] While Breiman, in a stimulating paper [217], voluntarily divided the statistical modelling into two cultures (i.e. data modelling vs algorithmic modelling), we should note that the dichotomization between traditional regression-based models and machine learning methods is somehow arbitrary. For instance, machine learning experts like Goodfellow, Bengio and Courville considered logistic regression as a legit machine learning approach in their notorious book about Deep Learning [218]. The distinction appears even more ambiguous considering models such as LASSO [219], ridge-regression [220] and Elastic-net [221], which are extensions of regression-based models with regularization terms to prevent overfitting and are commonly used by data scientists. Most of the machine learning software toolboxes include regression-based models among their arsenal [93, 222].

for a node split. In fact, many medical procedures are already summarized with decision trees [225]. However, the major drawback of decision trees is their instability, as a small change in the dataset can result in a completely different tree. Additionally, complex trees also easily overfit the training data and offer poor generalization performances. Optimization strategies such as the pruning of the tree or early stopping must often be implemented to guarantee good generalization performances. Decision trees also require binary splits of continuous variables, which results ultimately in a loss of information [226].

Tree-based ensembles

To overcome the instability of a single decision tree, a solution consists in combining multiple trees together to form an ensemble or committee of trees [8]. The predictions are then obtained from a consensus process over the ensemble (e.g. average or majority voting). As a result, one of the main advantages of using an ensemble of single classifiers is to lower the variance of the predictions [81]. Together with Random Forest, gradient boosted trees constitute flexible models for predictions on clinical datasets.

Beside tree-based algorithms, other eager machine learning approaches are available for supervised predictive modelling such as Artificial Neural Networks, Support Vector Machines [214] and Naïve Bayes classifier [227]. A detailed description of the vast catalogue of machine learning algorithms is out of scope of this thesis. In what follows, we examine the case-based reasoning framework, a special case of lazy learning very close to human reasoning.

5.4.3 Case-based reasoning

Case-based reasoning is a problem solving paradigm that relies on previous similar experiences to solve a current problem at hand. This concept was first introduced by Roger Shank in [228] as a model of dynamic memory. In contrast with the previously described algorithms, which are based on inductive inference, CBR exploits the notion of *transductive inference*.

The transduction paradigm was developed by Vapnik in the 1980s [229] to answer the following imperative:

> *When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you*

[8] In the ensemble terminology, the individual models (e.g. individual decision trees) that are aggregated together are often referred as *weak learners*.

> *really need but not a more general one.*
>
> V. Vapnik, 2006, *Estimation of Dependences Based on Empirical Data*, p.477, Information Science and Statistics [229]

The rationale behind transductive inference[9] is that it seems unnecessary to develop general solutions (induction) as an intermediary step to then produce individual predictions (deduction) when accurate predictions can be provided directly from the training data. In other words, the two steps, induction-deduction approach can be simplified as a single *transductive* step (Figure 5.3). Vapnik summarized this inference concept as moving *"from the particular to the particular"*. [26]. He added that *"this concept of inference appears when one would like to get the best result from a restricted amount of information"* [26]. Transductive learning seems therefore particularly suited for most of the current clinical settings.

The assumption behind CBR is that similar patients, in similar conditions, suffer from similar diagnosis and benefit from similar treatment, which seems a reasonable assumption in medicine, provided that one can measure a patient's condition accurately and can define a proper similarity metric.

A large part of a physician's medical knowledge relies on his/her past medical experience [27]. As physicians frequently reason over historical cases, CBR is much more intuitive to understand than most of the predictive models [205]. In addition, as CBR can leverage extensive repositories with a perfect recall, the queried databases can be seen as infallible and extended memory, shared among physicians [25].

Unlike traditional algorithms that only produce a prediction, CBR retrieves and displays complete past cases, which might include crucial information not required for the nearest patient search. Consequently, physicians have access to much more information than a unique prediction score [230]. In this regard, CBR can also be used as a retrospective data search engine as it can retrieve the whole set of similar patients to the physician. CBR are also particularly useful to retrieve rare cases [231].

However, because the process of CBR is very similar to human reasoning, it also inherits part of its cognitive biases. For instance, *anchoring* is well known cognitive bias whereby the decisions are abnormally influenced by a specific piece of information, called the anchor. In medicine, anchoring typically happens when a physician prematurely settles on a unique diagnosis but fails

[9] We should note that there is no consensus on the term usage in the machine learning community. For instance, as k-NN does not estimate a general function to produce outcomes, it can be considered as a transductive inference algorithm.

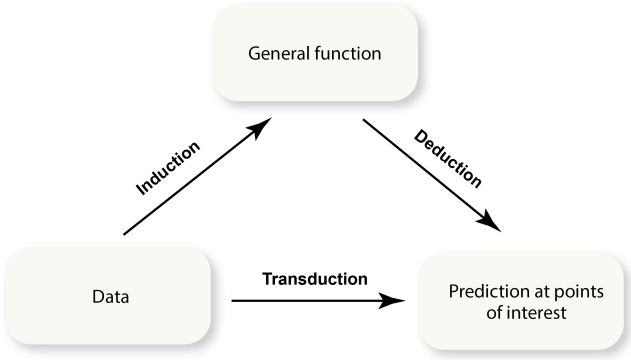


Figure 5.3: Relationship between induction, deduction and transduction. While the traditional *induction-deduction* process involves the estimation of a general function to compute new individual predictions, transductive inference reasons directly from individual instances to individual instances, without a *genealization* step. Adapted from *The Nature of Statistical Learning Theory* by V. Vapnik, 2013, Springer Science & Business Media, p169 [26]

to adjust and to question it when new, possibly contradictory, information is made available. As a result, CBR end-users should avoid relying solely on the most similar patient but integrate the provided information in a global context.

Secondly, the assumption that similar patients suffer from similar diagnoses might not always be true. Relevant factors for an accurate diagnosis can be omitted from the similarity metric definition. As a result, the similarity metric might fail to depict accurately a medical condition, producing only partial patient matches.

Finally, classical CBR does not provide any measure of uncertainty. At best, it can display some abstract similarity score with the current query. Therefore, its transductive nature makes CBR difficult to calibrate and validate, especially when only a few similar cases are retrieved.

Overall, carefully designed CBR with a relevant patient similarity metric, offer a simple, intuitive tool to assist the physicians' routine in a DDSS in complement with accurate and calibrated predictive model.

5.5 The interpretability component

The main goal of a DDSS is to support physicians' decisions with enhanced and accurate information. As a transfer of knowledge, from a mathematical model to a human user, this information should be clear, reliable and understandable. In particular, the reasons why a model predicts a given outcome should be *interpretable*. While machine learning models can achieve state-of-the-art performance in various settings, most of them remain *black boxes* where the inner mechanisms are opaque to the end-users. The lack of transparency regarding models predictions has often been advanced to explain the poor acceptance rate of advanced models, and therefore CDSS, in clinical practice[232–234]. In the following section, we introduce some important concepts related to the notion of model interpretability with regard to the DDSS specificities.

5.5.1 The notion of model interpretability

The field of machine learning devoted to facilitating the comprehension of AI solutions by human users is called *explainable AI* (often abbreviated as xAI). In the explainable AI literature, the terms *interpretability*, *explainability*, *transparency* are frequently encountered and often used interchangeably, as no consensus on the use, nor on the definitions of those terms exists [235]. Similarly, because there is currently no rigorous mathematical definition of *interpretability*, authors frequently use the term implying different meanings [160]. Dosilovic et al. argues that interpretability is often used *axiomatically* in scientific papers, as authors regularly mention it without first defining it [235]. Yet, some authors attempted to sketch a general definition of interpretability. For instance, Doshi-Velez and Kim define interpretability as the *"ability to explain or to present in understandable terms to a human"* [236]. Biran and Cotton defined model interpretability as *"the degree to which an observer can understand the cause of a decision"* [237]. Interpretability is therefore referred to the degree of human comprehensibility of a model or decision [238].

5.5.2 Interpretability : transparency & explainability

Interpretability is a complex concept that builds on several fields of science, especially on sociology, psychology and philosophy [239, 240]. In the field of machine learning, Lipton decomposes interpretability into two components: *transparency*, i.e. pre-hoc interpretability and *explainability*, i.e. post-hoc interpretability [241]. Mittelstadt et al. describe the differences between transparency and explainability, in terms of *model functioning* ("how the model

works?") vs *model's behavior* ("why has the model reached this conclusion ?"), respectively [242].

The concept of *transparency* states that the functionality of a model or an algorithm should be comprehended by a human [160]. It is mostly attained through constraints on the model complexity [243, 244]. For instance, researchers can constraint a model to be linear, additive and sparse to facilitate its understanding. An example of such approaches is the scoring system developed by Van Belle et al. which provides intuitive and simple rules using risk constraints on variable intervals, i.e. the risk is constant for each interval [245].

On the other hand, *post-hoc interpretability* addresses the model behavior as a whole or, more often, for a specific prediction. It attempts to understand how the model operates in practice, irrespective of its inner machinery. Post-hoc interpretability methods constitute therefore a form of knowledge extraction from the model [160].

This practical dichotomous taxonomy between *transparency* and *explainability* is now quite well accepted, as most of the interpretable methods fall in one of those two approaches [241, 243, 246].

5.5.3 Post-hoc interpretability builds on explanations

The notion of *post-hoc interpretability* relies on *explanations*. From a semantic perspective, *explaining* something is the act of making something clear or understandable to someone [247]. A simple but commonly used definition of explanations is that *an explanation is an answer to a "why-question"* [243, 248]. Montavon et al. describe an explanation as *"the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression)."* [246]

The implicit goal of an explanation is to facilitate the learning of mental causal models, i.e. if we observe that $y = y^*$ it is *probably* because $X = X^*$ (*probably*, because mental models are often approximate and can be wrong). Through that objective, explanations fulfill at least four different purposes:

1. Law requirement : e.g. *the right to explanation* [10]
2. Debugging : facilitate the model development.

[10] Model explainability is intensified by the European General Data Protection Regulation [36], which stipulates that every citizen has the right to be given an explanation when an algorithm uses his/her personal data. This right to explanation is however still subject to many ongoing debates regarding its applicability and its usefulness [249].

3. Improve transparency : facilitate the model understanding.
4. Increase the trust : facilitate the model acceptance.

The end-users of DDSS are mostly concerned with points 3 and 4.

5.5.4 Explanations in DDSS

Although explanations can be provided at various levels (e.g. model, data point, algorithmic process,...) [239, 242], in a DDSS setting, explanations are often required for individual prediction and should therefore be formulated at the *patient-level*[11]. In addition, since explanations should be understood by human end-users, they should refer to meaningful variables, i.e. variables expressed in a *commonly understood domain*. Explanations should also be *sparse* as the number of items a human mind can handle simultaneously is limited [250, 251]. Finally, they should be *contrasted* to a well-defined baseline. As highlighted by Miller [239], *"why"* questions are implicitly contrastive. This suggests that we often rely on an implicit expected value to compare the current outcome with. In a diagnostic setting, this could be translated as *why does a patient have diagnosis X instead of Y? Or why is he positive instead of negative ?*

Because explanations are contextual and user-dependent, it is impossible to define a universal approach to deliver them. Practically, they can be formulated as variable scores, simple rules, aggregated statistics, feature importance, heat-maps, etc.

In brief, interpretable DDSS should provide explanations that are patient-centered, sparse, contrastive in an accepted interpretable domain (commonly understood language).

5.5.5 Why are explanations needed ?

From a cognitive point of view, explanations are needed when an observation violates one's mental model [252]. The mental process engaged to find a correct explanation is called *abductive reasoning* [253]. It is a special form of inference

[11] Note that explanation can also facilitate the model transparency. However, obtaining a global understanding of a complex model behavior through human-understandable explanations appears to conflict with the very first goal of using *beyond-human-understanding* AI models to outperform human performance in predictive settings. Therefore, it seems counterintuitive and counterproductive to reduce complex algorithms to simple but understandable statements. The corollary is that explaining a complex model as a whole will often be approximate at best, inaccessible at worst.

focusing on cause-effect relationship [12] in an attempt to find the most likely conclusion from the observed data [254]. Abduction is seen as an intelligent guess. Unlike deduction, the plausible conclusions might not be correct and are not verified. Unlike induction, it does not aim to create a general rule but to explain a specific instance. Abductive reasoning is thought to play an important in the creative process, discoveries and the elaboration of hypothesis [255].

We can therefore justify the need for explainable models in DDSS, especially when the model's outcome is unexpected by the end-user. The need for explanations in a DDSS is further motivated by algorithmic accountability: although physicians remain the final decision makers, decisions that are erroneously influenced by an algorithm will raise the question of accountability.

5.5.6 Interpretability in machine learning: an unsolvable problem ?

Assessing the quality of an explanation heavily depends on the human mind and is therefore subject to many cognitive biases. For instance, the human mind tends to favor explanation (and explanation forms) that align with his own reasoning, even if a model reasons differently. Because there is no formal mathematical definition of interpretability, it is challenging to measure it objectively. In addition, explanations refer ultimately to our own mental model. Therefore, the same explanation might appear correct for someone but appear challenging for someone else.

Furthermore, while we require the models to clearly explain their reasoning, we should note that even some of our own decisions are difficult to rationalize and explain intelligibly. For instance, explaining precisely why a stranger's face looks similar to a relative is difficult to put into words. Such instinctive reasoning also appears in clinical decision-making. In 2009, Crosskerry elaborated a dual-process theory, which states that physicians rely on two different systems to make a diagnosis [256]. The first system is heuristic, intuitive and fast, while the second system is systematic, analytical and slow. Under this framework, many physicians make diagnoses based on system 1, suggesting that a complete rational understanding is not necessarily mandatory to perform medical decisions.

Some authors advocate to rely only on interpretable models for important decision making, i.e. favoring *transparency* over *explainability*, as black boxes models might still be at risk of miscalibration and miscalculation [244].

[12] In explainable AI, causality is always expressed with regard to a specific model, and not the real underlying causal mechanism(s). The causality between y and $\boldsymbol{X}$ is established through a model, i.e. $f(\boldsymbol{X} = \boldsymbol{X}^*) = y*$, and not the $\boldsymbol{X} = \boldsymbol{X}^* \implies y = y^*$, which suggests a direct causal relation. Remember that machine learning never guarantees causality.

However, this would deprive researchers from a large arsenal of powerful models. We should not be trapped by the illusion of model transparency. Because interpretability ultimately relies on the human capacity of understanding, the range of transparent models that can *really* be understood by human users is actually very limited. Even simple models such as logistic regression can be difficult to understand, especially for non-expert end-users.

While it might be of interest to have a comprehensive understanding of the model, we argue that physicians and patients are mostly interested in understanding individual predictions (similar to [160]). In addition, for many advanced models, reaching complete model transparency is likely utopic and is not entirely needed to perform decisions (see system 1 in clinical decision making [256]). Therefore, model explainability should be favored over model transparency in DDSS.

5.6 Conclusion

As DDSS mostly target specific medical conditions, research datasets often constitute good data candidates to develop DDSS, as they gather relevant data acquired under strict protocols. In addition, many of those datasets are already structured and curated to some extent, which facilitates the data pre-processing steps. Relevant medical knowledge, the core element of data-driven DDSS, can then be derived from these pre-processed databases to populate the system. This internal knowledge is often abstracted under predictive models or algorithms, which should be developed and evaluated following an appropriate model development strategy (model selection, hyper-parameters tuning, model validation). Beside predictive models, a supervised patient similarity metric is an important element to develop flexible case-based reasoning systems that can adapt to different clinical scenarios and augment physicians memory to facilitate the reasoning about similar patients. Finally, one should not overlook the importance of communicaing interpretable results to the end-users. To that end, individual explanations, provided at the patient-level, can be extracted from complex predictive models with post-hoc interpretability frameworks or interpretable surrogate models.

Chapter 6

Conclusion

6.1 Contributions

In this thesis, we studied solutions to develop tailored diagnostic decision support systems, driven by three medical challenges. These research projects resulted in the following contributions:

In the first part of this thesis, we demonstrated that the proximity matrices of tree-based ensemble provide excellent task-specific patient similarity metrics on incomplete datasets of heterogeneous variables. These metrics constitute a flexible approach to develop intelligent patient search engines on real clinical datasets, for various medical conditions while requiring little optimization.

We then developed and internally validated two simple but clinically useful predictive models for first trimester viability prediction based on multivariable logistic regression. We further demonstrated that the presence of two blood biomarkers, progesterone and β-hCG, used jointly with clinical and ultrasound features, increased the predictive performance compared to clinical and ultrasound variables alone. To extend the deployment of clinical predictive models to incomplete cases, we implemented an efficient online imputation method based on fixed chained equation and predictive mean matching, which allows queries with missing values to be processed by predictive models requiring complete data. By leveraging an additive feature attribution method, we improved the interpretability of statistical models for non-expert end-users. Under this framework, the individual predictions are explained as variable's contribution to the final model output and can be intuitively displayed to physicians with decision plots.

In the last use case of this thesis, we produced a novel classification of kidney transplant rejection that is both histologically and clinically relevant. Using a semi-supervised clustering algorithm, six new cluster phenotypes were discovered and validated on external data. Each of these phenotypes significantly associates with graft failure and overcomes the current limitations of intermediate and mixed phenotypes. In addition, a novel intuitive visualization tool that combines the biopsy's phenotype with the level of inflammation has been developed. This approach helps physicians unfamiliar with kidney transplantation to rapidly assess the rejection diagnosis and its severity on a single plot. This data-driven phenotyping of kidney transplant rejection, as successful proof of concept, opens future research directions for quantitative diseases reclassification.

6.2 Machine learning in clinical settings: useful or not ?

While advanced machine learning algorithms have demonstrated outstanding performance in many environments, including health-related problems (e.g. improving the prioritization of medical imaging [257] or diagnosing skin cancers from images [258]), their benefit over simpler models is yet to be demonstrated in real clinical predictive settings.

A recent and extensive review of machine learning approaches for clinical predictive modelling did not demonstrate any predictive performance benefit over logistic regression [259]. Additionally, many machine learning studies suffer from unfair evaluation strategies where typically a carefully crafted machine learning algorithm is benchmarked against an non-optimized baseline alternative [135]. Such concerns were already reported as early as 2000 [260].

Several authors have now warned medical practitioners and publishers not to fall for the irrational *hype* on A.I. and machine learning in medicine [135, 261–264]. Just because it has the words *A.I.* or *machine learning* in its title, it does not mean that a method necessary performs better [259], nor that it is more useful [265] than traditional approaches for clinical predictive modelling. As suggested by Steyerberg, modern machine learning algorithms might not be required as the nonlinearity and the nonadditive variable relations may not be pronounced enough for medical predictions [135]. Additionally, machine learning algorithms tend to be evaluated in terms of predictive performance instead of focusing on clinical utility [266]. Demonstrating that an algorithm has excellent discriminative power is not sufficient to claim that a model is useful [129, 135, 265].

Another charge addressed against the use of machine learning for predictive modelling is the lack of reproducibility as the complete architecture of the algorithm cannot be explicitly reported in papers. In comparison, it is easy to communicate the coefficients of a logistic regression model in a journal paper. To enforce the reproducibility of machine learning models, journals should now require operational codes and/or accessible platforms where the algorithms can be freely tested and externally validated [34, 267].

Some authors and developers also hide behind the black box aspect of machine learning to make unethical commercial use of their models [34, 244, 268]. This privatization of algorithms impedes the uptake by the scientific community and the reproducibility of results [34]. The black box aspect of machine learning also constitutes a cover to integrate potential unethical behavior into DDSS. As suggested by Char et al., opaque DDSS could be programmed to generate increased profits for their designers or purchasers (e.g. by recommending more drugs) [269].

Indubitably, machine learning algorithms offer some benefits over traditional statistical approaches. For instance, we demonstrated, in additional research not included in this thesis, that gradient boosted trees reached similar predictive performance in terms of discrimination and calibration as logistic regression models, while offering additional advantages in terms of missing values accommodation and non-linearity learning [270]. Nevertheless, the decision of selecting advanced algorithms over simpler regression-based models for clinical predictive modelling should be sufficiently motivated and documented.

6.3 Challenges associated with DDSS

Beside the DDSS core components such as a qualitative database, calibrated predictive models and interpretable interfaces, additional key challenges remain associated with the development of accurate and useful DDSS.

One of the major obstacles to the adoption of DDSS in clinical practice is the integration of the DDSS within the clinical workflow [14]. Efforts should be engaged to centralize and integrate disease-specific DDSS into the common EHR environment. In addition, DDSS should limit the amount of manual data entry for time-efficiency purposes. For instance, the prediction of a model could be made directly available once all the required variables are accessible in the EHR, without forcing the physicians to look for them and/or to manually re-enter them in the DDSS interface. Additional attention should also be paid in developing intuitive interfaces for DDSS, as efficient algorithms with complex user-interfaces will hardly find their place in clinical practice.

Another important consideration is the reliability of DDSS, as bugs and latency issues are two important factors known to negatively affect the adoption of DDSS by clinicians [12, 14]. DDSS should be designed and intensively tested to produce flawless products. Such practical concerns are especially true with computationally intensive algorithms. Optimizing and securing the DDSS implementation constitute therefore two additional but crucial steps towards the deployment of a DDSS in clinical practice.

In a later stage, the problem of DDSS maintenance should also be discussed. With the prospective collection of new data, additional key questions are posed : *when and how should the internal knowledge of a DDSS be revised and updated?* While lazy learning approaches are easy to update, as no model is generated, it is less clear when a predictive model should be updated to account for the additional batch of data collected [135]. Likewise, if a model is deployed in a slightly different setting than its development environment, it remains uncertain how and to what extent the DDSS should be adapted. To that end, logistic regression models can easily be re-calibrated to better fit the specificities of a local population [271].

While only a few percentage of studies has externally validated the performance of their DDSS approach, even fewer have validated the clinical usefulness in real condition. Even models that have been approved by the regulatory authorities are not necessary helpful in practice [265]. Objectively evaluate, in the vein of randomized clinical trials, the impact of a DDSS on the diagnostic errors, the morbidity/mortality reduction, the time- and cost-effectivness remains challenging and poorly studied [266]. The absence of extensive DDSS evaluations in practice warrants further in-depth studies (see below).

Finally, as briefly mentioned in the introductory section of this thesis, DDSS operate in a sensitive environment where privacy is highly valuable. In the area of decentralized data storing and computing strategies, this constitutes an additional challenge, as DDSS should be developed to preserve the privacy of the patients and its end-users, in particular, in accordance with the GDPR [36].

6.4 Future works

We conclude this work by discussing some important directions for future research.

6.4.1 Additional validation

Although our predictive models for first trimester viability prediction were internally validated, extensive validation of our models should be performed with external data to assess the reproducibility of our findings and study their clinical usefulness in a different setting. Similarly, our reclassification of kidney transplant biopsies would benefit from additional external validation. In this regard, we have provided a simple and freely accessible web-based interface where physicians can upload their own data to compute and visualize the new classification to facilitate external validation [272]. As discussed in the previous section, accurate DDSS are not synonym of clinically useful DDSS. To assess the real impact of DDSS, further studies should focus on objectively evaluating the usefulness of the present DDSS in a clinical environment [266] with appropriate and comprehensive metrics.

We demonstrated that patient-similarity metrics developed on gradient boosted tree proximity matrices constituted effective supervised approaches to retrieve similar patients on incomplete clinical databases. Although this method is particularly suited for clinical databases, its validity on different data types, such as longitudinal and high-dimensional data (e.g. omics data), warrants further investigations.

6.4.2 Interpretability

Although it is true that non-linear methods probably do not bring substantial performance benefit to clinical predictive modeling compared to simpler regression-based models [224, 259], advanced machine learning methods are often discarded for the sake of their poor interpretability. A common claim put forward is that regression-based models are more interpretable. However, no appropriate study exists on the understanding of logistic regression among end-users, especially by physicians users. Before promoting models based on *interpretability* criteria, further investigations should objectively evaluate the understanding of statistical and machine learning approaches, including logistic regression, by clinicians.

Given its strong theoretical background and its well-maintained repository [82], the SHAP framework has rapidly become one of the state-of-the-art methods for post-hoc interpretability. By fairly decomposing an individual prediction among the set of variables, the SHAP framework appears as an appealing choice to improve the interpretability of DDSS. However, to objectively validate its use within DDSS, further studies should assess how such recent methods are perceived and understood by non-expert physicians.

6.4.3 DDSS-physicians interactions

DDSS directly interact with the physicians' clinical decision making. While DDSS aim to improve patient outcome and reduce the misdiagnoses, the impact on physicians reasoning skills (positively or negatively) remains mostly unknown. For instance, accurate DDSS could indirectly decrease the critical appraisal of physicians who blindly rely on the DDSS outcome. Future research should study in greater details the interactions between physicians and DDSS and their implication on clinical practice.

6.4.4 Semi-supervised clustering and diagnostic continuums

Finally, the successful proof of concept of our reclassification system is anticipated to inspire future research far beyond kidney transplant rejection. Similar methodologies could be applied to phenotypic reclassification in other fields of medicine facing comparable limitations in their current classification system.

Semi-supervised clustering, by combining additional meaningful information to guide the clustering process, is expected to produce more clinically useful entities, no solely limited to a single modality (e.g. histological classification). As demonstrated with our intuitive visualization of kidney transplant rejection, there exist, for potentially many medical conditions, continuums of phenotypes and severity. However, for many diseases, the current diagnostic process remains based on simplistic hard classification (*yes/no*) [1]. We have seen with the recent Covid-19 outbreak, being infected by the SARS-CoV-2 virus triggers a whole range of responses, from being completely asymptomatic to life-threatening conditions [275]. To overcome the current limitations of hard-classification and develop more granular diagnostic approaches, future researches should focus on probabilistic extensions of semi-supervised clustering methods.

[1] with the notable exception of the oncology field where precise grading systems, including molecular phenotyping, provide a much finer granularity in the disease assessment [273, 274].

Bibliography

1. *The Cost of Sequencing a Human Genome* https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
2. Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. Are Whole-Exome and Whole-Genome Sequencing Approaches Cost-Effective? A Systematic Review of the Literature. *Genet Med* **20,** 1122–1130. ISSN: 1530-0366 (Oct. 2018).
3. HITInfrastructure. *Organizations See 878% Health Data Growth Rate Since 2016* https://hitinfrastructure.com/news/organizations-see-878-health-data-growth-rate-since-2016. May 2019.
4. Garba, S., Ahmed, A., Mai, A., Makama, G. & Odigie, V. Proliferations of Scientific Medical Journals: A Burden or A Blessing. *Oman Med J* **25,** 311–314. ISSN: 1999-768X (Oct. 2010).
5. *PubMed* https://pubmed.ncbi.nlm.nih.gov/.
6. *Cochrane* https://www.cochrane.org/welcome.
7. Medecine, S. *The Rise of the Data-Driven Physician* tech. rep. (Stanford Medicine, 2020).
8. McCoy, L. G., Nagaraj, S., Morgado, F., Harish, V., Das, S. & Celi, L. A. What Do Medical Students Actually Need to Know about Artificial Intelligence? *npj Digital Medicine* **3,** 1–3. ISSN: 2398-6352 (June 2020).
9. Esteva, A. *et al.* A Guide to Deep Learning in Healthcare. *Nature Medicine* **25,** 24–29. ISSN: 1546-170X (Jan. 2019).
10. Overhage, J. M. & McCallie, D. Physician Time Spent Using the Electronic Health Record During Outpatient Encounters. *Ann Intern Med* **172,** 169–174. ISSN: 0003-4819 (Jan. 2020).
11. Osheroff, J. A. *et al. Improving Outcomes with Clinical Decision Support: An Implementer's Guide, Second Edition* ISBN: 978-1-4987-5746-1 (CRC Press, Feb. 2012).

12. *Clinical Decision Support Systems* (ed Berner, E. S.) ISBN: 978-3-319-31911-7 978-3-319-31913-1 (Springer International Publishing, Cham, 2016).

13. Légat, L., Laere, S. V., Nyssen, M., Steurbaut, S., Dupont, A. G. & Cornu, P. Clinical Decision Support Systems for Drug Allergy Checking: Systematic Review. *Journal of Medical Internet Research* **20,** e8206 (Sept. 2018).

14. Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N. & Kroeker, K. I. An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success. *npj Digital Medicine* **3,** 1–10. ISSN: 2398-6352 (Feb. 2020).

15. Shortliffe, E. & Buchanan, B. A Model of Inexact Reasoning in Medicine (1975).

16. Metzger, J., MacDonald, K., First Consulting Group & California HealthCare Foundation. *Clinical Decision Support for the Independent Physician Practice* ISBN: 978-1-932064-14-8 (California HealthCare Foundation, Oakland, Calif., 2003).

17. Kwok, R., Dinh, M., Dinh, D. & Chu, M. Improving Adherence to Asthma Clinical Guidelines and Discharge Documentation from Emergency Departments: Implementation of a Dynamic and Integrated Electronic Decision Support System. *Emerg Med Australas* **21,** 31–37. ISSN: 1742-6723 (Feb. 2009).

18. Delvaux, N. *et al.* Clinical Decision Support Improves the Appropriateness of Laboratory Test Ordering in Primary Care without Increasing Diagnostic Error: The ELMO Cluster Randomized Trial. *Implementation Science* **15,** 100. ISSN: 1748-5908 (Nov. 2020).

19. Bell, C. M., Jalali, A. & Mensah, E. A Decision Support Tool for Using an ICD-10 Anatomographer to Address Admission Coding Inaccuracies: A Commentary. *Online J Public Health Inform* **5,** 222. ISSN: 1947-2579 (2013).

20. Wald, J. S. *et al.* A Patient-Controlled Journal for an Electronic Medical Record: Issues and Challenges. *Stud Health Technol Inform* **107,** 1166–1170. ISSN: 0926-9630 (2004).

21. Lo-Ciganic, W.-H. *et al.* Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. *JAMA Netw Open* **2,** e190968. ISSN: 2574-3805 (Mar. 2019).

22. Loupy, A. *et al.* Prediction System for Risk of Allograft Loss in Patients Receiving Kidney Transplants: International Derivation and Validation Study. *BMJ* **366.** ISSN: 0959-8138, 1756-1833 (Sept. 2019).

23. Spyridonos, P., Cavouras, D., Ravazoula, P. & Nikiforidis, G. A Computer-Based Diagnostic and Prognostic System for Assessing Urinary Bladder Tumour Grade and Predicting Cancer Recurrence. *Med Inform Internet Med* **27,** 111–122. ISSN: 1463-9238 (June 2002).

24. *AI Explainability Whitepaper* tech. rep. (Google LLC, 2019), 1–28.

25. Dussart, C., Pommier, P., Siranyan, V., Grelaud, G. & Dussart, S. Optimizing Clinical Practice with Case-Based Reasoning Approach: Clinical Practice and Case-Based Reasoning. *Journal of Evaluation in Clinical Practice* **14,** 718–720. ISSN: 13561294 (Oct. 2008).

26. Vapnik, V. *The Nature of Statistical Learning Theory* Second. ISBN: 978-0-387-98780-4 (Springer-Verlag, New York, 2000).

27. Norman, G. Research in Clinical Reasoning: Past History and Current Trends. *Med Educ* **39,** 418–427. ISSN: 0308-0110 (Apr. 2005).

28. Ng, K., Sun, J., Hu, J. & Wang, F. Personalized Predictive Modeling and Risk Factor Identification Using Patient Similarity. *AMIA Jt Summits Transl Sci Proc* **2015,** 132–136. ISSN: 2153-4063 (Mar. 2015).

29. Berner, E. S. Diagnostic Decision Support Systems: Why Aren't They Used More And What Can We Do About It? *AMIA Annu Symp Proc* **2006,** 1167–1168. ISSN: 1942-597X (2006).

30. Loupy, A. *et al.* Prediction System for Risk of Allograft Loss in Patients Receiving Kidney Transplants: International Derivation and Validation Study. *BMJ* **366,** l4923. ISSN: 1756-1833 (Sept. 2019).

31. *ADNEX Risk Model | Iota Group* https://www.iotagroup.org/iota-models-software/adnex-risk-model.

32. Van Calster, B. *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* **349,** g5920–g5920. ISSN: 1756-1833 (Oct. 2014).

33. Kunhimangalam, R., Ovallath, S. & Joseph, P. K. A Clinical Decision Support System with an Integrated EMR for Diagnosis of Peripheral Neuropathy. *J Med Syst* **38,** 38. ISSN: 1573-689X (Apr. 2014).

34. Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W. & Collins, G. S. Predictive Analytics in Health Care: How Can We Know It Works? *J Am Med Inform Assoc.*

35. Jalali, M. S. & Kaiser, J. P. Cybersecurity in Hospitals: A Systematic, Organizational Perspective. *J Med Internet Res* **20,** e10059. ISSN: 1438-8871 (May 2018).

36. *General Data Protection Regulation (GDPR) – Official Legal Text* https://gdpr-info.eu/.

37. Hustinx, P. Privacy by Design: Delivering the Promises. *IDIS* **3,** 253–255. ISSN: 1876-0678 (Aug. 2010).

38. Neamatullah, I. *et al.* Automated De-Identification of Free-Text Medical Records. *BMC Med Inform Decis Mak* **8,** 32. ISSN: 1472-6947 (July 2008).

39. Vaulet, T. *et al. Retrieving Similar Patients from Incomplete Databases with Tree-Based Ensembles*

40. Schacter, D. L. The Seven Sins of Memory. Insights from Psychology and Cognitive Neuroscience. *Am Psychol* **54,** 182–203. ISSN: 0003-066X (Mar. 1999).

41. Expert Panel on effective ways of, investing in Health & Directorate-General for Health and Food Safety (European Commission). *Assessing the Impact of Digital Transformation of Health Services* EU Publication (European Union, Feb. 2019), 108.

42. Bair, E. Semi-Supervised Clustering Methods: Semi-supervised Clustering Methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **5,** 349–361. ISSN: 19395108 (Sept. 2013).

43. Weinberger, K. Q. & Saul, L. K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research* **10,** 207–244 (2009).

44. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. *Statistical Science* **18,** 104–117. ISSN: 0883-4237 (Feb. 2003).

45. Xing, E., Jordan, M., Russell, S. J. & Ng, A. Distance Metric Learning with Application to Clustering with Side-Information. *Advances in Neural Information Processing Systems* **15** (2002).

46. Bellet, A., Habrard, A. & Sebban, M. A Survey on Metric Learning for Feature Vectors and Structured Data. *arXiv:1306.6709 [cs, stat].* arXiv: `1306.6709 [cs, stat]` (June 2013).

47. Burton, A. & Altman, D. G. Missing Covariate Data within Cancer Prognostic Studies: A Review of Current Reporting and Proposed Guidelines. *Br J Cancer* **91,** 4–8. ISSN: 0007-0920 (July 2004).

48. Löw, N., Hesser, J. & Blessing, M. Multiple Retrieval Case-Based Reasoning for Incomplete Datasets. *Journal of Biomedical Informatics* **92,** 103127. ISSN: 1532-0464 (Apr. 2019).

49. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data* ISBN: 978-1-118-59569-5 (John Wiley & Sons, Mar. 2019).

50. Greenland, S. & Finkle, W. D. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *Am J Epidemiol* **142,** 1255–1264. ISSN: 0002-9262 (Dec. 1995).

51. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32. ISSN: 1573-0565 (Oct. 2001).

52. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16,* 785–794. arXiv: 1603.02754 (2016).

53. Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* **15,** 3133–3181. ISSN: 1533-7928 (2014).

54. Lundberg, S. M. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence* **2,** 56–67. ISSN: 2522-5839 (Jan. 2020).

55. Breiman, L. *Manual for Setting Up, Using, And Understanding Random Forest* 2001.

56. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems* **30,** 3146–3154 (2017).

57. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29,** 1189–1232. ISSN: 0090-5364, 2168-8966 (Oct. 2001).

58. Friedman, J. H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* **38,** 367–378. ISSN: 0167-9473 (Feb. 2002).

59. Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A. & Rueckert, D. Random Forest-Based Similarity Measures for Multi-Modal Classification of Alzheimer's Disease. *NeuroImage* **65,** 167–175. ISSN: 1053-8119 (Jan. 2013).

60. Curtis, C. *et al.* The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. *Nature* **486,** 346–352. ISSN: 0028-0836 (Apr. 2012).

61. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov* **2,** 401–404. ISSN: 2159-8290 (May 2012).

62. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci Signal* **6,** pl1. ISSN: 1937-9145 (Apr. 2013).

63. Al-Memar, M. *et al.* Early-Pregnancy Events and Subsequent Antenatal, Delivery and Neonatal Outcomes: Prospective Cohort Study. *Ultrasound in Obstetrics & Gynecology* **54,** 530–537. ISSN: 1469-0705 (2019).

64. Senev, A. *et al.* Specificity, Strength, and Evolution of Pretransplant Donor-Specific HLA Antibodies Determine Outcome after Kidney Transplantation. *American Journal of Transplantation* **19,** 3100–3113. ISSN: 1600-6143 (2019).

65. Kroupa, P., Morton, C., Calvez, K. L. & Williams, M. *Assessing K-Nearest Neighbours Algorithm for Simple, Interpretable Time-to-Event Survival Predictions Over a Range of Simulated Datasets* in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* (June 2019), 367–372.

66. Rahutomo, F., Kitasuka, T. & Aritsugi, M. *Semantic Cosine Similarity* in (Oct. 2012).

67. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the Yield of Medical Tests. *JAMA* **247,** 2543–2546. ISSN: 0098-7484 (May 1982).

68. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems* **24** (2011).

69. van Buuren, S. Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Stat Methods Med Res* **16,** 219–242. ISSN: 0962-2802 (June 2007).

70. Hoogland, J. *et al.* Handling Missing Predictor Values When Validating and Applying a Prediction Model to New Patients. *Statistics in Medicine* **39,** 3591–3607. ISSN: 1097-0258 (2020).

71. Janssen, K. J. M. *et al.* Dealing with Missing Predictor Values When Applying Clinical Prediction Models. *Clinical Chemistry* **55,** 994–1001. ISSN: 0009-9147 (May 2009).

72. Newman, D. A. Missing Data: Five Practical Guidelines. *Organizational Research Methods* **17,** 372–411. ISSN: 1094-4281 (Oct. 2014).

73. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys* ISBN: 978-0-471-65574-9 (John Wiley & Sons, June 2004).

74. Rubin, D. B. Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics* **4,** 87–94. ISSN: 0735-0015 (1986).

75. van Buuren. *Flexible Imputation of Missing Data, Second Edition* ISBN: 978-0-429-96034-5 (CRC Press, July 2018).

76. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Stat Med* **15,** 361–387. ISSN: 0277-6715 (Feb. 1996).

77. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and Comparison of Prognostic Classification Schemes for Survival Data. *Stat Med* **18,** 2529–2545. ISSN: 0277-6715 (Sept. 1999).

78. van Buuren, S. & Groothuis-Oudshoorn, K. Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45,** 1–67 (2011).

79. Van Rossum, G. *Python Tutorial* tech. rep. CS-R9526 (Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995).

80. Sun, J., Sow, D., Hu, J. & Ebadollahi, S. *A System for Mining Temporal Physiological Data Streams for Advanced Prognostic Decision Support* in *2010 IEEE International Conference on Data Mining* (Dec. 2010), 1061–1066.

81. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* ISBN: 978-0-387-84858-7 (Springer Science & Business Media, Aug. 2009).

82. Lundberg, S. M. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions* in *Advances in Neural Information Processing Systems* (2017).

83. Suo, Q. *et al.* Deep Patient Similarity Learning for Personalized Healthcare. *IEEE Transactions on NanoBioscience,* 1–1. ISSN: 1536-1241 (2018).

84. Gonen, M., Alpaydın, E., Tr, B. E. & Tr, B. E. Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research,* 2211–2268 (2011).

85. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat].* arXiv: `1602.04938 [cs, stat]` (Aug. 2016).

86. Shapley, L. S. *A Value for N-Person Games* 307–318. ISBN: 978-1-4008-8197-0 (Princeton University Press, Mar. 2016).

87. Janzing, D., Minorics, L. & Blöbaum, P. Feature Relevance Quantification in Explainable AI: A Causal Problem. *arXiv:1910.13413 [cs, stat].* arXiv: `1910.13413 [cs, stat]` (Nov. 2019).

88. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* **8,** 25. ISSN: 1471-2105 (Jan. 2007).

89. Fisher, A. J., Rudin, C. & Dominici, F. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* (2019).

90. Lundberg, S. M. *et al.* Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery. *Nat Biomed Eng* **2,** 749–760. ISSN: 2157-846X (Oct. 2018).

91. Covert, I., Lundberg, S. & Lee, S.-I. *Understanding Global Feature Contributions With Additive Importance Measures* in *NeurIPS Proceedings* (Oct. 2020). arXiv: 2004.00668.

92. *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set* https://archive.ics.uci.edu/ml/datasets/.

93. Pedregosa, F. *et al.* Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

94. Al-Memar, M. *et al. Prediction of First Trimester Viability Using Simple Clinical and Ultrasound Variables*

95. National Institute for Health and Clinical Excellence. *NICE Clinical Guideline : Ectopic Pregnancy and Miscarriage: Diagnosis and Initial Management* tech. rep. (NICE, Manchester, Apr. 2019).

96. Abdallah, Y. *et al.* Limitations of Current Definitions of Miscarriage Using Mean Gestational Sac Diameter and Crown–Rump Length Measurements: A Multicenter Observational Study. *Ultrasound in Obstetrics & Gynecology* **38,** 497–502. ISSN: 1469-0705 (2011).

97. Chou, B. *The Johns Hopkins Manual of Gynecology and Obstetrics* (June 2020).

98. Magnus, M. C., Wilcox, A. J., Morken, N.-H., Weinberg, C. R. & Håberg, S. E. Role of Maternal Age and Pregnancy History in Risk of Miscarriage: Prospective Register Based Study. *BMJ* **364.** ISSN: 0959-8138, 1756-1833 (Mar. 2019).

99. Foo, L., Johnson, S., Marriott, L., Bourne, T., Bennett, P. & Lees, C. Peri-Implantation Urinary Hormone Monitoring Distinguishes between Types of First-Trimester Spontaneous Pregnancy Loss. *Paediatr Perinat Epidemiol* **34,** 495–503. ISSN: 1365-3016 (Sept. 2020).

100. Rossen, L. M., Ahrens, K. A. & Branum, A. M. Trends in Risk of Pregnancy Loss Among US Women, 1990-2011. *Paediatr Perinat Epidemiol* **32,** 19–29. ISSN: 1365-3016 (Jan. 2018).

101. Geller, P. A., Kerns, D. & Klier, C. M. Anxiety Following Miscarriage and the Subsequent Pregnancy: A Review of the Literature and Future Directions. *Journal of Psychosomatic Research* **56,** 35–45. ISSN: 0022-3999 (Jan. 2004).

102. Farren, J., Mitchell-Jones, N., Verbakel, J. Y., Timmerman, D., Jalmbrant, M. & Bourne, T. The Psychological Impact of Early Pregnancy Loss. *Hum Reprod Update* **24,** 731–749. ISSN: 1460-2369 (Nov. 2018).

103. Farren, J. *et al.* Posttraumatic Stress, Anxiety and Depression Following Miscarriage and Ectopic Pregnancy: A Multicenter, Prospective, Cohort Study. *Am J Obstet Gynecol* **222,** 367.e1–367.e22. ISSN: 1097-6868 (Apr. 2020).

104. Bottomley, C. *et al.* The Optimal Timing of an Ultrasound Scan to Assess the Location and Viability of an Early Pregnancy. *Human Reproduction* **24,** 1811–1817. ISSN: 0268-1161 (Aug. 2009).

105. Li, Q., Zhu, J. & Hua, K. Effects of subchorionic hematoma on pregnancy outcome: a meta analysis. *Zhonghua Yi Xue Za Zhi* **96,** 1383–1385. ISSN: 0376-2491 (May 2016).

106. Bottomley, C., Van Belle, V., Kirk, E., Van Huffel, S., Timmerman, D. & Bourne, T. Accurate Prediction of Pregnancy Viability by Means of a Simple Scoring System. *Human Reproduction* **28,** 68–76. ISSN: 1460-2350, 0268-1161 (Jan. 2013).

107. Guha, S. *et al.* External Validation of Models and Simple Scoring Systems to Predict Miscarriage in Intrauterine Pregnancies of Uncertain Viability. *Hum. Reprod.* **28,** 2905–2911. ISSN: 1460-2350 (Nov. 2013).

108. Elson, J., Salim, R., Tailor, A., Banerjee, S., Zosmer, N. & Jurkovic, D. Prediction of Early Pregnancy Viability in the Absence of an Ultrasonically Detectable Embryo. *Ultrasound Obstet Gynecol* **21,** 57–61. ISSN: 0960-7692 (Jan. 2003).

109. Lautmann, K., Cordina, M., Elson, J., Johns, J., Schramm-Gajraj, K. & Ross, J. A. Clinical Use of a Model to Predict the Viability of Early Intrauterine Pregnancies When No Embryo Is Visible on Ultrasound. *Hum. Reprod.* **26,** 2957–2963. ISSN: 1460-2350 (Nov. 2011).

110. Stamatopoulos, N. *et al.* Prediction of Subsequent Miscarriage Risk in Women Who Present with a Viable Pregnancy at the First Early Pregnancy Scan. *Aust N Z J Obstet Gynaecol* **55,** 464–472. ISSN: 1479-828X (Oct. 2015).

111. Papaioannou, G. I., Syngelaki, A., Maiz, N., Ross, J. A. & Nicolaides, K. H. Ultrasonographic Prediction of Early Miscarriage. *Hum Reprod* **26,** 1685–1692. ISSN: 1460-2350 (July 2011).

112. Oates, J. *et al.* A Prediction Model for Viability at the End of the First Trimester after a Single Early Pregnancy Evaluation. *Aust N Z J Obstet Gynaecol* **53,** 51–57. ISSN: 1479-828X (Feb. 2013).

113. Al-Memar, M. *et al.* Early-Pregnancy Events and Subsequent Antenatal, Delivery and Neonatal Outcomes: Prospective Cohort Study. *Ultrasound in Obstetrics & Gynecology* **54,** 530–537. ISSN: 1469-0705 (2019).

114. Higham, J. M., O'Brien, P. M. & Shaw, R. W. Assessment of Menstrual Blood Loss Using a Pictorial Chart. *Br J Obstet Gynaecol* **97,** 734–739. ISSN: 0306-5456 (Aug. 1990).

115. Hawker, G. A., Mian, S., Kendzerska, T. & French, M. Measures of Adult Pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). *Arthritis Care Res (Hoboken)* **63 Suppl 11,** S240–252. ISSN: 2151-4658 (Nov. 2011).

116. Pillai, R. N., Konje, J. C., Tincello, D. G. & Potdar, N. Role of Serum Biomarkers in the Prediction of Outcome in Women with Threatened Miscarriage: A Systematic Review and Diagnostic Accuracy Meta-Analysis. *Hum Reprod Update* **22,** 228–239. ISSN: 1460-2369 (2016 Mar-Apr).

117. al-Sebai, M. A., Kingsland, C. R., Diver, M., Hipkin, L. & McFadyen, I. R. The Role of a Single Progesterone Measurement in the Diagnosis of Early Pregnancy Failure and the Prognosis of Fetal Viability. *Br J Obstet Gynaecol* **102,** 364–369. ISSN: 0306-5456 (May 1995).

118. Homan, G., Brown, S., Moran, J., Homan, S. & Kerin, J. Human Chorionic Gonadotropin as a Predictor of Outcome in Assisted Reproductive Technology Pregnancies. *Fertility and Sterility* **73,** 270–274. ISSN: 00150282 (Feb. 2000).

119. Albrecht, E. D., Aberdeen, G. W. & Pepe, G. J. The Role of Estrogen in the Maintenance of Primate Pregnancy. *American Journal of Obstetrics & Gynecology* **182,** 432–438. ISSN: 0002-9378, 1097-6868 (Feb. 2000).

120. Fiegler, P., Katz, M., Kaminski, K. & Rudol, G. Clinical Value of a Single Serum CA-125 Level in Women with Symptoms of Imminent Abortion during the First Trimester of Pregnancy. *J Reprod Med* **48,** 982–988. ISSN: 0024-7758 (Dec. 2003).

121. Al Mohamady, M., Fattah, G. A., Elkattan, E., Bayoumy, R. & Hamed, D. A. Correlation of Serum CA-125 and Progesterone Levels with Ultrasound Markers in The Prediction of Pregnancy Outcome in Threatened Miscarriage. *Int J Fertil Steril* **9,** 506–511. ISSN: 2008-076X (2016).

122. Check, J. H., Nowroozi, K., Winkel, C. A., Johnson, T. & Seefried, L. Serum CA 125 Levels in Early Pregnancy and Subsequent Spontaneous Abortion. *Obstet Gynecol* **75,** 742–744. ISSN: 0029-7844 (May 1990).

123. Adeku, M. A., Adegbola, O. & Ajayi, G. O. The Predictive Value of Serum Cancer Antigen 125 (CA125) Levels on Pregnancy Outcome in Threatened Miscarriages. *Journal of Perinatal Medicine* **47,** 704–709. ISSN: 1619-3997 (Sept. 2019).

124. Wald, N. J., Cuckle, H., Brock, J. H., Peto, R., Polani, P. E. & Woodford, F. P. Maternal Serum-Alpha-Fetoprotein Measurement in Antenatal Screening for Anencephaly and Spina Bifida in Early Pregnancy. Report of U.K. Collaborative Study on Alpha-Fetoprotein in Relation to Neural-Tube Defects. *Lancet* **1,** 1323–1332. ISSN: 0140-6736 (June 1977).

125. DiMaio, M. S., Baumgarten, A., Greenstein, R. M., Saal, H. M. & Mahoney, M. J. Screening for Fetal Down's Syndrome in Pregnancy by Measuring Maternal Serum Alpha-Fetoprotein Levels. *New England Journal of Medicine* **317,** 342–346. ISSN: 0028-4793 (Aug. 1987).

126. Johns, J., Hyett, J. & Jauniaux, E. Obstetric Outcome after Threatened Miscarriage with and without a Hematoma on Ultrasound. *Obstetrics & Gynecology* **102,** 483–487. ISSN: 0029-7844 (Sept. 2003).

127. Seppälä, M. & Ruoslahti, E. Alpha-Fetoprotein In Abortion. *The British Medical Journal* **4,** 769–771. ISSN: 0007-1447 (1972).

128. Wald, N., Barker, S., Cuckle, H., Brock, D. J. & Stirrat, G. M. Maternal Serum Alpha-Fetoprotein and Spontaneous Abortion. *Br J Obstet Gynaecol* **84,** 357–362. ISSN: 0306-5456 (May 1977).

129. Harrell, F. E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* ISBN: 978-3-319-19425-7 (Springer, Aug. 2015).

130. Choong, S., Rombauts, L., Ugoni, A. & Meagher, S. Ultrasound Prediction of Risk of Spontaneous Miscarriage in Live Embryos from Assisted Conceptions. *Ultrasound Obstet Gynecol* **22,** 571–577. ISSN: 0960-7692 (Dec. 2003).

131. Steyerberg, E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* ISBN: 978-0-387-77243-1 (Springer-Verlag, New York, 2009).

132. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* ISBN: 978-0-412-04231-7 (CRC Press, May 1994).

133. Tantithamthavorn, C., McIntosh, S., Hassan, A. E. & Matsumoto, K. An Empirical Comparison of Model Validation Techniques for Defect Prediction Models. *IEEE Transactions on Software Engineering* **43,** 1–18. ISSN: 1939-3520 (Jan. 2017).

134. Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M. & Steinley, D. Evidence of Inflated Prediction Performance: A Commentary on Machine Learning and Suicide Research. *Clinical Psychological Science* **9,** 129–134. ISSN: 2167-7026 (Jan. 2021).

135. Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* ISBN: 978-3-030-16399-0 978-3-030-16398-3 (2019).

136. Ambler, G. & Royston, P. Fractional Polynomial Model Selection Procedures: Investigation of Type i Error Rate. *Journal of Statistical Computation and Simulation* **69,** 89–108. ISSN: 0094-9655 (Apr. 2001).

137. Long, Q. & Johnson, B. A. Variable Selection in the Presence of Missing Data: Resampling and Imputation. *Biostatistics* **16,** 596–610. ISSN: 1465-4644, 1468-4357 (July 2015).

138. Brand, J., van Buuren, S., le Cessie, S. & van den Hout, W. Combining Multiple Imputation and Bootstrap in the Analysis of Cost-Effectiveness Trial Data. *Statistics in Medicine* **38,** 210–220. ISSN: 1097-0258 (2019).

139. Van Hoorde, K., Van Huffel, S., Timmerman, D., Bourne, T. & Van Calster, B. A Spline-Based Tool to Assess and Visualize the Calibration of Multiclass Risk Predictions. *Journal of Biomedical Informatics* **54,** 283–293. ISSN: 1532-0464 (Apr. 2015).

140. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A Simple, Step-by-Step Guide to Interpreting Decision Curve Analysis. *Diagnostic and Prognostic Research* **3.** ISSN: 2397-7523 (Dec. 2019).

141. Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E. & Moons, K. G. M. Internal and External Validation of Predictive Models: A Simulation Study of Bias and Precision in Small Samples. *J Clin Epidemiol* **56,** 441–447. ISSN: 0895-4356 (May 2003).

142. Rousson, V. & Zumbrunn, T. Decision Curve Analysis Revisited: Overall Net Benefit, Relationships to ROC Curve Analysis, and Application to Case-Control Studies. *BMC Med Inform Decis Mak* **11,** 45. ISSN: 1472-6947 (June 2011).

143. Vickers, A. J. & Elkin, E. B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making* **26,** 565–574. ISSN: 0272-989X (2006).

144. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2017).

145. Quenby, S. *et al.* Miscarriage Matters: The Epidemiological, Physical, Psychological, and Economic Costs of Early Pregnancy Loss. *Lancet* **397,** 1658–1667. ISSN: 1474-547X (May 2021).

146. Feodor Nilsson, S., Andersen, P., Strandberg-Larsen, K. & Nybo Andersen, A.-M. Risk Factors for Miscarriage from a Prevention Perspective: A Nationwide Follow-up Study. *BJOG: An International Journal of Obstetrics & Gynaecology* **121,** 1375–1385. ISSN: 14700328 (Oct. 2014).

147. Zhou, Y., Li, H., Zhang, Y., Zhang, L., Liu, J. & Liu, J. Association of Maternal Obesity in Early Pregnancy with Adverse Pregnancy Outcomes: A Chinese Prospective Cohort Analysis. *Obesity* **27,** 1030–1036. ISSN: 1930-739X (2019).

148. Pei, L. *et al.* Effect of Periconceptional Folic Acid Supplementation on the Risk of Neural Tube Defects Associated with a Previous Spontaneous Abortion or Maternal First-Trimester Fever. *Human reproduction* (2019).

149. George, L. *et al.* Plasma Folate Levels and Risk of Spontaneous Abortion. *JAMA* **288,** 1867–1873. ISSN: 0098-7484 (Oct. 2002).

150. Gindler, J. *et al.* Folic Acid Supplements during Pregnancy and Risk of Miscarriage. *Lancet* **358,** 796–800. ISSN: 0140-6736 (Sept. 2001).

151. Ahmadi, R., Ziaei, S. & Parsay, S. Association between Nutritional Status with Spontaneous Abortion. *Int J Fertil Steril* **10,** 337–342. ISSN: 2008-076X (2017 Jan-Mar).

152. Gaskins, A. J. *et al.* Maternal Prepregnancy Folate Intake and Risk of Spontaneous Abortion and Stillbirth. *Obstet Gynecol* **124,** 23–31. ISSN: 1873-233X (July 2014).

153. Hu, C.-Y. *et al.* Risk Factors for Spontaneous Abortion from a Prevention Perspective in Rural China: A Population-Based Follow-up Study. *J Matern Fetal Neonatal Med* **34,** 2583–2591. ISSN: 1476-4954 (Aug. 2021).

154. Wald, N., Cuckle, H., Boreham, J., Terzian, E. & Redman, C. The Effect of Maternal Weight on Maternal Serum Alpha-Fetoprotein Levels. *Br J Obstet Gynaecol* **88,** 1094–1096. ISSN: 0306-5456 (Nov. 1981).

155. Farias, D. R. *et al.* Maternal Lipids and Leptin Concentrations Are Associated with Large-for-Gestational-Age Births: A Prospective Cohort Study. *Scientific Reports* **7.** ISSN: 2045-2322 (Dec. 2017).

156. Cook, N. R. Quantifying the Added Value of New Biomarkers: How and How Not. *Diagn Progn Res* **2,** 14. ISSN: 2397-7523 (July 2018).

157. Harrell, F. *Statistically Efficient Ways to Quantify Added Predictive Value of New Measurements* https://www.fharrell.com/post/addvalue/. Oct. 2018.

158. Curry, R. A. & Tempkin, B. B. *Sonography - E-Book: Introduction to Normal Structure and Function* ISBN: 978-0-323-32356-7 (Elsevier Health Sciences, Oct. 2015).

159. Marshall, G., Warner, B., MaWhinney, S. & Hammermeister, K. Prospective Prediction in the Presence of Missing Data. *Stat Med* **21,** 561–570. ISSN: 0277-6715 (Feb. 2002).

160. Lipton, Z. C. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat].* arXiv: `1606.03490 [cs, stat]` (June 2016).

161. Porta, M. *A Dictionary of Epidemiology* ISBN: 978-0-19-939004-5 (Oxford University Press, May 2014).

162. Grimes, D. A. The Nomogram Epidemic: Resurgence of a Medical Relic. *Ann Intern Med* **149,** 273–275. ISSN: 1539-3704 (Aug. 2008).

163. Mukri, F., Bourne, T., Bottomley, C., Schoeb, C., Kirk, E. & Papageorghiou, A. T. Evidence of Early First-Trimester Growth Restriction in Pregnancies That Subsequently End in Miscarriage. *BJOG* **115,** 1273–1278. ISSN: 1471-0528 (Sept. 2008).

164. Bottomley, C. *et al.* Functional Linear Discriminant Analysis: A New Longitudinal Approach to the Assessment of Embryonic Growth. *Hum. Reprod.* **24,** 278–283. ISSN: 1460-2350 (Feb. 2009).

165. de La Rochebrochard, E. & Thonneau, P. Paternal Age and Maternal Age Are Risk Factors for Miscarriage; Results of a Multicentre European Study. *Human Reproduction* **17,** 1649–1656. ISSN: 0268-1161 (June 2002).

166. Vaulet, T. *et al.* Data-Driven Derivation and Validation of Novel Phenotypes for Acute Kidney Transplant Rejection Using Semi-supervised Clustering. *JASN.* ISSN: 1046-6673, 1533-3450 (Mar. 2021).

167. Solez, K. *et al.* International Standardization of Criteria for the Histologic Diagnosis of Renal Allograft Rejection: The Banff Working Classification of Kidney Transplant Pathology. *Kidney Int.* **44,** 411–422. ISSN: 0085-2538 (Aug. 1993).

168. Haas, M. *et al.* The Banff 2017 Kidney Meeting Report: Revised Diagnostic Criteria for Chronic Active T Cell–Mediated Rejection, Antibody-Mediated Rejection, and Prospects for Integrative Endpoints for next-Generation Clinical Trials. *American Journal of Transplantation* **18,** 293–307. ISSN: 1600-6143 (2018).

169. Loupy, A. *et al.* The Banff 2019 Kidney Meeting Report (I): Updates on and Clarification of Criteria for T Cell– and Antibody-Mediated Rejection. *American Journal of Transplantation* **20,** 2318–2331. ISSN: 1600-6143 (2020).

170. Roufosse, C. *et al.* A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology: *Transplantation* **102,** 1795–1814. ISSN: 0041-1337 (Nov. 2018).

171. Racusen, L. C. *et al.* Antibody-Mediated Rejection Criteria - an Addition to the Banff 97 Classification of Renal Allograft Rejection. *Am. J. Transplant.* **3,** 708–714. ISSN: 1600-6135 (June 2003).

172. Haas, M. *et al.* Banff 2013 Meeting Report: Inclusion of C4d-Negative Antibody-Mediated Rejection and Antibody-Associated Arterial Lesions. *Am. J. Transplant.* **14,** 272–283. ISSN: 1600-6143 (Feb. 2014).

173. Loupy, A. *et al.* The Banff 2015 Kidney Meeting Report: Current Challenges in Rejection Classification and Prospects for Adopting Molecular Pathology. *Am. J. Transplant.* **17,** 28–41. ISSN: 1600-6143 (Jan. 2017).

174. Bullinger, L. *et al.* Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. *New England Journal of Medicine* **350,** 1605–1616. ISSN: 0028-4793, 1533-4406 (Apr. 2004).

175. Kickingereder, P. *et al.* Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. *Radiology* **280,** 880–889. ISSN: 0033-8419 (June 2016).

176. Senev, A. *et al.* Histological Picture of Antibody-Mediated Rejection without Donor-Specific Anti-HLA Antibodies: Clinical Presentation and Implications for Outcome. *Am. J. Transplant.* **19,** 763–780. ISSN: 1600-6143 (Mar. 2019).

177. Coemans, M. *et al.* Occurrence of Diabetic Nephropathy After Renal Transplantation Despite Intensive Glycemic Control: An Observational Cohort Study. *Diabetes Care.* ISSN: 0149-5992, 1935-5548 (Feb. 2019).

178. Bair, E. & Tibshirani, R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLOS Biology* **2,** e108. ISSN: 1545-7885 (Apr. 2004).

179. Strehl, A. & Ghosh, J. Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **3,** 583–617. ISSN: 1532-4435 (Mar. 2003).

180. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52,** 91–118. ISSN: 1573-0565 (July 2003).

181. Şenbabaoğlu, Y., Michailidis, G. & Li, J. Z. Critical Limitations of Consensus Clustering in Class Discovery. *Sci Rep* **4.** ISSN: 2045-2322 (Aug. 2014).

182. Royston, P. & Parmar, M. K. Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-Event Outcome. *BMC Medical Research Methodology* **13,** 152. ISSN: 1471-2288 (Dec. 2013).

183. Koenig, A. *et al.* Missing Self Triggers NK Cell-Mediated Chronic Vascular Rejection of Solid Organ Transplants. *Nat Commun* **10,** 1–17. ISSN: 2041-1723 (Nov. 2019).

184. Bestard, O. & Grinyó, J. Refinement of Humoral Rejection Effector Mechanisms to Identify Specific Pathogenic Histological Lesions with Different Graft Outcomes. *American Journal of Transplantation* **19,** 952–953. ISSN: 1600-6143 (2019).

185. Callemeyn, J. *et al.* Transcriptional Changes in Kidney Allografts with Histology of Antibody-Mediated Rejection without Anti-HLA Donor-Specific Antibodies. *J. Am. Soc. Nephrol.* ISSN: 1533-3450 (July 2020).

186. Madill-Thomsen, K. *et al.* Discrepancy Analysis Comparing Molecular and Histology Diagnoses in Kidney Transplant Biopsies. *Am. J. Transplant.* ISSN: 1600-6143 (Dec. 2019).

187. Furness, P. N. *et al.* International Variation in Histologic Grading Is Large, and Persistent Feedback Does Not Improve Reproducibility. *Am. J. Surg. Pathol.* **27,** 805–810. ISSN: 0147-5185 (June 2003).

188. Smith, B. *et al.* A Method to Reduce Variability in Scoring Antibody-Mediated Rejection in Renal Allografts: Implications for Clinical Trials - a Retrospective Study. *Transpl. Int.* **32,** 173–183. ISSN: 1432-2277 (Feb. 2019).

189. Vaulet, T. *et al. Data-Driven Chronic Rejection Phenotypes : A Novel and Validated Complement for the Complete Histological Assessment of Kidney Transplant Biopsies*

190. Sicard, A. *et al.* Computer-Assisted Topological Analysis of Renal Allograft Inflammation Adds to Risk Evaluation at Diagnosis of Humoral Rejection. *Kidney Int.* **92,** 214–226. ISSN: 1523-1755 (July 2017).

191. Fröhlich, H. *et al.* From Hype to Reality: Data Science Enabling Personalized Medicine. *BMC Medicine* **16.** ISSN: 1741-7015 (Dec. 2018).

192. *Blending Structured and Unstructured Data Develop Healthcare Insights* https://www.himss.org/resources/blending-structured-and-unstructured-data-develop-healthcare-insights.

193. Kong, H.-J. Managing Unstructured Big Data in Healthcare System. *Healthc Inform Res* **25,** 1–2. ISSN: 2093-3681 (Jan. 2019).

194. Ledford, H. Millions of Black People Affected by Racial Bias in Health-Care Algorithms. *Nature* **574,** 608–609 (Oct. 2019).

195. *Understanding Robust and Exploratory Data Analysis* 1st edition (eds Hoaglin, D. C., Mosteller, F. & Tukey, J. W.) ISBN: 978-0-471-38491-5 (Wiley-Interscience, New York, Jan. 2000).

196. Cook, R. D. Influential Observations in Linear Regression. *Journal of the American Statistical Association* **74,** 169–174. ISSN: 0162-1459 (1979).

197. Rochim, A. *Chauvenet's and Peirce's Criterion (Literature Review)* tech. rep. (Apr. 2016).

198. Grubbs, F. E. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **11,** 1–21. ISSN: 0040-1706 (Feb. 1969).

199. Barnett, V. & Lewis, T. *Outliers in Statistical Data* 3rd edition. ISBN: 978-0-471-93094-5 (Wiley, Chichester ; New York, Apr. 1994).

200. Wainer, H., Gessaroli, M. & Verdi, M. Visual Revelations - Finding What Is Not There through the Unfortunate Binning of Results: The Mendel Effect. *CHANCE* **19,** 49–52. ISSN: 0933-2480 (Jan. 2006).

201. Finkbeiner, C. Estimation for the Multiple Factor Model When Data Are Missing. *Psychometrika* **44,** 409–420. ISSN: 1860-0980 (Dec. 1979).

202. Parimbelli, E., Marini, S., Sacchi, L. & Bellazzi, R. Patient Similarity for Precision Medicine: A Systematic Review. *Journal of Biomedical Informatics* **83,** 87–96. ISSN: 1532-0464 (July 2018).

203. Brown, S.-A. Patient Similarity: Emerging Concepts in Systems and Precision Medicine. *Frontiers in Physiology* **7.** ISSN: 1664-042X (Nov. 2016).

204. Sharafoddini, A., Dubin, J. A. & Lee, J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Medical Informatics* **5,** e7. ISSN: 2291-9694 (Mar. 2017).

205. Gu, D., Liang, C. & Zhao, H. A Case-Based Reasoning System Based on Weighted Heterogeneous Value Distance Metric for Breast Cancer Diagnosis. *Artificial Intelligence in Medicine* **77,** 31–47. ISSN: 0933-3657 (Mar. 2017).

206. Park, Y.-J., Kim, B.-C. & Chun, S.-H. New Knowledge Extraction Technique Using Probability for Case-Based Reasoning: Application to Medical Diagnosis. *Expert Systems* **23,** 2–20. ISSN: 1468-0394 (2006).

207. Campillo-Gimenez, B., Jouini, W., Bayat, S. & Cuggia, M. Improving Case-Based Reasoning Systems by Combining K-Nearest Neighbour Algorithm with Logistic Regression in the Prediction of Patients' Registration on the Renal Transplant Waiting List. *PLOS ONE* **8,** e71991. ISSN: 1932-6203 (Sept. 2013).

208. Gottlieb, A., Stein, G. Y., Ruppin, E., Altman, R. B. & Sharan, R. A Method for Inferring Medical Diagnoses from Patient Similarities. *BMC Medicine* **11,** 194. ISSN: 1741-7015 (Sept. 2013).

209. Lee, J., Maslove, D. M. & Dubin, J. A. Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric. *PLOS ONE* **10,** e0127428. ISSN: 1932-6203 (May 2015).

210. Panahiazar, M., Taslimitehrani, V., Pereira, N. L. & Pathak, J. Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. *Stud Health Technol Inform* **210,** 369–373. ISSN: 0926-9630 (2015).

211. Chattopadhyay, S., Ray, P., Chen, H. S., Lee, M. B. & Chiang, H. C. *Suicidal Risk Evaluation Using a Similarity-Based Classifier* in *Advanced Data Mining and Applications* (eds Tang, C., Ling, C. X., Zhou, X., Cercone, N. J. & Li, X.) (Springer, Berlin, Heidelberg, 2008), 51–61. ISBN: 978-3-540-88192-6.

212. Domingos, P. A Few Useful Things to Know about Machine Learning. *Commun. ACM* **55,** 78–87. ISSN: 0001-0782 (Oct. 2012).

213. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. *On the Surprising Behavior of Distance Metrics in High Dimensional Space* in *Database Theory — ICDT 2001* (eds Van den Bussche, J. & Vianu, V.) (Springer, Berlin, Heidelberg, 2001), 420–434. ISBN: 978-3-540-44503-6.

214. Boser, B. E., Guyon, I. M. & Vapnik, V. N. *A Training Algorithm for Optimal Margin Classifiers* in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (Association for Computing Machinery, New York, NY, USA, July 1992), 144–152. ISBN: 978-0-89791-497-0.

215. Platt, J. C. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods* in *Advances in Large Margin Classifiers* (MIT Press, 1999), 61–74.

216. Harrell, F. *Classification vs. Prediction* fharrell.com/post/classification/. Jan. 2017.

217. Breiman, L. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statist. Sci.* **16,** 199–231. ISSN: 0883-4237, 2168-8745 (Aug. 2001).

218. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* ISBN: 978-0-262-03561-3 (MIT Press, Nov. 2016).

219. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** 267–288. ISSN: 0035-9246 (1996).

220. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12,** 55–67. ISSN: 0040-1706 (Feb. 1970).

221. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 301–320. ISSN: 1467-9868 (Apr. 2005).

222. Martín Abadi *et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* (2015).

223. Jordan, M. I. & Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **349,** 255–260. ISSN: 0036-8075, 1095-9203 (July 2015).

224. Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* ISBN: 978-3-030-16398-3 (2019).

225. Podgorelec, V., Kokol, P., Stiglic, B. & Rozman, I. Decision Trees: An Overview and Their Use in Medicine. *Journal of Medical Systems,* 19 (2002).

226. Fedorov, V., Mannino, F. & Zhang, R. Consequences of Dichotomization. *Pharm Stat* **8,** 50–61. ISSN: 1539-1612 (2009 Jan-Mar).

227. Hand, D. J. & Yu, K. Idiot's Bayes: Not So Stupid after All? *International Statistical Review / Revue Internationale de Statistique* **69,** 385–398. ISSN: 0306-7734 (2001).

228. Schank, R. C. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People* ISBN: 978-0-521-27029-8 (Cambridge University Press, Cambridge Cambridgeshire : New York, Jan. 1983).

229. Vapnik, V. *Estimation of Dependences Based on Empirical Data* ISBN: 978-0-387-30865-4 (Springer-Verlag, New York, 2006).

230. Bichindaritz, I. Case-Based Reasoning in the Health Sciences. *Artificial Intelligence in Medicine* **36,** 121–125. ISSN: 09333657 (Feb. 2006).

231. Buske, O. J. *et al.* PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. *Human Mutation* **36,** 931–940. ISSN: 10597794 (Oct. 2015).

232. He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X. & Zhang, K. The Practical Implementation of Artificial Intelligence Technologies in Medicine. *Nature Medicine* **25,** 30–36. ISSN: 1546-170X (Jan. 2019).

233. Towards Trustable Machine Learning. *Nat Biomed Eng* **2,** 709–710. ISSN: 2157-846X (Oct. 2018).

234. Khairat, S., Marc, D., Crosby, W. & Al Sanousi, A. Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. *JMIR Med Inform* **6,** e24. ISSN: 2291-9694 (Apr. 2018).

235. Došilović, F. K., Brčić, M. & Hlupić, N. *Explainable Artificial Intelligence: A Survey* in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (May 2018), 0210–0215.

236. Doshi-Velez, F. & Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*. arXiv: 1702.08608 [cs, stat] (Mar. 2017).

237. Cotton, C. & Or, B. Explanation and Justification in Machine Learning: A Survey, 6.

238. Lisboa, P. J. G. *Interpretability in Machine Learning – Principles and Practice* in *Fuzzy Logic and Applications* (eds Masulli, F., Pasi, G. & Yager, R.) (Springer International Publishing, Cham, 2013), 15–21. ISBN: 978-3-319-03200-9.

239. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* **267,** 1–38. ISSN: 0004-3702 (Feb. 2019).

240. Lombrozo, T. The Structure and Function of Explanations. *Trends Cogn Sci* **10,** 464–470. ISSN: 1364-6613 (Oct. 2006).

241. Lipton, Z. C., Kale, D. C. & Wetzel, R. Modeling Missing Data in Clinical Time Series with RNNs. *arXiv:1606.04130 [cs, stat]*. arXiv: 1606.04130 [cs, stat] (June 2016).

242. Mittelstadt, B., Russell, C. & Wachter, S. *Explaining Explanations in AI* in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, Atlanta GA USA, Jan. 2019), 279–288. ISBN: 978-1-4503-6125-5.

243. Molnar, C. *Interpretable Machine Learning* ().

244. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* **1,** 206. ISSN: 2522-5839 (May 2019).

245. Van Belle, V. M. C. A. *et al.* A Mathematical Model for Interpretable Clinical Decision Support with Applications in Gynecology. *PLoS ONE* **7** (ed Brusic, V.) e34312. ISSN: 1932-6203 (Mar. 2012).

246. Montavon, G., Samek, W. & Müller, K.-R. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing* **73,** 1–15. ISSN: 1051-2004 (Feb. 2018).

247. *"explain"* https://www.merriam-webster.com/dictionary/explain. 2021.

248. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]*. arXiv: 1706.07269 [cs] (Aug. 2018).

249. Selbst, A. D. & Powles, J. Meaningful Information and the Right to Explanation. *International Data Privacy Law* **7,** 233–242. ISSN: 2044-3994 (Nov. 2017).

250. Miller, G. A. The Magical Number Seven, plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review* **63,** 81–97. ISSN: 1939-1471(Electronic),0033-295X(Print) (1956).

251. Cowan, N. The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Curr Dir Psychol Sci* **19,** 51–57. ISSN: 0963-7214 (Feb. 2010).

252. Hilton, D. J. Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance. *Thinking & Reasoning* **2,** 273–308. ISSN: 1354-6783 (Nov. 1996).

253. *Abductive Inference: Computation, Philosophy, Technology* (eds Josephson, J. R. & Josephson, S. G.) ISBN: 978-0-521-43461-4 (Cambridge University Press, Cambridge, 1994).

254. Harman, G. H. The Inference to the Best Explanation. *Philosophical Review* **74,** 88–95 (1965).

255. Popper, K. *The Logic of Scientific Discovery* 2nd edition. ISBN: 978-0-415-27844-7 (Routledge, London, Mar. 2002).

256. Croskerry, P. A Universal Model of Diagnostic Reasoning. *Acad Med* **84,** 1022–1028. ISSN: 1938-808X (Aug. 2009).

257. Lynch, C. J. & Liston, C. New Machine-Learning Technologies for Computer-Aided Diagnosis. *Nat Med* **24,** 1304–1305. ISSN: 1546-170X (Sept. 2018).

258. Esteva, A. *et al.* Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **542,** 115–118. ISSN: 1476-4687 (Feb. 2017).

259. Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y. & Calster, B. V. A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology* **110,** 12–22. ISSN: 0895-4356, 1878-5921 (June 2019).

260. Schwarzer, G., Vach, W. & Schumacher, M. On the Misuses of Artificial Neural Networks for Prognostic and Diagnostic Classification in Oncology. *Stat Med* **19,** 541–561. ISSN: 0277-6715 (Feb. 2000).

261. Harrell, F. *Is Medicine Mesmerized by Machine Learning?* fharrell.com/post/medml. Feb. 2018.

262. Panch, T., Mattie, H. & Celi, L. A. The "Inconvenient Truth" about AI in Healthcare. *npj Digit. Med.* **2,** 1–3. ISSN: 2398-6352 (Aug. 2019).

263. Toh, T. S., Dondelinger, F. & Wang, D. Looking beyond the Hype: Applied AI and Machine Learning in Translational Medicine. *EBioMedicine* **47,** 607–615. ISSN: 2352-3964 (Sept. 2019).

264. Fröhlich, H. *et al.* From Hype to Reality: Data Science Enabling Personalized Medicine. *BMC Medicine* **16,** 150. ISSN: 1741-7015 (Aug. 2018).

265. Park, S. H. Regulatory Approval versus Clinical Validation of Artificial Intelligence Diagnostic Tools. *Radiology* **288,** 910–911. ISSN: 0033-8419 (Sept. 2018).

266. Shah, N. H., Milstein, A. & Steven C. Bagley, P. Making Machine Learning Models Clinically Useful. *JAMA* (Aug. 2019).

267. Reps, J. M., Schuemie, M. J., Suchard, M. A., Ryan, P. B. & Rijnbeek, P. R. Design and Implementation of a Standardized Framework to Generate and Evaluate Patient-Level Prediction Models Using Observational Healthcare Data. *J Am Med Inform Assoc* **25,** 969–975. ISSN: 1527-974X (Aug. 2018).

268. Vayena, E., Blasimme, A. & Cohen, I. G. Machine Learning in Medicine: Addressing Ethical Challenges. *PLOS Medicine* **15,** e1002689. ISSN: 1549-1676 (Nov. 2018).

269. Char, D. S., Shah, N. H. & Magnus, D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N Engl J Med* **378,** 981–983. ISSN: 0028-4793 (Mar. 2018).

270. Vaulet, T. *et al.* Gradient Boosted Trees with Individual Explanations : An Alternative to Logistic Regression for Viability Prediction in the First Trimester of Pregnancy. *Computer Methods and Programs in Biomedicine,* 106520. ISSN: 0169-2607 (Nov. 2021).

271. Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J. & Steyerberg, E. W. A Calibration Hierarchy for Risk Models Was Defined: From Utopia to Empirical Data. *Journal of Clinical Epidemiology* **74,** 167–176. ISSN: 0895-4356 (June 2016).

272. *RejectClass* https://rejectclass.pythonanywhere.com/. 2021.

273. Sobin, L. H., Gospodarowicz, M. K. & Wittekind, C. *TNM Classification of Malignant Tumours* ISBN: 978-1-4443-5896-4 (John Wiley & Sons, Aug. 2011).

274. *Molecular Diagnosis of Cancer: Methods and Protocols* Second (eds Roulston, J. E. & Bartlett, J. M. S.) ISBN: 978-1-58829-160-8 (Humana Press, 2004).

275. Huang, C. *et al.* Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *The Lancet* **395,** 497–506. ISSN: 0140-6736, 1474-547X (Feb. 2020).

276. Appiah, V., Taylor, S., Vaulet, T., Howarth, N., Gevenois, P. A. & Tack, D. Are Referral Guidelines for CT Examinations Addressing All Clinical Scenarios? A Comparison of EURO-2000 Guidelines and ESR iGuide. *Eur Radiol.* ISSN: 1432-1084 (Feb. 2021).

277. Al-Memar, M. *et al.* First-Trimester Intrauterine Hematoma and Pregnancy Complications. *Ultrasound Obstet Gynecol* **55,** 536–545. ISSN: 1469-0705 (Apr. 2020).

278. Vaulet, T., Thaunat, O. & Naesens, M. Authors' Reply. *JASN* **32,** 2388–2389. ISSN: 1046-6673, 1533-3450 (Sept. 2021).

279. Fourie, H. *et al. The Ability of Plasma Proteins to Predict Pregnancy Outcome in the First Trimester* in *Annual Academic Meeting of the Royal College of Obstetricians and Gynaecologists (RCOG)* (London, United Kingdom, Feb. 2021).

280. Vaulet, T. *et al.* Building Gestational Age Independent Features for First Trimester Viability Prediction. *Ultrasound in Obstetrics & Gynecology* **56,** 27–27. ISSN: 1469-0705 (2020).

281. Al-Memar, M. *et al.* OC16.07: Early Pregnancy Biomarkers for the Prediction of First Trimester Viability. *Ultrasound in Obstetrics & Gynecology* **54,** 41–41. ISSN: 1469-0705 (2019).

282. Vaulet, T., De Moor, B. & Naesens, M. *Modelling Kidney Transplant Biopsies with a Mixture of Independent Poisson Distributions* in *Annual IEEE EMBS Benelux Chapter Symposium - Artificial Intelligence in Healthcare* (Leuven, Nov. 2019).

283. Vaulet, T., Lerut, E., Sprangers, B., Kuypers, D., De Moor, B. & Naesens, M. *Reclassification of Kidney Allograft Biopsies Using Semi-Supervised Machine Learning* in *19th Congress of the European Society for Organ Transplantation (ESOT)* (Copenhagen, Denmark, Sept. 2019).

284. Fourie, H. *et al.* P14.07: First Trimester Alpha-Fetoprotein Levels and Ultrasound Measurements of the Yolk Sac in Women Who Miscarry Compared to Those That Continue to a Term Pregnancy. *Ultrasound in Obstetrics & Gynecology* **54,** 200–200. ISSN: 1469-0705 (2019).

285. Al-Memar, M. *et al.* EP03.07: Early Pregnancy Biomarkers and Prediction of Hypertensive Disorders of Pregnancy. *Ultrasound in Obstetrics & Gynecology* **54,** 247–247. ISSN: 1469-0705 (2019).

286. Al-Memar, M. *et al.* OC19.07: Impact of Intrauterine Hematoma on Pregnancy Outcomes. *Ultrasound in Obstetrics & Gynecology* **52,** 46–47. ISSN: 1469-0705 (2018).

287. Al-Memar, M. *et al. Pain and Bleeding in the First Trimester and Long-Term Pregnancy Outcomes: A Prospective Cohort Study* in (WILEY, 2019), 18–18.

288. Al-Memar, M. *et al. Impact of Intrauterine Haematoma on Pregnancy Outcomes.* in (SAGE PUBLICATIONS INC, 2019), 378A–378A.

List of publications

International journal papers (published)

Vaulet, T., Al-Memar, M., Fourie, H., *et al.* Gradient Boosted Trees with Individual Explanations : An Alternative to Logistic Regression for Viability Prediction in the First Trimester of Pregnancy. *Computer Methods and Programs in Biomedicine,* 106520. ISSN: 0169-2607 (Nov. 2021)

Vaulet, T., Divard, G., Thaunat, O., *et al.* Data-Driven Derivation and Validation of Novel Phenotypes for Acute Kidney Transplant Rejection Using Semi-supervised Clustering. *JASN.* ISSN: 1046-6673, 1533-3450 (Mar. 2021)

Appiah, V., Taylor, S., Vaulet, T., Howarth, N., Gevenois, P. A. & Tack, D. Are Referral Guidelines for CT Examinations Addressing All Clinical Scenarios? A Comparison of EURO-2000 Guidelines and ESR iGuide. *Eur Radiol.* ISSN: 1432-1084 (Feb. 2021)

Al-Memar, M., Vaulet, T., Fourie, H., *et al.* Early-Pregnancy Events and Subsequent Antenatal, Delivery and Neonatal Outcomes: Prospective Cohort Study. *Ultrasound in Obstetrics & Gynecology* **54,** 530–537. ISSN: 1469-0705 (2019)

Al-Memar, M., Vaulet, T., Fourie, H., *et al.* First-Trimester Intrauterine Hematoma and Pregnancy Complications. *Ultrasound Obstet Gynecol* **55,** 536–545. ISSN: 1469-0705 (Apr. 2020)

International journal papers (submitted)

Al-Memar, M., Fourie, H., Vaulet, T., *et al. Prediction of First Trimester Viability Using Simple Clinical and Ultrasound Variables*, submitted to the

American Journal of Obstetrics and Gynecology.

Vaulet, T., Al-Memar, M., Fourie, H., *et al. Retrieving Similar Patients from Incomplete Databases with Tree-Based Ensembles*, submitted to the *Journal of Biomedical Informatics.*

International journal papers (to be submitted)

Vaulet, T., Aubert, O., Callemeyn, J., *et al. Data-Driven Chronic Rejection Phenotypes : A Novel and Validated Complement for the Complete Histological Assessment of Kidney Transplant Biopsies* to be submitted to the *American Journal of the Society of Nephrology*

Letters and replies in international journals

Vaulet, T., Thaunat, O. & Naesens, M. Authors' Reply. *JASN* **32,** 2388–2389. ISSN: 1046-6673, 1533-3450 (Sept. 2021)

International conferences abstracts and presentations

Fourie, H., Al-Memar, M., Vaulet, T., *et al. The Ability of Plasma Proteins to Predict Pregnancy Outcome in the First Trimester* in *Annual Academic Meeting of the Royal College of Obstetricians and Gynaecologists (RCOG)* (London, United Kingdom, Feb. 2021)

Vaulet, T., Al-Memar, M., Fourie, H., *et al.* Building Gestational Age Independent Features for First Trimester Viability Prediction. *Ultrasound in Obstetrics & Gynecology* **56,** 27–27. ISSN: 1469-0705 (2020)

Al-Memar, M., Fourie, H., Vaulet, T., *et al.* OC16.07: Early Pregnancy Biomarkers for the Prediction of First Trimester Viability. *Ultrasound in Obstetrics & Gynecology* **54,** 41–41. ISSN: 1469-0705 (2019)

Vaulet, T., De Moor, B. & Naesens, M. *Modelling Kidney Transplant Biopsies with a Mixture of Independent Poisson Distributions* in *Annual IEEE EMBS Benelux Chapter Symposium - Artificial Intelligence in Healthcare* (Leuven, Nov. 2019)

Vaulet, T., Lerut, E., Sprangers, B., Kuypers, D., De Moor, B. & Naesens, M. *Reclassification of Kidney Allograft Biopsies Using Semi-Supervised Machine Learning* in *19th Congress of the European Society for Organ Transplantation (ESOT)* (Copenhagen, Denmark, Sept. 2019)

Fourie, H., Al-Memar, M., Vaulet, T., *et al.* P14.07: First Trimester Alpha-Fetoprotein Levels and Ultrasound Measurements of the Yolk Sac in Women Who Miscarry Compared to Those That Continue to a Term Pregnancy. *Ultrasound in Obstetrics & Gynecology* **54,** 200–200. ISSN: 1469-0705 (2019)

Al-Memar, M., Fourie, H., Vaulet, T., *et al.* EP03.07: Early Pregnancy Biomarkers and Prediction of Hypertensive Disorders of Pregnancy. *Ultrasound in Obstetrics & Gynecology* **54,** 247–247. ISSN: 1469-0705 (2019)

Al-Memar, M., Vaulet, T., Fourie, H., *et al.* OC19.07: Impact of Intrauterine Hematoma on Pregnancy Outcomes. *Ultrasound in Obstetrics & Gynecology* **52,** 46–47. ISSN: 1469-0705 (2018)

Al-Memar, M., Vaulet, T., Nikolic, G., *et al. Pain and Bleeding in the First Trimester and Long-Term Pregnancy Outcomes: A Prospective Cohort Study* in (WILEY, 2019), 18–18

Al-Memar, M., Vaulet, T., Fourie, H., *et al. Impact of Intrauterine Haematoma on Pregnancy Outcomes.* in (SAGE PUBLICATIONS INC, 2019), 378A–378A



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
ESAT-STADIUS
Celestijnenlaan 200A box 2402
B-3001 Leuven
thibaut.vaulet@esat.kuleuven.be