

Subnational Analysis and Modelling of the Ebola
Epidemic in West Africa, 2013-2016: application of
machine learning algorithms for case fatality
imputation

Alpha Forna

Imperial College London

School of Public Health

Department of Infectious Disease Epidemiology

Thesis submitted for PhD examination

2020

Declaration of Originality

I declare this work to be my own under the supervision of Prof. Christl A. Donnelly, Dr Pierre Nouvellet and Dr Ilaria Dorigatti. All participants involved in this thesis have been acknowledge in relevant sections of this thesis.

Alpha Forna

March 2020

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Acknowledgements

I would like to thank my supervisors, Prof Christl Ann Donnelly, Dr Pierre Nouvellet, and Dr Ilaria Dorigatti for their untiring guidance and support throughout my PhD. Their incredible teaching and mentorship have helped tremendously to develop my statistical and research skills for which I am eternally grateful.

The Department of Infectious Disease Epidemiology has been a wonderful environment to learn and work in and I am grateful to everyone who has taken time out of their busy schedule to help and advise me throughout. I would like to thank my UG2 office colleagues past and present who have listened with endless patience to each new problem I have encountered. Lucia, my reliable 'call a friend' for all R issues. Obi, with whom I never felt I was doing anything for the first time. Helen, whose sweets from conference trips I would dearly miss. Paul, with whom I always share the disappointment of England losing football matches. Tom, whose free trips to the Zoo and the VIP treatment they came with, always left me and my family feeling like 'connected Londoners'. Thank you.

I am so fortunate to be part of the Donnelly Group—once a Donnelly always a Donnelly, we are now made up of PhD students and Postdocs at both Imperial College London and the University of Oxford. Group members have also moved on to other exciting opportunities in academia and industry. Thanks, mates, for all the support, friendship, food and laughter.

I would like to thank the World Health Organisation (WHO) for kindly allowing me to use the comprehensive anonymised case line list from the West African Ebola Epidemic throughout my PhD and in the publications that came out of this project.

I would like to thank the Commonwealth Scholarship Commission for funding my studentship and for all the opportunities they have provided me.

Finally, but most importantly, I thank Swaray, Alvetta and the rest of my family, particularly my Papa, Mr Alie B. Forna who have never been less than 100% supportive not just during this PhD but through all the challenges life has thrown at me. This PhD is as much your effort as it is mine.

Dedication

This project is dedicated to my late Mama, Mrs. Lovetta Christiana Forna who passed away on April 30, 2012. She always instilled in me the importance of trusting God and her favourite quote was “It is well”. In her memory I will always strive to make my good even better and transform my better into my best. I would also like to acknowledge all the lives that were lost during the West African Ebola epidemic of 2013-2016. May all your souls rest in perfect peace.

It is well!

Relevant Publications

1. **Forna A**, Nouvellet P, Dorigatti I, Donnelly CA. Case fatality ratio estimates for the 2013 – 2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation. *Clinical Infectious Diseases* 2019. <https://doi.org/10.1093/cid/ciz678> [Published]

Attributable content to Alpha Forna: Study design, data analysis, interpretation of the results and writing of the manuscript.

2. **Forna A**, Dorigatti I, Nouvellet P, Donnelly CA. Spatiotemporal Variability in Case Fatality Ratios for 2013–2016 Ebola epidemic in West Africa. *International Journal of Infectious Diseases* 2020. <https://doi.org/10.1016/j.ijid.2020.01.046> [Published]

Attributable content to Alpha Forna: Study design, data analysis, interpretation of the results and writing of the manuscript.

3. **Forna A**, Dorigatti I, Nouvellet P, Donnelly CA. Comparison of machine learning inference methods for estimating case fatality ratios: An Ebola outbreak simulation study. [In preparation]

Attributable content to Alpha Forna: Study design, data analysis, interpretation of the results and writing of the manuscript.

Presentations

1. Poster presentation: ‘Imputing survival using a Boosted Regression Tree (BRT) model: an analysis of the West African Ebola epidemic data’. (2017) *Imperial College London School of Public Health Annual Symposium*.

2. Oral presentation: 'Case fatality ratio estimates for the 2013-2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation'. (2018) *Imperial College London School of Public Health Annual Symposium*.
3. Poster presentation: 'Case fatality ratio estimates for the 2013 - 2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation'. (2018) *International Meeting on Emerging Diseases and Surveillance (IMED)*, Vienna, Austria.

Abstract

Background

The 2013-2016 West African Ebola epidemic has been the largest to date with more than 11,000 deaths in the affected countries. The data collected have provided more insight into the case fatality ratio (CFR) and how it varies with age and other characteristics. However, the accuracy and precision of the naïve CFR remain limited because 44% of survival outcomes were unreported.

Methods

Using a machine learning (MaLe) model, Boosted Regression Tree (BRT), I imputed survival outcomes (i.e. survival or death) when unreported, corrected for model imperfection to estimate the CFR without imputation, with imputation and adjusted with imputation. I used semivariogram analysis and kriging to investigate subnational heterogeneities in CFR estimates. I used simulations to evaluate the performance of various MaLe inference methods for the estimation of CFR under different outbreak data scenarios.

Results

The adjusted CFR estimates were 82.8% (95% CI 45.6-85.6%) overall and 89.1% (95% CI 40.8%-91.6%), 65.6% (95% CI 61.3%-69.6%) and 79.2% (95% CI 45.4%-84.1%) for Sierra Leone, Guinea and Liberia, respectively. BRT modelling accounted for most of the spatiotemporal variation and interactions in CFR, but moderate spatial autocorrelation remained. Combining district-level CFR estimates and kriged district-level residuals provided the best linear unbiased map of CFR. Temporal autocorrelation was not observed in the

district-level residuals from the BRT estimates. Finally, I observed that the performance of MaLe inference methods for CFR imputation varies under different outbreak data scenarios.

Conclusions

Adjusted CFR estimates improved the naïve CFR estimates obtained without imputation and were more representative. Used in conjunction with other resources, adjusted CFR estimates and the unbiased CFR maps will inform future public health response to Ebola outbreaks. I confirm that, across the board, data imputation with adjustment for the sensitivity and specificity of MaLe inference methods reduces the bias in CFR estimates.

Contents

Declaration of Originality.....	2
Copyright Declaration.....	3
Acknowledgements.....	4
Dedication.....	6
Relevant Publications	7
Abstract	9
Contents.....	11
List of figures.....	16
List of tables	25
List of abbreviations and acronyms.....	28
Chapter 1: Introduction	30
1.1 Background.....	30
1.1.1 Ebola Virus Disease Epidemiology.....	34
1.1.2 Detection and diagnosis of Ebola.....	36
1.1.3 Case fatality ratio of Ebola.....	37
1.2 Ebola: from outbreak to epidemic	39
1.2.1 Global Policy on Ebola	41
1.2.2 Treatment Options for Ebola	42
1.2.3 Ebola Vaccines and Clinical trial designs	44
1.2.4 Prevention, surveillance and control	45
1.3 Machine learning and Geostatistical methods	46
1.3.1 Machine learning in Infectious disease modelling	47
1.3.2 Geostatistics in infectious disease modelling	48
1.3.3 Characterising the spatiotemporal epidemiology of Ebola	49
1.4 Justification.....	49
1.4.1 Thesis aims and objectives	50
1.5 Thesis Outline	50
Chapter 2: Understanding the 2013–2016 West Africa Ebola Dataset	53
2.1 Overview	53
2.2. Background.....	53
2.3 Geographical description of West Africa	55
2.4 Case Classification System used in West Africa.....	56
2.5 Identification of Predictors from the Data	58
2.5.1 Survival outcome Variable.....	58

2.5.2 Potential Predictors.....	58
2.6 Missingness in the West African Ebola dataset.....	60
2.6.1 Comparison of age and reporting delay between cases with known and unknown outcome	61
2.6.2 Proportion of predictors with known and unknown survival outcomes.....	63
2.7 Discussion.....	73
Chapter 3: Methods and algorithms for Ebola case fatality ratio estimation, geospatial analysis and simulation study.....	75
3.1 Overview	75
3.2 Conceptual description of BRT	76
3.3 Selection of BRT threshold	77
3.3.1 Optimisation of Prevalence	77
3.3.2 Maximisation Percentage Correctly classified.....	77
3.3.3 Maximisation of Specificity and Sensitivity	78
3.3.4 Equal Sensitivity and Specificity.....	78
3.4 Measure of BRT performance	78
3.4.1 Area Under the ROC Curve	79
3.4.2 Sensitivity	79
3.4.3 Specificity.....	79
3.4.4 Percentage Correctly Classified PCC	79
3.5 K-fold cross validation.....	79
3.6 Description of BRT hyperparameters used to optimize performance.....	80
3.6.1 Tree Complexity (<i>tc</i>)	81
3.6.2 Learning Rate (<i>lr</i>).....	81
3.6.3 Bag fraction (<i>bf</i>)	81
3.6.4 Data Partitioning Ratio	82
3.7 Non-parametric bootstrapping.....	82
3.8 Algorithm implemented for hyperparameter optimisation.....	84
3.9 Algorithm implemented for CFR with imputation estimation	85
3.10 Algorithm implemented for CFR adjusted for BRT model bias	87
3.11 Algorithm implemented to validate sensitivity/specificity adjustment.....	88
3.12 Algorithm implemented for spatial CFR prediction and residual estimation.....	90
3.13 Algorithm implemented for temporal CFR prediction and residual estimation	91
3.14 Algorithms for data simulation.....	91
3.14.1 Generating Missingness	92

3.15 Algorithm implemented to investigate the performance of models as a function of outbreak data characteristics.....	93
3.16 Algorithm implemented for CFR adjusted for bias in the models	95
Chapter 4: Case fatality ratio estimates for the 2013 – 2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation	96
4.1 Introduction.....	96
4.2 Methods	98
4.2.1 Data sources	98
4.2.2 Statistical methods.....	99
4.3 Results.....	103
4.4 Discussion.....	111
4.5 Additional Results	115
4.5.1 Identification of hyperparameters.....	115
4.5.2 Comparison of CFR estimates obtained on the training set, without and with imputation	117
4.5.3 Comparison of CFR estimates obtained on the validation data, with and without imputation.....	124
4.5.4 CFR adjusted with imputation by gender and age (categories).....	131
4.5.5 CFR estimates without, unadjusted and adjusted with imputation for other predictors in the simplified model.....	133
4.5.6 CFR by fever occurrence and date of case reporting.....	137
4.5.7 Robustness of BRT imputation to data down-sampling.....	138
4.5.8 Sampling to explore simulated missingness.....	140
4.6 Sensitivity Analysis.....	141
4.6.1 Comparison of known and unknown outcome data for ‘confirmed’ and ‘confirmed and probable’ cases.....	141
4.6.2 BRT model parameterisation for confirmed case data	142
4.6.3 Minimal model predictors for variation in case classification	145
4.6.4 CFRs for variations in case classification	147
4.6.5 BRT imputation performance for variations in case classification.....	149
4.6.6 CFR estimates by age group, delay, country and fever for other case classifications	149
Chapter 5: Spatiotemporal variability in Case Fatality Ratios for 2013–2016 Ebola epidemic in West Africa.....	153
5.1 Introduction.....	153
5.2 Methods	154
5.2.1 Geographical characteristics of the study setting.....	154
5.2.2 Data sources	155

5.2.3 Spatiotemporal Analysis.....	155
5.3 Results.....	159
5.4 Discussion.....	167
5.5 Additional Results.....	171
5.5.1 Moran's I indices for CFR residuals.....	171
5.5.2 Selecting a spatial model for the analysis.....	175
5.5.3 Spatial semivariogram for observed CFR residuals.....	177
5.5.4 Temporal semivariogram for observed CFR residuals.....	178
5.5.5 Complete map of predicted CFR.....	179
5.5.6 Directional semivariograms.....	180
5.5.7 Comparing Gaussian and exponential models in mapping CFR.....	181
5.5.8 Number of cases and number of known deaths per district.....	183
Chapter 6: Comparison of machine learning inference methods for estimating case fatality ratios: An Ebola outbreak simulation study	185
6.1 Introduction.....	185
6.2 Methods.....	186
6.2.1 Simulated Data.....	186
6.2.2 Logistic regression.....	188
6.2.3 Random Forest.....	188
6.2.4 Boosted Regression Trees.....	189
6.2.5 Bayesian Additive Regression Trees.....	189
6.2.6 Artificial Neural Networks.....	190
6.2.7 Outbreak data characteristics investigated for this study.....	194
6.2.8 Experimental Setup.....	195
6.3 Results.....	197
6.4 Discussion.....	208
6.5 Additional results.....	211
6.5.1 Survival outcome, demographic and clinical predictors considered for this simulation study.	211
6.6 Sensitivity analysis.....	215
6.6.1 Comparing the unadjusted and adjusted CFR estimates for MCAR missingness in survival outcomes.....	215
6.6.2 Comparing the unadjusted and adjusted CFR estimates for MNAR missingness in survival outcomes.....	216
Chapter 7: Discussion of findings: Subnational Analysis and Modelling of the Ebola Epidemic in West Africa, 2013-2016	218
7.1 Summary of Findings.....	218

7.2 Future Work and Limitations.....	221
7.3 Implications of Research	224
7.4 Conclusions.....	227
8 References	228

List of figures

Figure 1. 1: Geographic distribution of Ebola in Africa. Reproduced from [5].	31
Figure 1. 2: A microscopic structure of the Ebola virus: Reproduced from [16].	33
Figure 1. 3: Clinical features of the 2013–2016 Ebola virus disease epidemic in West Africa. Reproduced from [18].	36
Figure 1. 4: Schematic summary of the chapters in this thesis.	52
Figure 2. 1: The geographical location and distribution of administrative districts in Sierra Leone, Guinea and Liberia.	56
Figure 4. 1: Schematic summary of the analysis steps used in this chapter.	100
Figure 4. 2: Proportion of known survival outcomes (i.e. Dead and Alive) and unknown survival outcomes (i.e. Missing) for ‘confirmed, probable and suspected’ cases. a: Proportion of deaths, survivals, and entries with unknown outcome by age group (in years). b: Proportion of deaths, survivals, and entries with unknown outcome by reporting delay (in days).	107
Figure 4. 3: CFR by age, delay, country, and fever. Median and 95% confidence intervals (CI) are plotted (based on 1000 bootstrap realisations for ‘confirmed, probable and suspected cases’).	109
Figure 4. 4: Out-of-sample predictive accuracy measures (sensitivity, specificity, proportion of percentage correctly classified (PCC) and area under the receiver operating curve (AUC)) on the validation data obtained using 72 different hyperparameterisations of the BRT	

model. p represents the proportion of data to train the model with $p=0.5$, $p=0.65$ and $p=0.8$ explored..... 116

Figure 4. 5: Within-sample accuracy measures (sensitivity, specificity, proportion of percentage correctly classified (PCC) and area under the receiver operating curve (AUC)) on the training data obtained using 72 different hyperparameterisations of the BRT model. p represents the proportion of data to train the model with $p=0.5$, $p=0.65$ and $p=0.8$ explored. 117

Figure 4. 6: CFR estimates without (black) and with (orchid) imputation for age, delay, country and fever, for confirmed, probable and suspected cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations). 118

Figure 4. 7: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed and probable cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations). 119

Figure 4. 8: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed and probable cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations)..... 120

Figure 4. 9: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed and probable cases obtained for the training data using the (simplified) BRT

model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).	121
Figure 4. 10: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).	122
Figure 4. 11: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).	123
Figure 4. 12: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed, probable and suspected cases for the validation data obtained using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).	124
Figure 4. 13: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed, probable and suspected cases obtained for the validation data using the (simplified) BRT model with tree complexity =10, learning rate=0.01, bag fraction=0.5, data partitioning ratio=0.8. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).	126

Figure 4. 14: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed and probable cases for the validation data obtained using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).....	127
Figure 4. 15: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed and probable cases obtained for the validation data using the (simplified) BRT model with tree complexity =10, learning rate=0.01, bag fraction=0.5, data partitioning ratio=0.8. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).	128
Figure 4. 16: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed cases for the validation data obtained using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).....	129
Figure 4. 17: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed probable cases obtained for the validation data using the (simplified) BRT model with tree complexity =10, learning rate=0.01, bag fraction=0.5, data partitioning ratio=0.8. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).	130
Figure 4. 18: CFR for the other predictors in the simplified model without imputation (black), with imputation unadjusted(orchid) and with imputation adjusted (orange) for estimated BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed, probable and suspected' cases).....	134

Figure 4. 19: CFR for the other predictors in the simplified model without imputation (black), with imputation unadjusted (orchid) and with imputation adjusted (orange) for estimated BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed and probable' cases).	135
Figure 4. 20: CFR for the other predictors in the simplified model without imputation (black), with imputation unadjusted (orchid) and with imputation adjusted (orange) for estimated BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed').	136
Figure 4. 21: Performance (i.e. sensitivity, specificity, PCC, and AUC) of BRT model after down-sampling of data for all Ebola cases (i.e. confirmed, probable and suspected).....	139
Figure 4. 22: Imputed CFR and simulated missingness CFR for CPS (i.e. confirmed, probable and suspected) cases, CP (i.e. confirmed and probable) cases and C (i.e. confirmed) cases.	141
Figure 4. 23: A, B: EVD case proportions for confirmed and probable cases by age and reporting delay. C, D: EVD case proportion for confirmed cases by age and reporting delay.	142
Figure 4. 24: Within-sample accuracy for confirmed cases. p is the proportion of data used as training set. p=0.5, p=0.65 and p=0.8, respectively. PCC=Percentage Correctly Classified, AUC area under the receiver operating characteristic (ROC) curve. This shows an optimistic training performance but also indicates that I am not overfitting the model.	143
Figure 4. 25: Out-of-sample accuracy for confirmed and probable cases. p is the proportion of data used as training set. p=0.5, p=0.65 and p=0.8, respectively. PCC=Percentage Correctly Classified, AUC area under the receiver operating characteristic (ROC) curve. This shows, overall lower performance for confirmed only data compared to confirmed,	

probable and suspected data. The hyperparameters broadly remain the same as for confirmed, probable and suspected case data.	144
Figure 4. 26: CFR by age group, delay, country, and fever without imputation (black), with imputation unadjusted (orchid) and with imputation adjusted (orange) for BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed and probable' cases).	151
Figure 4. 27: CFR by age group, delay, country, and fever without imputation (black), with imputation (orchid) unadjusted and with imputation adjusted (orange) for BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed and probable' cases).	152
Figure 5. 1: Choropleth spatial distribution of case fatality ratio (CFR) for the 2013-2016 Ebola epidemic in West Africa. a) Median predicted CFR adjusted for imputation (and standard deviation) for each district. b) The median observed district-level CFR. Dark grey shading denotes districts for which data were unavailable. c) Median district-level residuals (i.e. observed CFR minus predicted CFR adjusted for imputation). In each case, the median is on the left and the standard deviation (sd) is on the right.	163
Figure 5. 2: Semivariograms fitted with a Gaussian model to the residuals of district-level CFR adjusted for imputation based on the Boosted Regression Tree model (BRT). The red line is the fitted model for the region as a whole and the blue lines are fitted models for individual countries (Sierra Leone, Guinea and Liberia). Note that the x-axes vary.	164
Figure 5. 3: Isopleth case fatality ratio (CFR) map for West Africa of a) median and b) standard deviation (sd) of kriged residuals (i.e. observed CFR minus predicted CFR adjusted	

for imputation). c) Isopleth map for kriged CFR (i.e. predicted CFR adjusted for imputation plus the kriged residuals). 165

Figure 5. 4: Temporal distribution of case fatality ratio (CFR) for the 2013-2016 Ebola epidemic in West Africa. a) Median predicted CFR adjusted for imputation (and 95% CI) for each quarter. b) The median observed quarterly CFR. c) Quarterly residuals (i.e. observed CFR minus predicted CFR adjusted for imputation). Note that there were too few cases in late 2013 and early 2016 to be meaningfully aggregated into quarter..... 166

Figure 5. 5: Semivariograms fitted with a Gaussian model to the residuals of quarter-specific CFR adjusted for imputations based on the Boosted Regression Tree model (BRT). The red line is the fitted model for the region as a whole and the blue lines are fitted models for individual countries (Sierra Leone, Guinea and Liberia). Note that the x-axes vary..... 167

Figure 5. 6: Theoretical semivariograms fitted to the residual empirical semivariogram for Sierra Leone, Guinea and Liberia combined..... 177

Figure 5. 7. Semivariograms fitted to the residuals of Observed CFR. Red line fitted model for overall, blue lines fitted models for individual countries. 178

Figure 5. 8. Semivariograms fitted to the residual of observed CFRs. Red line fitted model for overall, blue lines fitted models for individual countries. Note that the x-axes vary. 179

Figure 5. 9. Isopleth map for predicted CFR adjusted for imputation..... 180

Figure 5. 10. Directional semivariograms for overall (i.e. Sierra Leone, Guinea, and Liberia). 181

Figure 5. 11: Isopleth maps for kriged residuals. [a–b] median and standard deviation for the overall semivariogram fitted with a Gaussian model. [c–d] median and standard deviation for the overall semivariogram fitted with an exponential model. 182

Figure 5. 12: Isopleth map for kriged CFRs (i.e. predicted CFR adjusted for imputation plus the kriged residuals). a) Kriged CFR for semivariogram fitted with a Gaussian model. b) Kriged CFR for semivariogram fitted with an exponential model. 183

Figure 5. 13: Ebola virus disease cases (i.e. confirmed, probable and suspected) with known deaths/ total number of cases (i.e. confirmed, probable and suspected) per district for Sierra Leone, Guinea and Liberia. 184

Figure 6. 1: Inferential performances (mean squared error [mse]) for all 20 predictors of case fatality ratio (CFR) estimated for each inference method (i.e. LR, RF, BRT, BART, ANN). The dotted horizontal line separates the ten predictors used in the simulation process from the ten predictors added for model fitting. 198

Figure 6. 2: Survival outcomes and predictors missing completely at random (MCAR): Imputation performance (sensitivity, specificity, percentage correctly classified [PCC], area under the receiver operating characteristic curve [AUC]) as a function of the proportion of data used for model training and a) the proportion of outbreak data (p); where $p=1$ corresponds to the full dataset of 12,049 cases (100%); b) the proportion of simulated missingness of 0.1 (10%), 0.2 (20%) and 0.4 (40%). The imputation performance at which imputations are as good as chance is indicated by the dotted horizontal line at 0.5..... 200

Figure 6. 3: Survival outcome missing at random (MAR) and predictors missing completely at random (MCAR): Imputation performance (sensitivity, specificity, percentage correctly classified [PCC], area under the receiver operating characteristic curve [AUC]) as a function of the proportion of data used for model training and a) the proportion of outbreak data (p); where $p=1$ corresponds to the full dataset of 12,049 cases (100%); b) the proportion of simulated missingness of 0.1 (10%), 0.2 (20%) and 0.4 (40%). The imputation performance at

which imputations are as good as chance is indicated by the dotted horizontal line at 0.5.

..... 201

Figure 6. 4: Survival outcome missing not at random (MNAR) and predictors missing completely at random (MCAR): Imputation performance (sensitivity, specificity, percentage correctly classified [PCC], area under the receiver operating characteristic curve [AUC]) as a function of the proportion of data used for model training and a) the proportion of outbreak data (p); where $p=1$ corresponds to the full dataset of 12,049 cases (100%); b) the proportion of simulated missingness of 0.1 (10%), 0.2 (20%) and 0.4 (40%). The imputation performance at which imputations are as good as chance is indicated by the dotted horizontal line at 0.5. 203

Figure 6. 5: The relationship between the unadjusted CFR estimate for each missingness type (MCAR, MAR, MNAR) and the proportion of data used for model training and a) proportion of outbreak data (p); where $p=1$ corresponds to the full dataset of 12,049 cases (100%); b) proportion of simulated missingness of 0.1 (10%), 0.2 (20%) and 0.4 (40%). The true CFR of the complete simulated data (without simulated missingness) is indicated by the red dotted horizontal line. 205

Figure 6. 6: The unadjusted and adjusted CFR estimates for MAR and MCAR missingness in survival outcomes and predictors missing completely at random (MCAR). The true CFR of the complete simulated data (without simulated missingness) is indicated by the red dotted horizontal line. **Error! Bookmark not defined.**

Figure 6. 7: Comparing the unadjusted and adjusted CFR estimates for MCAR missingness in survival outcomes and predictors missing completely at random (MCAR). The true CFR of the complete simulated data (without simulated missingness) is indicated by the red dotted horizontal line. 216

Figure 6. 8: Comparing the unadjusted and adjusted CFR estimates for MNAR missingness in survival outcomes and predictors missing completely at random (MCAR). The true CFR of the complete simulated data (without simulated missingness) is indicated by the red dotted horizontal line. 217

List of tables

Table 2. 1: Number of confirmed, probable and suspected cases with reported survival outcomes/Number of confirmed, probable and suspected cases with unreported survival outcomes. 60

Table 2. 2: Summary statistics for known outcomes and unknown outcomes data for age and reporting delay. Known outcomes data are data without imputation and unknown outcomes data are data for which I conducted imputation. Reporting delay: delay between the date of reporting and the date of the onset of symptoms (in days). I used an independent two-sample t-test (p-values) to compare known outcomes data and unknown outcomes data.. 61

Table 2. 3: Proportion of missingness in candidate predictors for confirmed, probable and suspected cases. 63

Table 2. 4: Proportion of missingness in candidate predictors for confirmed and probable cases. 66

Table 2. 5: Proportion of missingness in candidate predictors for confirmed cases. 70

Table 4. 1: CFR estimates without imputation, unadjusted with imputation and adjusted with imputation for confirmed, probable and suspected cases. All CFR estimates and corresponding CIs were calculated using a non-parametric bootstrap of the BRT model. *

Adjusted CFR estimated using tp function to correct for bias in BRT model performance.. 106

Table 4. 2: Relative contributions (%) of the predictors in the minimal BRT model using 10-fold cross validation, $tc = 27$, $lr = 0.001$ and $bf = 0.75$ and trained on 1000 training sets generated by randomly sampling 65% cases with known survival outcomes without bootstrapping. The minimal model used these 10 predictors.	108
Table 4. 3: BRT model performance. ^a the full BRT model used all 43 candidate predictors. ^b the simplified BRT model used the 24 predictors retained after model simplification. ^c the minimal BRT model used the ten most important predictors. Median and 95% confidence intervals (CI) reported, based on 1,000 bootstrap realisations using confirmed, probable, and suspected cases.....	110
Table 4. 4: CFR estimates with imputation and confidence intervals are calculated from 1000 bootstraps of a BRT model. The BRT parameters used for imputation are $tc=27$, data partitioning=0.65, $lr=0.001$ and $bf=0.75$	131
Table 4. 5: CFR by fever occurrence and date of reporting. The cut-off date (25 th November 2014) was informed by previous work using the same data but for an earlier stage of the epidemic.	138
Table 4. 6: Relative contributions (%) of the important predictors used in BRT models using 10-fold cross validation, $tc = 27$, $lr = 0.001$ and $bf = 0.75$, for confirmed and probable cases, and trained on 1000 training sets generated by randomly sampling 65% cases with known survival outcomes. The minimal model used 10 predictors.	145
Table 4. 7: Relative contributions (%) of the important predictors used in BRT models using 10-fold cross validation, $tc = 27$, $lr = 0.001$ and $bf = 0.75$, for confirmed cases, and trained on 1000 training sets generated by randomly sampling 65% cases with known survival outcomes. The minimal model used 10 predictors.	146

Table 4. 8: CFR for confirmed cases. Overall and country level estimates. All CFR estimates and corresponding CIs were calculated using a non-parametric bootstrap of the BRT model.	147
Table 4. 9: CFR for confirmed and probable cases. Overall and country level estimates. All CFR estimates and corresponding CIs were calculated using a non-parametric bootstrap of the BRT model.....	148
Table 4. 10: BRT Imputation performance (simplified model) for confirmed cases and confirmed and probable cases.	149
Table 5. 1: Parameter estimates for the best-fit Gaussian semivariogram model fitted to the spatial residuals (observed CFR minus predicted CFR adjusted for imputation) from the BRT model.....	162
Table 5. 2: Spatial autocorrelation analysis for case fatality ratio (CFR) residuals in Sierra Leone, Guinea and Liberia from 2013 to 2016, with/without district predictor in the BRT model. The global Moran's I is for the three countries combined. The p-values in red are statistically significant, <0.05.....	172
Table 5. 3: Spatial semivariogram models fitted to overall district-level CFR residuals.....	176
Table 6. 1: Summary of the machine learning algorithms and hyper-parameters investigated.	191
Table 6. 2: Survival outcome, demographic and clinical predictors considered for this simulation study.....	211

List of abbreviations and acronyms

Ad26	recombinant adenovirus serotype 26
AIC	Akaike's information criterion
ANN	artificial neural network
AUC	Area under the curve
BART	Bayesian additive regression tree
CFR	Case fatality ratio
CI	confidence interval
CP	confirmed and probable
CPS	confirmed, probable and suspected
CYD	chimeric yellow fever-dengue
DRC	Democratic Republic of the Congo
D-W	Durbin-Watson
EBOVAC	Ebola vaccine project
ETC	Ebola treatment centre
EVD	Ebola Virus Disease
GADM	Database of Global Administrative Areas
GDP	gross domestic product
HCW	Health care worker
HIV	Human immunodeficiency virus
IPC	infection prevention and control
IQR	interquartile range

LISA	local indicator of spatial association
LR	logistic regression
LSHTM	London School of Hygiene and Tropical Medicine
MAR	missing at random
MCAR	missing completely at random
MNAR	missing not at random
MaLe	machine learning
MVA	modified vaccinia Ankara
PCC	percentage correctly classified
PPE	personal protective equipment
RDT	rapid diagnostic test
RF	random forest
RI	relative influence
RNA	ribonucleic acid
ROC	receiver operating characteristic
TDV	tetravalent dengue vaccine
TPOT	Tree-based Pipeline Optimization Tool
UK	United Kingdom
US	United States of America
VHF	viral haemorrhagic fever
VSV	vesicular stomatitis virus
WHO	World Health Organisation
ZEBOV	Zaire Ebola virus

Chapter 1: Introduction

1.1 Background

The Ebola virus belongs to the family of *Filoviridae* which include the Cuevavirus, Marburgvirus and Ebolavirus [1]. Six species of Ebola have been identified including Zaire ebolavirus, Sudan ebolavirus, Tai Forest ebolavirus, Bundibugyo ebolavirus, Reston ebolavirus and, most recently, the Bombali ebolavirus [2]. The Zaire, Sudan, Tai Forest and Bundibugyo are four of the ebolaviruses known to cause disease [3]. The first Ebola virus Disease (EVD) reported in Zaire (now the Democratic Republic of Congo) in 1976 was caused by the Zaire ebolavirus [4]. Since then, ebolaviruses have been named after the geographical locations in which they are first serologically identified. Figure 1.1 shows the geographic distribution of EVD in Africa [5].



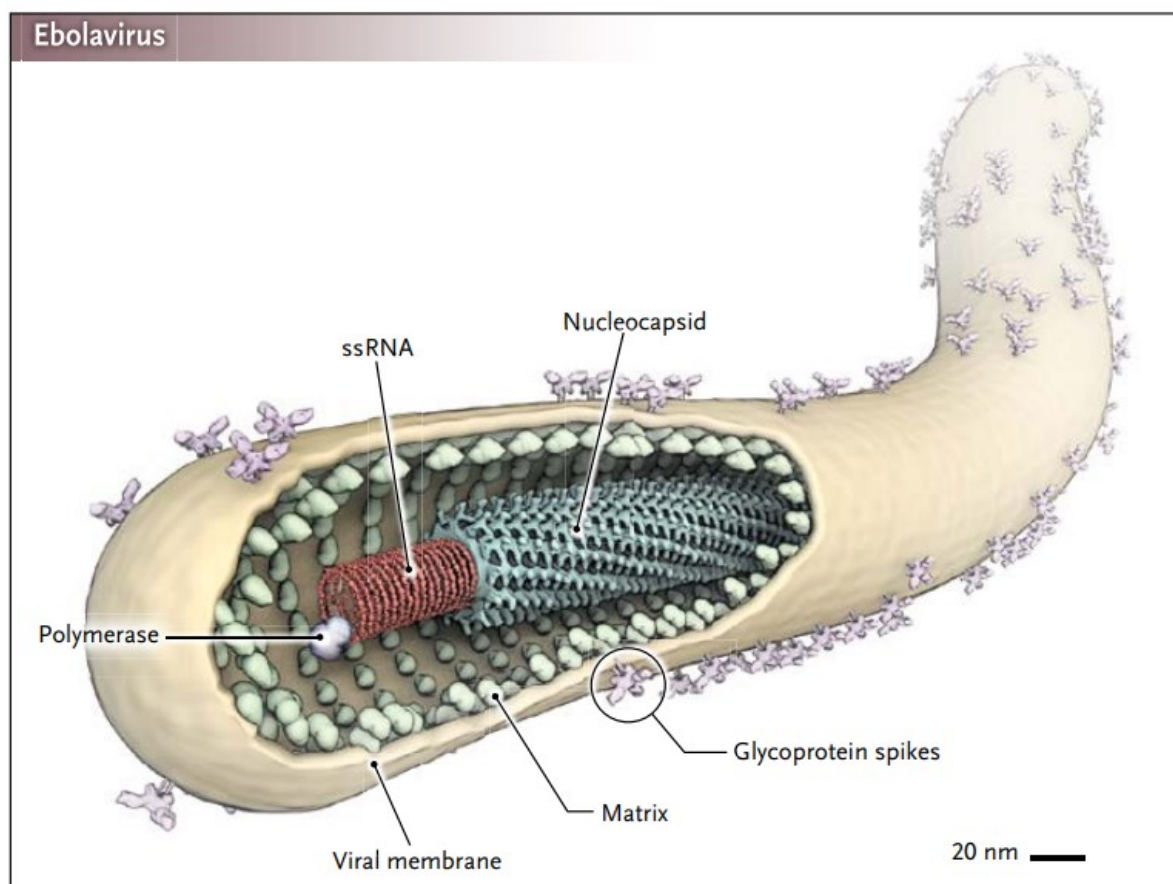
Figure 1. 1: Geographic distribution of Ebola in Africa. Reproduced from [5].

EVD caused by the Zaire ebolavirus is more severe compared to that caused by Tai Forest, Bundibugyo and Sudan [6]. There are inter-species and intra-species variations in the severity of these haemorrhagic fevers. In this thesis, my focus is on the severity of EVD caused by the Zaire ebolavirus and from here on, except when otherwise stated, by EVD I mean the Zaire ebolavirus disease.

EVD is a zoonotic disease initially transmitted to humans from animals. Fruit bats have been considered as the natural reservoir of the disease [7]. EVD enters the human host through mucosal surfaces and abrasions of the skin. Once the virus gets into the human body, it rapidly replicates and invades the immune system [8]. The pathological consequences that lead to death are the result of a damaged immune system [8]. Furthermore, the sexual transmission of EVD has been described in the literature. For example, the Ebola virus has been isolated from the semen of an EVD survivor on day 83 after symptom onset, and in 1 of 25 survivors (4%) up to month 18 [9, 10]. Sexual transmission can cause sporadic EVD cases long after an outbreak has peaked.

Individuals who eventually develop antibodies against the virus can survive the disease and lifelong immunity is conferred [11]. However, this theory of life-long immunity was recently challenged as a few survivors in the recent Democratic Republic of the Congo outbreak are reported to have been re-infected with Ebola. What is still unclear and subject to further investigation, is whether these reinfectees were misdiagnosed Ebola cases or whether reinfectee viral load at initial infection was suboptimal to elicit long-term immunity [12]. Immunologically, it is not uncommon for immunocompromised individuals (including pregnancies, diabetes and HIV) to develop a waning immunity. Immunocompromised individuals could therefore potentially be at-risk of reinfection. Comorbidities could also contribute to Ebola re-infection. For example, a recent study suggests that measles infection causes immune amnesia whereby, essentially, there is a reset of the human antibody repertoire acquired for every previous infection which could include Ebola [13]. Additionally, immune privileged sites, the eyes and the central nervous system can harbour the virus for months after systemic infection. For instance, the Ebola virus was detected in a survivor who developed meningoencephalitis 9 months after the acute infection [14].

Figure 1.2 summarises the microscopic structure of the Ebola virus. Ebola virus ribonucleic acid (RNA) has been detected in semen, genital secretions, skin, body fluids and nasal secretions of infected patients and so transmission is mainly sustained from human-to-human when families and unprotected health workers care for the sick, and when communities bury their dead [15]. Thus, control strategies implemented during EVD outbreaks are primarily meant to stop transmission chains by initiating safe patient care and safe burials.



Structure of Ebolavirus.

Shown is an ebolavirus particle and its characteristic filamentous shape. The negative-strand RNA genome is found in the center of particles in an encapsidated form as the nucleocapsid, together with the polymerase complex. Embedded in the virus membrane are trimeric glycoprotein spikes. Beneath the membrane is the matrix protein, which facilitates morphogenesis and budding of virus particles. The image is based on Protein Data Bank identifiers 3CSY and 1ES6 (www.rcsb.org) and Electron Microscopy Data Bank identifier EMD-2043 (www.emdatabank.org). The abbreviation ssRNA denotes single-stranded RNA.

Figure 1. 2: A microscopic structure of the Ebola virus: Reproduced from [16].

1.1.1 Ebola Virus Disease Epidemiology

From August 1976 to date, every EVD outbreak has provided us an opportunity to better understand the epidemiology of the disease. Ngoy Mosholu, the Chief Medical Officer of the Bumba Zone in Zaire, gave the first epidemiological description of EVD in his report to the central government of the Democratic Republic of the Congo in 1976 [4]. Ngoy Mosholu described a 'mystery disease' that caused fever, headache, abdominal pain and bleeding. Forty-three years later for both the epidemic in West Africa and the recent outbreak in the Democratic Republic of Congo, the signs and symptoms of EVD have remained broadly unchanged [4]. The World Health Organisation (WHO) Ebola Response Team in West Africa identified other symptoms that include anorexia, difficult swallowing, eye pain and hiccups among others [17]. Figure 1.3 shows a detailed description of the clinical features of EVD in West Africa [18]. Similar symptomatic patterns have been identified for the outbreak in the Democratic Republic of the Congo [19].

Epidemiological parameters that characterise the determinants, occurrence and distribution of EVD have been corroborated for successive outbreaks. For instance, nine months after the West Africa Ebola epidemic, the basic reproduction number which measures the number of secondary cases originating from one primary case was estimated as 1.71 (95% CI, 1.44 to 2.01) for Guinea, 1.83 (95% CI, 1.72 to 1.94) for Liberia, and 2.02 (95% CI, 1.79 to 2.26) for Sierra Leone [17]. These basic reproduction numbers were consistent with previous outbreak estimates. For example, the basic reproduction number for the 2000 outbreak in Uganda was estimated to be 1.34 (standard deviation = 0.03) and the basic reproduction number for the 1995 outbreak in the Democratic Republic of the Congo was estimated to be 1.83 (standard deviation = 0.06) [20]. These consistent estimates in West Africa were early indicators that

the large West Africa Ebola epidemic was not as the result of biological changes in the Ebola virus [17].

To have an idea of beds needed in hospitals, quarantine timelines and contact tracing strategies, distributional parameters for the serial interval, incubation period and delays (e.g. from onset to hospitalisation) were also estimated in West Africa. For instance, incubation period ranging from 2 to 21 days, has broadly remained consistent for successive EVD outbreaks informing estimates of a likely exposure period on which to focus interventions and timelines for contact tracing and case hospitalisation, thus helping to prevent the onward transmission of the virus. Public health measures like public advice, case isolation, personal hygiene and infection control precautions, when implemented effectively, ultimately reduce transmission of EVD in communities [21].

Severity estimates of EVD are not always consistent for different EVD outbreaks. In 1976, Ngoy reported 22 deaths from 30 EVD cases. In their first published report for the 2016 outbreak in the Democratic Republic of the Congo, the WHO Ebola Response Team reported 25 deaths from 50 cases [4, 17]. These discrepancies in the case numbers highlight the challenge of identifying cases and quantifying EVD severity, which is the central theme of this thesis.

Clinical Features of Ebola Virus Disease.		
Phase of Illness	Time since Symptom Onset	Clinical Features
Early febrile	0–3 days	Fever, malaise, fatigue, body aches
Gastrointestinal	3–10 days	Primary: epigastric pain, nausea, vomiting, diarrhea Associated: persistent fever, asthenia, headache, conjunctival injection, chest pain, abdominal pain, arthralgias, myalgias, hiccups, delirium
Shock or recovery	7–12 days	Shock: diminished consciousness or coma, rapid thready pulse, oliguria, anuria, tachypnea Recovery: resolution of gastrointestinal symptoms, increased oral intake, increased energy
Late complications	≥10 days	Gastrointestinal hemorrhage, secondary infections, meningoencephalitis, persistent neurocognitive abnormalities*

* Secondary infections are presumptive diagnoses based on clinical features of distributive shock, oral or esophageal candidiasis, and oral ulcers; meningoencephalitis is a presumptive diagnosis based on clinical features of unconsciousness and stiff neck.

Figure 1. 3: Clinical features of the 2013–2016 Ebola virus disease epidemic in West Africa.
Reproduced from [18].

1.1.2 Detection and diagnosis of Ebola

To provide treatment effectively and gauge severity, we must efficiently detect and diagnosis the disease. A major challenge since EVD discovery is case confirmation as most outbreaks emerge in communities with limited human and material resources. In 1976, the hospital spread of EVD, at least in part because of shared needles and syringe,s would have been halted if the hospital had the capacity to isolate and promptly identify the Ebola virus as well as stopping the sharing of the needles and syringes. Tellingly, these diagnostic challenges persisted during the 2013-2016 epidemic in West Africa [22]. Case diagnosis during the epidemic relied on polymerase chain reaction (PCR) tests and because of limited laboratory capacity and weak local transportation infrastructure, there were delays from when the samples were collected to when the results were available which often led further delays in freeing up beds and isolating the infected from the uninfected cases.

A rapid diagnostic test has been shown to reduce delays in EVD case confirmation [22, 23]. Nouvellet *et al.* in a modelling study showed that, if rapid diagnostic tests had been used in combination with PCR, the scale of the West African epidemic could have been reduced by over a third [22]. Rapid diagnostic tests are being used in the recent Democratic Republic of the Congo outbreak and the initial outcomes are encouraging [24]. Work is still ongoing to develop further diagnostic technology for EVD diagnosis and case confirmation. If these efforts are successful, efficient case confirmation would improve our ability to gauge reliably the severity of EVD, especially during ongoing outbreaks.

1.1.3 Case fatality ratio of Ebola

Epidemiologically, the severity of a disease can be quantified using the following parameters:

i) mortality rates, ii) proportionate mortality, iii) years of potential life lost and iv) case fatality ratios (CFR) [25]. The choice of the parameter used to express disease severity quantitatively depends on other clinical characteristics of the disease. For instance, severity is better expressed for chronic diseases (e.g. arthritis) using mortality rates wherein deaths from a chronic disease in a year are divided by the midyear population (an approximation to account for change in population over time) since, at least in theory, everybody in a population is liable to contract the chronic disease [25]. On the other hand, for acute infectious diseases like EVD, CFR is more appropriate because only infected individuals would die from their infection before the disease is eventually eliminated through public health interventions or naturally.

The CFR is defined as the proportion of cases who eventually die from a disease [25].

Mathematically, CFR is calculated as follows:

$$\text{Case fatality ratio (percent)} = \frac{\text{Number of individuals dying during a specified period of time after disease onset or diagnosis}}{\text{Number of individuals with the specified disease}} \times 100, (1)$$

Although conceptually a simple epidemiological parameter, CFR is one of the most important quantities used to characterise diseases. CFR can pinpoint the risk of dying from disease between people in different geographical areas and population sub-groups; help to determine the effectiveness of treating a disease over time and, in some circumstances, serve as an incidence rate surrogate [25].

Using the case numbers reported by Ngoy Mosholu in 1976, and the simple equation above, the case fatality ratio of EVD could have been naively estimated as 73.3 % (Binomial 95% CI 54.1%–87.1%). Similarly, the naïve CFR for the recent Democratic Republic of Congo outbreak could be estimated as 50.0% (Binomial 95% CI 35.5%–64.5%). Naïve CFR estimates can be misleading if case outcomes are unknown for a nonnegligible proportion of cases [4]. For instance, while Ngoy’s estimates could have alerted a functioning health system to respond swiftly to the disease in Yambuku, in the Democratic Republic of the Congo, for the 2018 outbreak in the same country, a naïve CFR of about 50% could have been a spurious underestimation or overestimation of CFR [4]. Spurious CFR underestimation could occur as a result of delayed reporting where, at time w , the persons who have died up to time w will not be the only ones to die of Ebola among those who became cases by time w [26]. In contrast, spurious CFR overestimation could occur as a result of preferential ascertainment of severe cases as surveillance systems are more likely to capture severe cases [26]. Lipsitch *et al.* highlighted some possible solutions to these biases in estimating CFR, solutions mainly based on either improving the quality of surveillance data (e.g. use sentinel surveillance sites to estimate multipliers between various levels of severity and extrapolate to a larger

population) or implementing statistical techniques that account for some of the biases (e.g. apply a competing-risk Kaplan-Meier-like method or a parametric mixture model to the full dataset) [26].

Using statistical techniques, WHO Ebola Response Team found that the CFR estimate of 56% (95% CI 39–72) for the 2018 Democratic Republic of the Congo outbreak was consistent with estimates for the 2014–2016 West African Ebola virus disease epidemic ($p=0.427$) [19]. Even with concordant CFR estimates from West Africa and the Democratic Republic of the Congo, data quantity and quality could lead to variation in CFR estimates and issues like right censoring of available data (where the available data are used to estimate CFR before all deaths have occurred for a given time period) must be accounted for to estimate CFR reliably.

When EVD outbreaks grow to become epidemics, we are even more challenged by the characteristics the available data, and the statistical techniques to reliably estimate CFR.

1.2 Ebola: from outbreak to epidemic

The three possible population-level stages of EVD after a spill-over event from animals to humans are outbreaks, epidemics and pandemics. These population-level stages have explicit epidemiological differences.

Outbreaks are sudden increases in the number of disease cases above what is expected in a specific population for a specific geographical area [27]. Outbreaks become epidemics on expansion of the geographical area over which cases occur but still limited to within the borders of a specific country and/or region [27]. Epidemics become pandemics when cases spread over several countries or continents, usually affecting a large number of people. Thankfully, in West Africa the spill-over event that infected a boy in Gueckedou village (in

Guinea) was prevented from becoming a global pandemic (although cases were isolated in Spain, UK, United States of America) but, sadly, it developed from an outbreak to a serious epidemic affecting Sierra Leone, Guinea and Liberia. A potentially serious expansion into Nigeria was stopped in its tracks thanks to the bravery and professionalism of health professionals, and what was clearly a better funded and managed public health infrastructure in Nigeria compared to Sierra Leone, Guinea and Liberia [28].

In 1976, the first outbreak of Ebola virus disease (EVD) was reported in Zaire and Sudan [29]. Although some subsequent outbreaks reported more than 200 cases (confirmed, probable, suspected), the collective number of EVD cases before 2014 had not exceeded 3400 [3], and the disease was confined to small rural communities away from the cities. However, the epidemiology of the disease was about to change at the onset of the 2013 epidemic in West Africa.

The West African epidemic caused severe socioeconomic problems with the broadest geographical reach. Over 28,000 cases and 11,000 deaths were reported at the end of the epidemic [30]. The fast-growing economies of Sierra Leone, Guinea and Liberia saw significant drops in gross domestic product (GDP)s. The pre-EVD projections for growth in GDP for Liberia 5.9%, Sierra Leone 11.3% and Guinea 4.5% were later revised to 2.5%, 8.0% and 2.4% respectively [31]. EVD had reached dense urban areas where people moved with relative ease [32]. The weak health systems in the affected countries were exposed, but so were bureaucracies and funding limitations of global health agencies [33].

Although the initial response was slow, it picked up with the allocation of material and human resources to the affected countries. Social mobilisation, contact tracing, treatment beds and safe burials gradually improved throughout the epidemic [34], as did the expert support from

global health organisations. For example, the WHO and other international partners launched a US\$100 million-dollar response plan towards epidemic control [35]. However, the focus and coordination of this human and material support were, at least initially, inadequate. There was an obvious need for a paradigm shift to i) review global health policy on Ebola, ii) develop new EVD treatment options and iii) novel vaccines and design clinical trials to fast-track vaccine development.

1.2.1 Global Policy on Ebola

After years of weak development and civil unrest, health systems in Sierra Leone, Guinea and Liberia were fragile at the onset of the EVD epidemic in 2014 [36]. Thus, national governments and local public health agencies could not manage a disease that was emerging in the region for the first time [37]. A sustainable solution to the emerging epidemic needed global leadership and coordination but this was not forthcoming, at least in the early stages of the epidemic [38].

When, eventually, the WHO director general declared the West African Ebola epidemic a public health emergency of international concern [36], it underscored the urgency of a coordinated international response and the imperative need to improve capacity in affected countries. The declaration triggered temporary recommendations to the affected states, land-border States and the international community [36]. The affected states (i.e. Sierra Leone, Guinea and Liberia) were asked to declare a national emergency, activate disaster management plans and establish emergency operation centres. The affected countries were expected to mobilise all sectors of society, including, health workers and traditional leaders, and protective equipment was to be provided to all staff associated with the Ebola response. Land-Border States (e.g. Ivory Coast, Guinea Bissau, Mali and Senegal) were expected to

conduct rigorous surveillance to identify promptly clusters of unexplained fevers and deaths. Qualified laboratories and rapid response teams were put on standby for contact tracing and case management [36]. The declaration cautioned against international traveling restrictions except for EVD cases and their established contacts. All countries were expected to implement risk communication, laboratory diagnostics and prepare for medical evacuations. These directives went a long way to intensify the global response to the epidemic. However, capacity building in the affected countries themselves was not included in the recommendations and local public health professionals were ill-prepared to take ownership of the Ebola response [36]. Treatment and vaccination options were also limited and the few available at the time were not enough adequately to serve the response [39]. The global community was never ready for an epidemic of that magnitude.

1.2.2 Treatment Options for Ebola

Before the West African Ebola epidemic, there was little incentive, at least on the part of the pharmaceutical industry, promptly to develop treatment options for EVD. When 2 USAID workers got infected with EVD in Liberia, the only treatment available was an experimental Ebola antibody [40]. This experimental antibody, made available for international staff only, fuelled discontent about global justice [36]. However, global health experts were quick to point out that the experimental serum had been previously used only in nonhuman primates and its safety and efficacy were still unknown [36]. Thus, even with limited quantities of the serum available, there were still ethical issues around using experimental therapies in a large population.

As the epidemic expanded across the region, research teams started carrying out clinical trials to further test the safety and efficacy of other treatment options. For instance, in Sierra

Leone, in phase 2 single arm clinical trial, TKM-130803, a small interfering RNA lipid nanoparticle product, administered at a dose of 0.3 mg/kg/d by intravenous infusion to adult patients with severe EVD, showed no improvement in survival when compared to historic controls [41]. With little safety and efficacy data available, ZMapp, a cocktail of antibody-based drugs targeting both distinct and overlapping parts of the Ebola virus glycoprotein, showed promise for immediate treatment in an epidemic situation. However, with limited quantities, and challenges in storing and administering the drug in tropical and under-resourced environments (shipping and storage temp is -20°C), the isolated individual successes could not be replicated in the wider infected population [42]. Antivirals (e.g. Favipiravir (T-705) and Brincidofovir) and the repurposing of pre-existing therapies (e.g. Artemisinin and anticoagulant therapy) were proposed as treatment options but none proved effective throughout the West African epidemic [43].

Recent studies from the outbreak in the Democratic Republic of the Congo are looking more promising. Two experimental drugs, Regeneron's REGN-EB3 and a monoclonal antibody called mAb114, harvested from the antibodies of survivors did better than both ZMapp and the antiviral drug-remdesivir [44]. For 499 patients, an independent data safety and monitoring board found that only 29% of patients on REGN-EB3, 34% of patients on mAb114 had died, compared with 49% on ZMapp and 53% on remdesivir [44]. However, while these results are promising, they must be interpreted with caution as sample sizes are still comparatively small and there could be underlining confounding factors driving these initial estimates. With research still ongoing, there is a need for better treatment options to reduce Ebola CFR during outbreaks and/or epidemics.

1.2.3 Ebola Vaccines and Clinical trial designs

In 2013, there were no licensed vaccines for EVD pre-exposure and postexposure prophylaxis. The pharmaceutical and research communities were playing catch up and, eventually, only small quantities of vaccines were available for experimentation and potential use during the West African epidemic [45]. Regional and global guidelines on the use of these experimental vaccines in outbreaks settings were unclear and so governments and global health agencies like WHO and the United States Centre for Disease Control (US-CDC) were left with the challenge of dealing with the ethical issues arising at the time. Ultimately, ethical issues around patient consent and whether to withhold potential vaccines from some sections of the affected population became incontestable. Fast-tracked clinical studies were designed and implemented to enable all at-risk individuals to access the vaccines.

In Guinea, an open-label, cluster-randomised ring vaccination trial was carried out as the epidemic was ongoing [46]. Cases reported from health care facilities and the community, and confirmed in designated laboratories were used to define clusters of all contacts and contacts of contacts (including absent residents), and the recombinant, replication-competent vesicular stomatitis virus-based vaccine expressing a surface glycoprotein of Zaire Ebolavirus (rVSV-ZEBOV) was randomly administered immediately and after a delay within each cluster [46]. This 'ring' design ensured that all at-risk individuals got the vaccine but also provided a baseline to test for vaccine efficacy. The concluding findings from the Guinea study show that rVSV-ZEBOV offers substantial protection against Ebola virus disease [47].

As case numbers began to fall, the vaccine trials that followed were primarily safety and immunogenicity studies [48]. For instance, in Sierra Leone, the EBOVAC-Salone study was established and implemented to validate further the safety and immunogenicity of the

Ad26.ZEBOV/MVA-BN-Filo prime-boost Ebola vaccine regimen as results of a similar phase 1 trial in Kenya had shown that it was well tolerated and immunogenic [49].

Encouragingly, moving from the West African epidemic, the WHO and partners in the Democratic Republic of the Congo are implementing similar vaccine strategies in response to the ongoing outbreak. A ring vaccination programme, similar to that implemented in Guinea, is being deployed as part of a new international approach for rapid mitigation of outbreaks through multiple interventions [50]. Hopefully, the development of Ebola vaccines will continue and, importantly, clinical trial designs and vaccines will continue to evolve to address the unique challenges each Ebola outbreak brings.

1.2.4 Prevention, surveillance and control

Aside from treatment and vaccines for EVD, more traditional approaches are still required to mount a robust response during an Ebola outbreak. These traditional approaches are cost-effective as they only require the skills and commitment of frontline health and social care workers. Early case diagnosis contact tracing and case isolation are at the heart of this frontline response. Early diagnosis usually requires effective diagnostics as described earlier, but importantly field workers must be incentivised to provide such services even in hard to reach areas. Successful contact tracing involves the assessment of EVD symptoms, interviewing techniques and counselling [51]. Contact tracers should not only have investigative and analytical skills but should also build trust between themselves, the affected communities and the public health system. Case isolation and safe burials are very sensitive stages of an Ebola response. For these to be effective, health communication and social mobilisation efforts should be encouraged to improve the public's knowledge, attitudes, and practices regarding EVD [51]. Safe burials must not only be dignified but

burial teams should be appropriately trained to use personal protective equipment (PPEs) and the PPEs incinerated after use [51].

A key pillar of an Ebola response is the prevention of health care worker (HCW) infections particularly when care is being provided for individuals with EVD. For instance, an HCW case fatality ratio of 57% has been reported for the West Africa Ebola epidemic suggesting that at least half of the carers who attended to the patients and became infected died of EVD themselves [52]. The fact that HCWs died from EVD with a similar CFR as the community may not necessarily reflect the lack of training. However, HCWs with skills training before taking up frontline roles in EVD care would do better at protecting themselves and providing adequate care for their patients. These skills should include competency in performing all Ebola-related infection control practices and procedures, and the proper use of infection prevention and control (IPC) tools and equipment, specifically facial masks and PPEs [51].

1.3 Machine learning and Geostatistical methods

From John Snow's classic use of dot-maps to show the location of deaths in the 1854 Cholera outbreak to modern day use of mobile phone data to predict the spatial spread of cholera [53, 54], the need for methods and tools to collect and analyse outbreak data remains a priority for epidemiologists and public health professionals. Modern day tools like mobile phones and smartwatches are revolutionising data collection but, particularly in resource-limited settings, standardised paper-based data capture tools are still relevant and useful.

Methods for data analysis seen in the literature range from ubiquitous geostatistical methods to the use of artificial intelligent systems [55, 56]. Artificial intelligent systems, including machine learning, learn complex patterns in large datasets while geostatistical methods are used widely to characterise the spatiotemporal patterns in large datasets.

In this thesis, I focus on the use of machine learning and geospatial methods for the analysis of the 2013-2016 West African Ebola dataset. Data for this work were collected using the World Health Organisation (WHO) viral haemorrhagic fever (VHF) case reporting forms which are standard paper-based forms used for infectious disease surveillance. Data used for my work are described in detail in Chapter 2 of this thesis. I would now introduce, with a few examples, the use of machine learning and geostatistical methods in infectious disease modelling.

1.3.1 Machine learning in Infectious disease modelling

Machine learning is a major branch of artificial intelligence whereby less restrictive algorithms are used to learn patterns in data [57]. In a machine learning framework, a functional relationship between a response and a predictor is not assumed. Instead, flexible algorithms are used to learn the relationship between an outcome and its proposed predictors [58]. Because of this flexible approach, machine learning techniques are generally very effective in predicting outcomes at the individual level. Machine learning techniques are employed in a range of scientific disciplines, including actuarial science, economics and ecology [59-61].

Since their early development, machine learning algorithms have been designed and used to analyse medical datasets made possible by the digital revolution that provides relatively inexpensive and available means to collect and store a large set of data. Datasets from different sub-fields (e.g. epidemiology, genetics, oncology) are being used to improve diagnostics, disease treatment and prevention [62-64]. Because machine learning techniques show high predictive performances, inherently handle missing and noisy data, identify influential predictors and interactions, they are ideal to answer relevant questions in infectious disease epidemiology. Bhatt *et al.* used Boosted Regression Trees, a machine

learning technique to map the global distribution and burden of the dengue virus [65]. Gu *et al.* used random forest, to attribute Salmonella illnesses to sources of infection involving many causal pathways [66]. These uses of machine learning illustrate their effectiveness in estimating epidemiological parameters.

In the future, it is anticipated that machine learning will play a major role in infectious disease outbreak preparedness and response [56]. In this thesis, I demonstrate the power of machine learning in an epidemic analysis by refining the CFR estimates for the 2013–2016 West African Ebola epidemic.

1.3.2 Geostatistics in infectious disease modelling

Gaining a better understanding of the spatiotemporal structure of infectious diseases is increasingly being recognised as a key to improving our knowledge about their population dynamics and control [67]. Geostatisticians continue to develop methods that provide powerful frameworks to characterise the spatial patterns in infectious diseases. For instance, within a Bayesian framework, Raso *et al.* used a geostatistical model to map accurately the infection intensity of schistosomiasis in the Ivory Coast [68].

Other common approaches are also being used by infectious disease researchers. For example, Srividya *et al.* used geostatistical methods (i.e. semivariograms and kriging) to analyse the geographical distribution of lymphatic filariasis in Southern India [69]. Carla Nunes used a similar approach to identify spatiotemporal clusters of tuberculosis incidence in Portugal [70].

With the ubiquitous use of different geostatistical methods for different infectious diseases, in different countries, I am motivated to apply similar approaches to characterise the spatiotemporal heterogeneity of EVD CFR.

1.3.3 Characterising the spatiotemporal epidemiology of Ebola

Geostatistical methods have previously been used to understand the epidemiological parameters of the EVD [71, 72]. The literature points to quite a few studies that spatiotemporally characterise EVD transmission [73, 74]. For example, Kramer *et al.* used network models to estimate the spatial process of EVD transmission and found that the probability of transmission between locations depended on distance, population density and the international cross-border importations of EVD cases [73]. For Sierra Leone, Fang *et al.* used Poisson transmission models to map spatiotemporal transmission patterns, identify influential factors, and assess the effects of interventions at sub-administrative levels (chiefdom-level in Sierra Leone) [74].

However, the spatiotemporal characterisation of EVD CFR has seldomly been reported in the literature. Using EVD cases with unreported survival outcomes, Garske *et al.* detected outliers in CFR estimates between districts of residence and treatment centres during the West African epidemic, but I haven't found any more studies that characterise the spatiotemporal heterogeneity of EVD CFR. In this thesis, I intend to leverage the strength of both machine learning and geostatistical methods to further characterise the spatial epidemiology of CFR.

1.4 Justification

My thesis will focus on refining the CFR estimates of EVD and characterising the sub-national heterogeneities in EVD CFR. I will particularly focus on using machine learning to impute for Ebola cases with unknown survival outcomes and use geostatistical approaches to characterise the spatiotemporal heterogeneities in CFR. As outlined in the previous sections, there are knowledge gaps in understanding the epidemiology of EVD CFR. No previous study has included data with unreported survival outcomes to estimate EVD CFR and only a few

studies have examined in detail the spatiotemporal heterogeneities EVD CFR. Machine learning and geostatistical methods could be useful tools to fill in these gaps. The new epidemiological knowledge could be used to support future outbreak/epidemic response. There is also a need for robust analytical frameworks which could be deployed in real time to improve the situational awareness around ongoing outbreaks/epidemics.

1.4.1 Thesis aims and objectives

The aims of my thesis are to re-estimate EVD CFR for the 2013-2016 epidemic, to characterise spatiotemporal heterogeneity in CFR at fine spatiotemporal scales and investigate the general utility of machine learning in estimating CFR. Addressing these aims will provide new insight into the epidemiology of the 2013–2016 West African Ebola epidemic with a view of informing future public health contingency planning, resource allocation and impact assessment. These aims will be reached through the following objectives:

1. To calibrate a machine learning model (i.e. Boosted Regression Trees) to impute for unreported survival outcomes.
2. To combine the imputed and reported survival outcomes to re-estimate CFRs for the West African Ebola epidemic.
3. To characterise key sources of spatiotemporal heterogeneity in EVD CFRs.
4. To compare various machine learning algorithms in estimating EVD CFR as a function of outbreak data characteristics.

1.5 Thesis Outline

Following this introductory chapter, Chapter 2 introduces the 2013–2016 West African EVD case line list dataset, which is referenced throughout this thesis. Chapters 3 describe details of the machine learning (i.e. Boosted Regression Trees) method, geostatistical methods (i.e.

Semivariogram and Kriging) and other machine learning inference methods used in Chapters 4, 5 and 6. Chapter 4 reports the detailed findings from the use of machine learning (i.e. Boosted Regression Trees) to i) identify the important predictors that explain the variance in observed CFR, ii) impute for unreported survival outcomes, iii) adjust imputations for model biases and iv) re-estimate more representative EVD CFR. Chapter 5 reports the detailed findings from the use of geostatistical methods (i.e. semivariogram analysis and kriging) to analyse the patient mix in detail to characterise key sources of spatiotemporal heterogeneity in EVD CFR. In Chapter 6, I evaluate with simulations how various machine learning inference methods perform in estimating EVD CFR under different outbreak scenarios. Lastly, Chapter 7 discusses the findings, limitations, and future direction of this work and how it provides 1) new insight into the epidemiology of the 2013–2016 West African Ebola epidemic that can 2) inform future public health contingency planning, resource allocation and impact assessment. Figure 1.4 presents a summary schematic outline of the thesis.

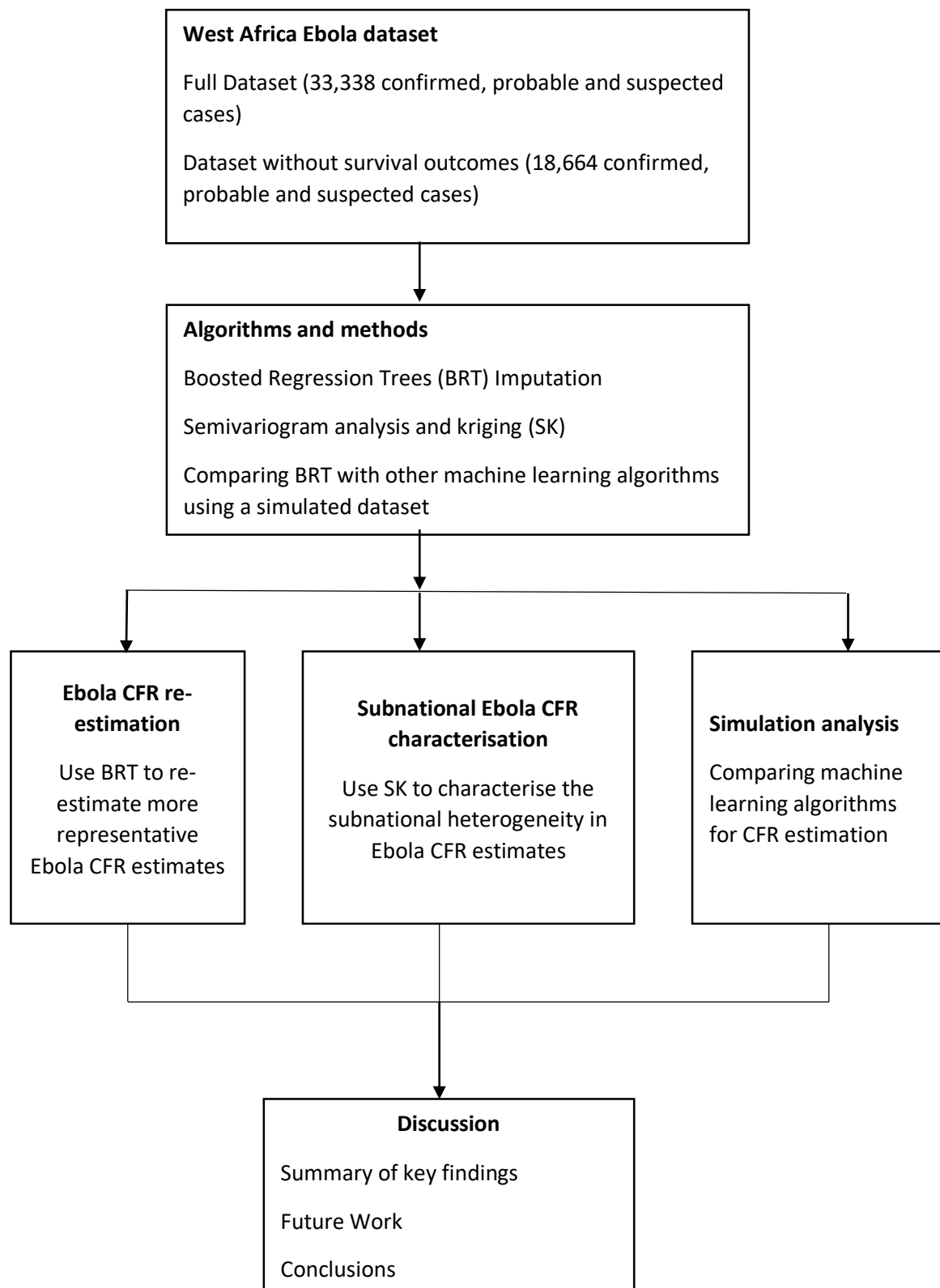


Figure 1. 4: Schematic summary of the chapters in this thesis.

Chapter 2: Understanding the 2013–2016 West Africa Ebola Dataset

Some work in this chapter forms the basis of: **Forna A**, Nouvellet P, Dorigatti I, Donnelly CA.

Case fatality ratio estimates for the 2013 – 2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation. *Clinical Infectious Diseases* 2019 and **Forna A**, Dorigatti I, Nouvellet P, Donnelly CA. Spatiotemporal Variability in Case Fatality Ratios for the 2013–2016 Ebola epidemic in West Africa. *International Journal of Infectious Diseases* 2020.

2.1 Overview

The data used throughout this thesis came from the WHO's comprehensive case line list collated during the 2013–2016 West African Ebola epidemic. With technical input from the WHO, data were collected by field staff at the health ministries in Sierra Leone, Guinea and Liberia. As a WHO collaborating centre for infectious disease modelling, the Department of Infectious Disease Epidemiology, at Imperial College London, had access to the data which it then used to produce timely situational reports. These situational reports were then used to help inform the Ebola response.

2.2. Background

Routinely collected during surveillance and response activities for EVD in Guinea, Liberia, and Sierra Leone, the West African Ebola dataset is the largest dataset ever curated for an Ebola epidemic and allows for a detail investigation of EVD epidemiology. For example, nine months after the West African epidemic had started, epidemiological data published by the WHO Ebola Response Team indicated that, without drastic improvements in control measures, the numbers of cases and of deaths from EVD were expected to increase from hundreds to thousands per week in just a few months [17]. These findings and others

helped to galvanise the global response that eventually led to the elimination of Ebola in West Africa. Subsequent publications from the WHO Ebola Response Team and other research groups continue to improve our understanding of the West African epidemic. Findings from West Africa provide useful insights to help researchers, public health professionals and policymakers understand the epidemiology of the recent outbreak in the Democratic Republic of the Congo [19].

The work that went into data collection and curation for the West African Ebola epidemic was no mean feat. Surveillance teams had to take risks in sometimes hostile communities to follow epidemiological chains, do contact tracing and move suspected cases to the nearest health facilities [51]. Clinicians had to provide the best possible care with limited resources and with the risk of being infected themselves [51]. Despite these challenges, surveillance teams and clinicians managed to fill in the standardised viral haemorrhagic fever (VHF) forms for Ebola cases in the communities and the health facilities. These data, just as those collected even in ideal settings, needed cleaning and curation for further analysis. Extensive data cleaning and curation were done by the WHO Ebola Team at Imperial College London and I am grateful they allowed me to use the curated data in this thesis.

In this chapter, I have described the 2013-2016 Ebola data and identified their characteristics that motivated the epidemiological questions and methodological approaches I used in answering the questions. Specifically, I define the case classification system used during the West African epidemic as this underpins the data structure, identify predictors of interest from the data and I characterise the nature of missingness in the data (i.e. missingness in survival outcomes and missingness in the potential predictors of CFR reported by clinicians and surveillance teams).

2.3 Geographical description of West Africa

The map shows the 63 districts of Sierra Leone, Guinea and Liberia from where data were collected throughout the epidemic (Figure 2.1). The epidemic started in Gueckedou, a district (prefecture) in Guinea which became the initial epicentre of the epidemic [75]. As seen in Figure 2.1, Gueckedou is bordered by districts in both Sierra Leone (i.e. Kono and Kailahun) and Liberia (i.e. Lofa) and, beyond the political boundaries, inhabitants (i.e. the Kissi people) in these districts share a common language and similar cultures [76]. This inter-country homogeneity together with clinical, social and demographic factors contributed greatly to magnitude of the West African Ebola epidemic. In Chapter 1, I described some of these clinical and sociodemographic factors that contributed towards the making of the epidemic in West Africa.

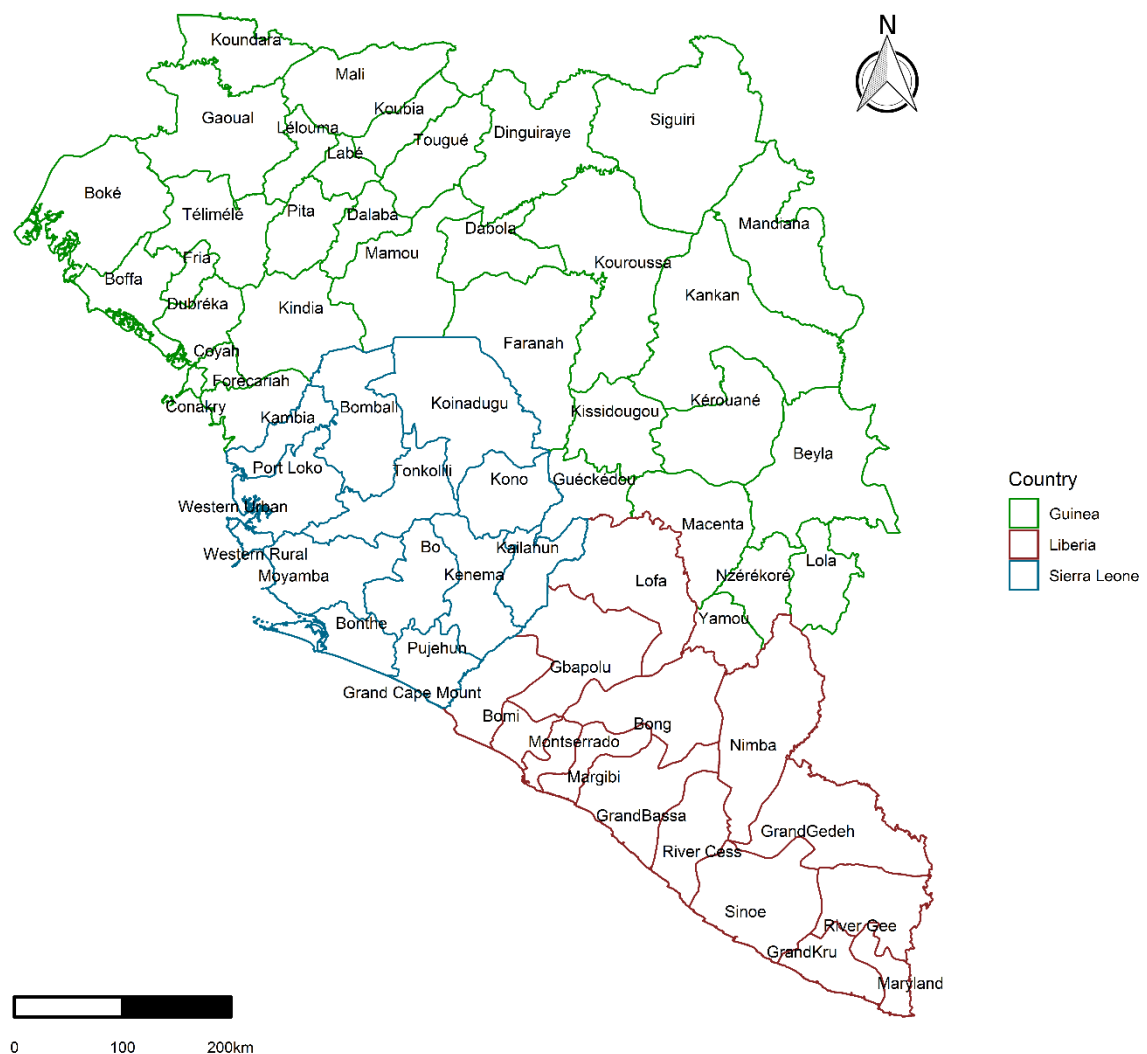


Figure 2. 1: The geographical location and distribution of administrative districts in Sierra Leone, Guinea and Liberia.

2.4 Case Classification System used in West Africa

Case classification systems can vary depending on the data collection practices of the authorities involved in an Ebola response. For instance, data are being collected for confirmed and probable cases in the Democratic Republic of the Congo whereas data were mainly collected for confirmed, probable and suspected cases in West Africa. In West Africa,

data collection practices changed over the course of the epidemic. For example, Liberia switched from an aggregate system of reporting to an individual-based system of reporting that could have affected data curation at the country level [77]. Eventually, in West Africa, the case classification system was harmonised and scaled on the degree of certainty by which an individual was considered a case. Individual cases were then classified as either confirmed, probable or suspected.

In almost all instances where Ebola data are to be used to estimate epidemiological parameters, case classification directly influences epidemiological parameters. For instance, the CFR estimates for confirmed cases could be completely different from estimates for confirmed, probable and suspected cases combined [78]. Therefore, in this thesis, I have considered case classification an important determinant of EVD CFR. In all instances in this thesis, confirmed, probable and suspected cases are defined as follows:

1. **Confirmed Cases:** Confirmed cases are defined as persons who returned a positive laboratory test (i.e. polymerase chain reaction test) for EVD [17].
2. **Probable Cases:** Probable cases are defined as persons suspected to have EVD as evaluated by a clinician or persons who died from suspected Ebola and had an epidemiologic link to confirmed cases. Such persons need not have been tested or confirmed by a laboratory [17].
3. **Suspected Cases:** Suspected cases are defined as persons, alive or dead, with sudden onset of high fever and who had contact with suspected, probable, or confirmed EVD cases or with a dead or sick animal; persons with sudden onset of high fever and a minimum of three of the following symptoms: headache, vomiting, anorexia or loss

of appetite, diarrhoea, lethargy, stomach pain, aching muscles or joints, difficulty swallowing, breathing difficulties, or hiccapping; or any person who had unexplained bleeding or who died suddenly from an unexplained cause [17].

2.5 Identification of Predictors from the Data

The West African EVD case line list is a large dataset that captures 1,211 potential predictors and 33,338 confirmed, probable and suspected cases. From this large dataset, I selected the predictors required for CFR estimation and spatiotemporal analysis. Some predictors required for the analysis were created from existing predictors. For instance, the predictor reporting delay is the difference between the date of a case was reported and the date of disease onset. Continuous predictors like reporting delay and patient age had to be converted to categorical predictors for some aspects of the analysis. The outcome variable and potential predictors are described in detail.

2.5.1 Survival outcome Variable

The survival outcome for this dataset was reported as a “final status” of patients. The final status was defined as alive, dead or missing. The fate of patients with missing final status is unknown e.g. patients could be alive in the communities or later died in the communities but were not reported to the appropriate authorities. Reported (alive and dead) data account for 56% of our final status for patients, thus 44% of the final status of patients were missing.

2.5.2 Potential Predictors

We selected 43 potential predictors of EVD CFR (Table 2.3 – Table 2.5). This selection was partly informed by what had already been documented in the literature [17]; partly by a judgement on the potential influence of predictors on EVD CFR. Non-numerical predictors

were converted into factors. Conveniently, no categorisation of continuous predictors was required as machine learning accommodates non-linear relationships. However, health worker positions were re-coded to reduce the 25 health worker positions recorded, broadly ranging from ward porters, drivers to consultant medical doctors.

2.5.2.1 Demographic Predictors

For this analysis, demographic predictors include country of origin, age, gender, health care worker, health worker position and the delay between the date of the onset of symptoms and the date the case was reported. Additionally, to evaluate the risk of exposure, I defined the health care workers' high-contact, mid-contact, low-contact and other. The classification of the risk was based on our knowledge of these roles.

2.5.2.2 Clinical Predictors

The clinical predictors are being further divided into two sub-groups; the bleeding predictors and the non-bleeding predictors.

Bleeding predictors

These predictors specifically define bleeding observed in patients. They include gum bleeding, nose bleeding, skin bleeding, stool bleeding, urine bleeding, hematemesis, vomiting blood, coughing blood, other bleeding, bleeding from an injection site, vaginal bleeding, and unexplained bleeding.

Non-Bleeding Predictors

These predictors specifically define all the general signs and symptoms usually associated with EVD. They include fever, anorexia, vomiting, abdominal pain, muscle pain, joint pain,

chest pain, cough, difficulty breathing, difficulty swallowing, conjunctivitis, sore throat, confusion, hiccups, jaundice, painful eyes, rash and unconsciousness.

2.6 Missingness in the West African Ebola dataset

In this thesis, my focus is to use potential clinical and demographic predictors to explain the observed variance in CFR and impute for unreported survival outcomes. Thus, I describe missingness in the dependent variable (i.e. survival outcome) and the potential predictors. The entire dataset has 33,338 confirmed, probable and suspected cases. There are 14,694 cases with unreported survival outcomes representing about 44% of total case numbers. There was a marked difference in completeness at the country level, with data recorded much more completely in Guinea than in the other countries (i.e. Sierra Leone and Liberia) (Table 2.1). Except for case classification, all potential predictors have some missing data (Table 2.3 – Table 2.5).

Table 2. 1: Number of confirmed, probable and suspected cases with reported survival outcomes/Number of confirmed, probable and suspected cases with unreported survival outcomes.

Case classification	All Countries	Guinea	Liberia	Sierra Leone
confirmed	9,471/6,973	3,297/7	2,384/1,359	3,790/5,607
probable	1,821/2,089	443/0	747/853	631/1,234
suspected	7,352/5,632	0/10	1,493/1,294	5,859/4,328
Total	18,644/14,694	3,740/17	4,624/3,509	10,280/11,169

2.6.1 Comparison of age and reporting delay between cases with known and unknown outcome

To identify the nature of the missingness in survival outcomes, I used the central limit theorem (i.e. for comparison of means based on large sample sizes) to compare cases with known and unknown survival outcomes in the same confirmation status categories (Table 2.2) and characterised the distribution of the data. This allowed us to investigate whether survival outcome missingness is independent of the characteristics of the cases.

Table 2. 2: Summary statistics for known outcomes and unknown outcomes data for age and reporting delay. Known outcomes data are data without imputation and unknown outcomes data are data for which I conducted imputation. Reporting delay: delay between the date of reporting and the date of the onset of symptoms (in days). I used an independent two-sample t-test (p-values) to compare known outcomes data and unknown outcomes data.

Predictors	Confirmed Cases			Confirmed and Probable Cases			Confirmed, Probable and Suspected		
	Known	Unknown	p-value	Known	Unknown	p-value	Known	Unknown	p-value
	Outcomes*	Outcomes ^a		Outcomes*	Outcomes ^a		Outcomes*	Outcomes ^a	
	Median (IQR)	Median (IQR)		Median (IQR)	Median (IQR)		Median (IQR)	Median (IQR)	

Age (in years)	30.0 (19.0-44.0)	28 (17.0-40.0)	<0.0001	30.0 (19.0-45.0)	28.0 (16.0-40.0)	<0.0001	30.0 (14.0-46.0)	29.0 (17.0-43.0)	<0.0001
Reporting delay (in days)	4.0 (2.0-6.0)	3.0 (2.0-5.0)	<0.0001	4.0 (2.0-6.0)	3.0 (2.0-6.0)	<0.0001	4.0 (2.0-7.0)	3.0 (2.0-6.0)	<0.0001

2.6.2 Proportion of predictors with known and unknown survival outcomes

I report the missingness in the potential predictors used this analysis for i) confirmed, probable and suspected cases, ii) confirmed and probable cases iii) and confirmed cases. As stated in Chapter 1, machine learning algorithms inherently handle this missingness in predictors when used for data imputation [58]. These inherent missingness handlers include mean/median imputation for the missingness in each continuous predictor; frequent value imputation for each categorical predictor, and the replacement of missing values with surrogates.

Confirmed, probable and suspected cases

Table 2. 3: Proportion of missingness in candidate predictors for confirmed, probable and suspected cases.

Predictors	Known Survival Outcomes			Unknown Survival Outcomes		
	Non-	Missing	%	Non-	Missing	%
	missing	Cases	Missingness	missing	Cases	Missingness
	Cases			Cases		
Bleeding gums	5326	13318	71.4	3296	11398	77.6
Bleeding nose	5310	13334	71.5	3280	11414	77.7
Bleeding skin	5274	13370	71.7	3254	11440	77.9
Bleeding stool	5320	13324	71.5	3279	11415	77.7
Bleeding urine	5271	13373	71.7	3250	11444	77.9

Hematemesis	5123	13521	72.5	3207	11487	78.2
Bloody vomit	5091	13553	72.7	3185	11509	78.3
Bloody cough	5280	13364	71.7	3256	11438	77.8
Bleeding other	5168	13476	72.3	3215	11479	78.1
Bleeding injection site	5297	13347	71.6	3257	11437	77.8
Bleeding vagina	5190	13454	72.2	3212	11482	78.1
Unexplained bleeding	11580	7064	37.9	7976	6718	45.7
Fever	14010	4634	24.9	10238	4456	30.3
Fatigue	13736	4908	26.3	10037	4657	31.7
Anorexia	13255	5389	28.9	9937	4757	32.4
Vomiting	13126	5518	29.6	9508	5186	35.3
Diarrhoea	12948	5696	30.6	9370	5324	36.2
Headache	12675	5969	32.0	9455	5239	35.7
Abdominal pain	12482	6162	33.1	9295	5399	36.7
Muscle pain	12477	6167	33.1	9368	5326	36.2
Joint pain	12309	6335	34.0	9404	5290	36.0
Chest pain	9289	9355	50.2	6502	8192	55.8

Cough	9047	9597	51.5	6262	8432	57.4
Difficult breathing	11497	7147	38.3	8757	5937	40.4
Difficult swallowing	11125	7519	40.3	8573	6121	41.7
Conjunctivitis	11134	7510	40.3	8669	6025	41.0
Sore throat	8822	9822	52.7	6350	8344	56.8
Confused	4975	13669	73.3	3202	11492	78.2
Hiccups	11558	7086	38.0	8368	6326	43.1
Jaundice	5056	13588	72.9	3288	11406	77.6
Painful eyes	4878	13766	73.8	3154	11540	78.5
Rash	9747	8897	47.7	7788	6906	47.0
Unconsciousness	4921	13723	73.6	3151	11543	78.6
Districts in Sierra Leone	18358	286	1.5	13969	725	4.9
District in Liberia	18358	286	1.5	13969	725	4.9
Districts in Guinea	18358	286	1.5	13969	725	4.9
Gender	18096	548	2.9	14372	322	2.2

Age	18171	473	2.5	13629	1065	7.2
Quarter	16610	2034	10.9	10730	3964	27.0
Hospitalisation status	14811	3833	20.6	9457	5237	35.6
Case classification	18644	-	-	14694	-	-
Health care worker position	748	17896	96.0	284	14410	98.1
Delay	14918	3726	20.0	9699	4995	34.0

Confirmed and probable cases

Table 2. 4: Proportion of missingness in candidate predictors for confirmed and probable cases.

Predictors	Known Survival Outcomes			Unknown Survival Outcomes		
	Non-	Missing	%	Non-	Missing	%
	missing	Cases	Missingness	missing	Cases	Missingness
	Cases			Cases		
Bleeding gums	3932	7360	65.2	2260	6802	75.1
Bleeding nose	3923	7369	65.3	2252	6810	75.1

Bleeding skin	3895	7397	65.5	2237	6825	75.3
Bleeding stool	3936	7356	65.1	2250	6812	75.2
Bleeding urine	3897	7395	65.5	2239	6823	75.3
Hematemesis	3744	7548	66.8	2186	6876	75.9
Blood vomiting	3717	7575	67.1	2178	6884	76.0
Bloody cough	3901	7391	65.5	2238	6824	75.3
Bleeding other	3778	7514	66.5	2193	6869	75.8
Bleeding	3915	7377	65.3	2242	6820	75.3
injection site						
Bleeding vagina	3838	7454	66.0	2209	6853	75.6
Unexplained	6054	5238	46.4	4291	4771	52.6
bleeding						
Fever	8021	3271	29.0	5619	3443	38.0
Fatigue	7808	3484	30.9	5496	3566	39.4
Anorexia	7342	3950	35.0	5443	3619	39.9
Vomiting	7286	4006	35.5	5226	3836	42.3
Diarrhoea	7169	4123	36.5	5161	3901	43.0
Headache	6987	4305	38.1	5198	3864	42.6
Abdominal pain	6749	4543	40.2	5067	3995	44.1

Muscle pain	6795	4497	39.8	5101	3961	43.7
Joint pain	6640	4652	41.2	5115	3947	43.6
Chest pain	5011	6281	55.6	3588	5474	60.4
Cough	4840	6452	57.1	3444	5618	62.0
Difficult breathing	5740	5552	49.2	4688	4374	48.3
Difficult swallowing	5486	5806	51.4	4593	4469	49.3
Conjunctivitis	5611	5681	50.3	4717	4345	47.9
Sore throat	4555	6737	59.7	3455	5607	61.9
Confused	3571	7721	68.4	2225	6837	75.4
Hiccups	5988	5304	47.0	4508	4554	50.3
Jaundice	3634	7658	67.8	2264	6798	75.0
Painful eyes	3518	7774	68.8	2206	6856	75.7
Rash	5032	6260	55.4	4204	4858	53.6
Unconscious	3545	7747	68.6	2197	6865	75.8
Districts in Sierra Leone	11065	227	2.0	8519	543	6.0

Districts in Liberia	11065	227	2.0	8519	543	6.0
Districts in Guinea	11065	227	2.0	8519	543	6.0
Gender	11237	55	0.5	8882	180	2.0
Age	11023	269	2.4	8353	709	7.8
Quarter	10090	1202	10.6	5874	3188	35.2
Hospitalisation status	9222	2070	18.3	5440	3622	40.0
Case classification	11292	-	-	9062	-	-
Health care worker position	641	10651	94.3	177	8885	98.0
Delay	9617	1675	14.8	5356	3706	40.9

Confirmed cases

Table 2. 5: Proportion of missingness in candidate predictors for confirmed cases.

Predictors	Known Survival Outcomes			Unknown Survival Outcomes		
	Non-	Missing	%	Non-	Missing	%
	missing	Cases	Missingness	missing	Cases	Missingness
	Cases			Cases		
Bleeding gums	3063	6408	67.7	1333	5640	80.9
Bleeding nose	3057	6414	67.7	1328	5645	81.0
Bleeding skin	3038	6433	67.9	1314	5659	81.2
Bleeding stool	3073	6398	67.6	1322	5651	81.0
Bleeding urine	3039	6432	67.9	1319	5654	81.1
Hematemesis	2876	6595	69.6	1264	5709	81.9
Blood vomiting	2857	6614	69.8	1256	5717	82.0
Blood cough	3041	6430	67.9	1317	5656	81.1
Bleeding other	2906	6565	69.3	1269	5704	81.8
Bleeding inject site	3052	6419	67.8	1320	5653	81.1
Bleeding vagina	2995	6476	68.4	1295	5678	81.4

Unexplained bleeding	4766	4705	49.7	2818	4155	59.6
Fever	6523	2948	31.1	3862	3111	44.6
Fatigue	6336	3135	33.1	3770	3203	45.9
Anorexia	5903	3568	37.7	3730	3243	46.5
Vomiting	5843	3628	38.3	3533	3440	49.3
Diarrhoea	5734	3737	39.5	3490	3483	49.9
Headache	5591	3880	41.0	3532	3441	49.3
Abdominal pain	5401	4070	43.0	3431	3542	50.8
Muscle pain	5440	4031	42.6	3465	3508	50.3
Joint pain	5292	4179	44.1	3477	3496	50.1
Chest pain	3942	5529	58.4	2247	4726	67.8
Cough	3784	5687	60.0	2134	4839	69.4
Difficult breathing	4524	4947	52.2	3100	3873	55.5
Difficult swallowing	4299	5172	54.6	3005	3968	56.9
Conjunctivitis	4444	5027	53.1	3130	3843	55.1
Sore throat	3528	5943	62.7	2135	4838	69.4

Confused	2761	6710	70.8	1278	5695	81.7
Hiccups	4693	4778	50.4	2952	4021	57.7
Jaundice	2820	6651	70.2	1308	5665	81.2
Painful eyes	2713	6758	71.4	1261	5712	81.9
Rash	3933	5538	58.5	2770	4203	60.3
Unconsciousness	2731	6740	71.2	1251	5722	82.1
Districts in Sierra Leone	9260	211	2.2	6478	495	7.1
Districts in Liberia	9260	211	2.2	6478	495	7.1
Districts in Guinea	9260	211	2.2	6478	495	7.1
Gender	9428	43	0.5	6827	146	2.1
Age	9246	225	2.4	6356	617	8.8
Quarter	8397	1074	11.3	4212	2761	39.6
Hospitalisation status	7821	1650	17.4	3864	3109	44.6
Case classification	9471	-	-	6973	-	-

Health care	577	8894	93.9	128	6845	98.2
worker position						
Delay	8010	1461	15.4	3793	3180	45.6

2.7 Discussion

In this chapter, I have described the case classification system used in West Africa, characterised the missingness in survival outcomes and potential predictors that explain the observed variance in EVD CFR. The case classification system informs the entire analysis and CFR refinement. In subsequent chapters, I have presented estimates for different combinations of confirmed, probable and suspected cases. Missingness characterisation informs the modelling approaches used in this thesis. For instance, although statistical inference methods like logistic regression can identify the important predictors that explain the observed variance in EVD CFR, they can only be used to estimate CFRs that includes cases with no missingness in predictors or wherein missing cases are either deleted or imputed. Machine learning inference methods should consider both observed and missing predictors in CFR estimation. In this thesis, I used machine learning inference methods to impute missing data and re-estimate EVD CFR.

To justify the use of machine learning for missing data imputation, the types of data missingness must be well characterised. The three main types of missing data are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [79]. When data are MCAR, missing values are not systematically different from observed values and missingness is considered a chance process. When data are MNAR, missing values are systematically different from observed values as a result of the missing

value themselves influencing the probability of missingness or that some unmeasured predictor predicts both the values of the missing predictors and the probability of missingness [80]. In this instance, machine learning could be useful if the unmeasured predictors are identified and later included in the analysis.

Hypothetically, machine learning is most useful when data are MAR. For MAR data, the missing values are systematically different from the observed values, but the systematic differences are fully accounted for by measured predictors. Thus, machine learning can be used to infer patterns from observed data (with both survival outcomes and predictors) and these inferences can be used to impute survival outcomes for data with only measured predictors. In Table 2.3, for the West African EVD data, I show that survival outcomes are not MCAR and that the missingness depends on the measured predictors. With this characterisation of the missingness in the 2013–2016 West African EVD dataset, I have leveraged the strength of machine learning to refine EVD CFR estimates and analysed the detail case mix to characterise key sources of spatiotemporal heterogeneity in the refined estimates. Also, to better understand how outbreak data characteristics influence CFR estimation, I carried out an outbreak simulation analysis that compares the performance and imputed CFRs of machine learning inference methods given various outbreak data scenarios.

Chapter 3: Methods and algorithms for Ebola case fatality ratio estimation, geospatial analysis and simulation study

Work in this chapter form the basis of: **Forna A**, Nouvellet P, Dorigatti I, Donnelly CA. Case fatality ratio estimates for the 2013 – 2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation. *Clinical Infectious Diseases* 2019 and **Forna A**, Dorigatti I, Nouvellet P, Donnelly CA. Spatiotemporal Variability in Case Fatality Ratios for 2013–2016 Ebola epidemic in West Africa. *International Journal of Infectious Diseases* 2020.

3.1 Overview

This chapter describes details of the methods and algorithms used in subsequent chapters. I described boosted regression trees as the main ensemble machine learning algorithm used in this thesis. Other techniques that underpin the analytic framework are described. These include hyperparameter optimisation, cross validation, non-parametric bootstrapping and adjustment for model sensitivity and specificity. While the geostatistical approach used in this thesis is adequately described in Chapter 5, algorithms built on the boosted regression tree framework are also described in this chapter. Finally, the algorithmic framework for the simulation analysis is also described in Chapter 3. Thus, Chapter 3 compliments Chapters 4, 5 and 6 where the focus is on applying these algorithms to i) impute and re-estimate CFR for the 2013–2016 West Africa epidemic (Chapter 4) ii) characterise key sources of spatiotemporal heterogeneity in EVD CFR (Chapter 5) and iii) evaluate data imputation model performance and the resulting CFR estimates while varying outbreak data characteristics (Chapter 6).

3.2 Conceptual description of BRT

Boosted Regression Tree (BRT) is a machine learning algorithm that combines regression/classification trees with “boosting” [58]. For regression trees, a set of decision rules is built on the predictors. These decision rules are constructed by recursively splitting the data into successively smaller subgroups using binary splits. For all predictors, splits are repeatedly applied to output until the best splits are identified. The best splits are those that maximise the homogeneity of the smaller subgroups with respect to the outcome (e.g. death/recovery).

For the classification trees used in this study, the splits are the tree branches and the homogeneous subgroups are terminal nodes or “leaves” of the tree. The boosting part of the model is a form of gradient descent that adds new trees to reduce the gradient of the loss function– a measure indicating the loss in imputation performance due to a suboptimal model, thus further reducing variance in the outcomes. The final model is an additive model in which each term is a tree [58]. Because this is a forward stage-wise fitting process, cross validation (more details provided in section 3.5) was applied to further optimise the number of trees and prevent overfitting.

The BRT algorithm, implemented in the “gbm package” and used in this study can be found here [81]. Briefly, this algorithm depends on four hyperparameters (tree-complexity, data partitioning ratio, learning rate and bag fraction). Changing these hyperparameters affects the model performance. A short description of each hyperparameter is given below, with more details provided in section 3.4. Tree complexity determines the number of nodes in each tree. The data partitioning ratio determines the proportion of data used for model training and out-of-sample validation. The learning rate is a scaling factor used to scale the

predictions of each model. The smaller the learning rate, the less importance given to each single tree in the model. Bag fraction introduces stochasticity to the model thus preventing overfitting. To optimise model performance, the tree complexity, data partitioning ratio, learning rate and bag fraction were all varied.

3.3 Selection of BRT threshold

BRT models return probabilities for the outcome. These probabilities can be interpreted if they are converted into binary values. This conversion is possible when a threshold probability is defined. Different threshold optimisation techniques have been proposed in the literature. For example, the researcher might choose an arbitrary threshold above and below which a binary classification is made. A default threshold probability is 0.5 but others have been recommended [82]. Here, I briefly describe the study-specific threshold optimisation strategies.

The BRT probabilities returned for the outcome could also be used to randomly sample outcomes as part of a bootstrap process. While this approach was not considered for this thesis, it does provide a methodological direction to be considered in future work.

3.3.1 Optimisation of Prevalence

Optimisation of prevalence sets the threshold at the observed prevalence of survival. It is most useful when prevalence is the most important factor being studied.

3.3.2 Maximisation Percentage Correctly classified

This method maximises total accuracy. Intuitively, maximising the accuracy should be the main goal. However, where prevalence is very low, maximising the percentage correctly classified may entirely neglect the existence of the less prevalent class.

3.3.3 Maximisation of Specificity and Sensitivity

Sensitivity is the proportion of positives(alive) that are correctly classified as such whereas specificity is the proportion of negatives(dead) correctly classified as such. By maximising the sum of the specificity and the sensitivity, the mean error rates for alive and dead observations are minimised. This strategy is independent of prevalence and over-predicts rare observations.

3.3.4 Equal Sensitivity and Specificity

The strategy here is to ensure that the sensitivity is equal to the specificity. It optimises the threshold so that death predictions are just as likely to be wrong as alive predictions. This method is correlated with prevalence thus rare observations are given a low threshold. It has been reported in the literature that this method is not the most suitable for mapping predictions [83].

Having examined the different optimisation strategies in detail, throughout this, I have selected the equal sensitivity and specificity threshold optimisation. This optimisation approach allows for CFR predictions without trading off dead and alive outcomes. The consensus in the literature is to minimise the false positives and negatives rates and/or maximise the true negative and positive rates [84].

3.4 Measure of BRT performance

The four model performance measures used in this thesis are the Area under the Receiver Operator Characteristic (ROC) Curve, Specificity, and Sensitivity, Percentage Correctly classified (PCC). They are described as follows:

3.4.1 Area Under the ROC Curve

One measure used to evaluate BRT performance is the ROC curve, a plot of the true positive predictions of the outcome against the false positive prediction of outcome [85]. The area under the ROC curve take values that range from 0 to 1. As a measure of performance, ROC values between 0.9 are regarded as good, values between 0.7 to 0.9 are regarded as reasonable, values between 0.5 to 0.7 are poor and values equal to 0.5 indicate a random performance that is only due to chance [86].

3.4.2 Sensitivity

The sensitivity is the proportion of positives(alive) predicted to be positive by the BRT model [85].

3.4.3 Specificity

Specificity is the proportion of negatives(dead) predicted to be negative by the BRT model [85].

3.4.4 Percentage Correctly Classified PCC

The percentage correctly classified (PCC) is a measure of the percentage of predictions that are correctly classified, those could be either positive (alive) or the negative (Dead).

3.5 *K*-fold cross validation

K-fold cross validation is the way overfitting is monitored and avoided when fitting BRT models. In this thesis, I implemented a 10-fold cross validation, which means that the training data used in building the BRT model were divided into ten subsets and at each fold, the model was trained on 9 (out of the 10) unique subsets and the omitted subset was used for holdout testing [58]. The average prediction error across the testing subsets is a measure

of goodness-of-fit. This allows us to identify overfitting and determine the performance of the model.

3.6 Description of BRT hyperparameters used to optimize performance

Hyperparameters are configuration variables that are external to the machine learning model and whose values cannot be estimated from data [87]. Hyperparameters cannot be learned directly from the data in standard model training and therefore hyperparameters must be specified before training a model [87]. The right hyperparameter values will eliminate the chances of overfitting and underfitting a machine learning model.

To eliminate overfitting or underfitting a model, hyperparameters are optimised by performing a Grid search or Random search. With a grid search, I specify the hyperparameters I want to experiment with, and the range of values to try and grid search will perform all the possible combinations of the hyperparameter values using cross validation. With random search, however, a larger hyperparameter space is searched to randomly picking out the hyperparameters and combinations of these are used at each iteration [87].

In this thesis, I used a grid search for hyperparameter optimisation. Grid search allowed me extensively to search the literature to inform the choices of hyperparameters to experiment with. Also, findings from a previous study suggest that grid and random search could produce similar performances for BRT models. Thus, hyperparameter combinations were gridded to optimise the BRT model used in this thesis. Hyperparameters optimised in the BRT model are i) tree complexity ii) learning rate iii) bag fraction and iv) data partitioning ratio.

3.6.1 Tree Complexity (tc)

The tree complexity (tc) sets the number of nodes in each tree of the model. $tc=1$ implies that each tree contains one node and two leaves and $tc=2$ implies that the tree contains two nodes and four leaves. The optimal number of trees is a function of both the learning rate and tree complexity [58]. In the grid search used for optimal hyperparameterisation, I included tc between 5 to 40 at intervals of 5.

Furthermore, the tc defines the interactions in the BRT model, for a $tc=1$, each tree contains only one stump (predictor) and so a collection of trees will behave as just an additive model of individual predictors. However, for $tc>1$, more predictors are used for each tree and each tree could be regarded as an interaction term of predictors. Interactions are inherently included when building BRT models and the probabilities at each terminal node reflects these interactions in the model.

3.6.2 Learning Rate (lr)

The learning rate (lr), also called the shrinkage factor, determines the contribution of each tree to the final model [58]. The smaller the learning rate, the less each tree influences the final model and, therefore, a larger number of trees is generally needed to minimise the holdout deviance and larger computation time is needed. In the grid search used for optimal hyperparameterisation, I included lr of 0.01, 0.001 and 0.0005, as suggested in the literature [58].

3.6.3 Bag fraction (bf)

Bag fraction (bf) introduces stochasticity into a BRT model which improves accuracy and reduces overfitting. Reducing overfitting introduces variance to the fitted values [87]. With a default of 0.5, 50% of observations are selected without replacement at each iteration.

When the $bf=1$, no stochasticity is introduced. Optimal bf (0.5 to 0.75) has been suggested in the literature [58]. In the grid search used for optimal hyperparameterisation, I included bf between 0.5 to 1 at intervals of 0.25.

3.6.4 Data Partitioning Ratio

In this analysis, I tested three data partitioning approaches to select the optimal hyperparameterisation. These were as follows:

Pareto's Principle (80/20 partitioning)

I used 80% of the data for training the model and the remaining 20% for out-of-sample validation.

Freidman's (65/35 or 70/30 partitioning)

I used 65% of the data for training the model and the remaining 35% for out-of-sample validation [88]. A slight variant of this is the 70% training and 30% validation [89]. In our grid search, I used the 65/35 partitioning.

Half and Half Partitioning (50/50 partitioning)

I used 50% of the data for training the model and remaining 50% for out-of-sample validation.

3.7 Non-parametric bootstrapping

The bootstrap is a widely applicable and very powerful statistical tool to estimate the uncertainty around estimates. Non-parametric bootstrapping allows us to calculate uncertainty using the original dataset without predefining the underlying distribution function. For an original dataset of n cases, I randomly sample n cases with replacement to produce a bootstrap dataset. This random sampling is repeated B (e.g. 1000 times) times

and the final mean/median estimates and corresponding confidence interval are calculated from these B bootstrap datasets [87]. In the context of this study, I applied the non-parametric bootstrap to our modelling framework, generated 1000 CFR estimates with imputation and from these calculated the median and 95% confidence interval.

Within our modelling framework, I divided the data into data with known survival outcomes and data with unknown survival outcomes. The data with known survival outcomes were further divided into training and out-of-sample datasets. I fitted the model on the training data and measured out-of-sample performance on the out-of-sample data. Relying on this out-of-sample performance, I used the BRT model to impute the data with unknown outcomes. CFR estimates were then calculated combining both the imputed and known survival outcomes.

Bootstrapping allowed us to quantify the variance of these CFR estimates, as I repeatedly (1000 realizations) sampled all the data with replacement before I divided the data into the initial two parts (i.e. data with known and data with unknown outcomes), fitted the model and imputed for the unknown outcomes. In bootstrapping, because I sampled with replacement each time, cases in the training data could have appeared in the out-of-sample data thus producing over optimistic out-of-sample performance and unreliable CFR estimates that are not a true reflection of the imputation potential of the model. In my implementation, by ensuring that for each bootstrap dataset, cases in training data were excluded from the out-of-sample dataset, the out-of-sample performance and corresponding robust CFR estimates I estimated are a true reflection of the imputation potential of the BRT model.

While this approach that excluded cases in training data from the out-of-sample dataset reflected the potential of the BRT model, it did not exclude completely the risk of over-optimistic BRT model performance in genuinely new data. For large datasets, performance is better gauged using a holding-back option whereby a portion of data is not used at any stage during model training.

3.8 Algorithm implemented for hyperparameter optimisation

For given values of tc , lr and bf , and the proportions (p) of cases in training set, BRT models were built and validated using the following algorithm:

1. Generate the training set by randomly sampling without replacement the specified proportions of cases with death and survival outcomes.
2. Generate the out-of-sample validation set, consisting of the cases with observed survival outcomes excluded from the training set (step 1).
3. Build the BRT model through the `gbm.step` function in the 'dismo' package, using the training set built in step 1, the specified tc , lr , bf , 10-fold cross validation and increasing the model in steps of 50 trees at each iteration, using survival outcome (i.e. death or survival) as response predictor and gender, health care worker, health care worker position, bleeding gums, bleeding nose, bleeding skin, bleeding stool, bleeding urine, haematemesis, vomiting, cough, bleeding from other sites, bleeding injection site, bleeding vagina, unexplained bleeding, fever, jaundice, painful eyes, rash, unconsciousness, fatigue, anorexia, vomiting, diarrhoea, headache, abdominal pain, muscle pain, joint pain, chest pain, coughing blood, difficult

breathing, conjunctivitis, sore throat, confused, hiccup, difficulty swallowing, case classification, quarter, age, delay, current hospitalisation, districts and country of origin as predictors. The optimal number of trees minimising the holdout (cross validation) deviance is computed.

4. Use the BRT model built in step 3 to predict the survival outcomes of the cases in the validation set. For each subject i in the validation set, the BRT model gives the probability p that subject i is dead.
5. Define the cut-off threshold p_T by which cases with $p_i > p_T$ are classified as dead or alive otherwise. I chose cut-off values giving equal sensitivity and specificity using the 'optimal.threshold' function in the 'PresenceAbsence' package.
6. Use the cut-off threshold p_T computed in step 5 to classify cases in the validation set as dead or alive.
7. Compute the sensitivity (proportion of deaths correctly classified), specificity (proportion of survival correctly classified) and percentage of predictions correctly classified in the validation set PCC and the area under the receiver operating characteristic curve (AUC).

The choice of optimising the cut-off thresholds to have equal sensitivity and specificity in step 5 was made to avoid introducing bias in CFR estimated with imputation.

3.9 Algorithm implemented for CFR with imputation estimation

Mean CFR estimates with imputation

The mean CFR estimates with imputation were calculated from 1,000 realisations of steps 1– 7, followed by steps 8, 9 and 10 below:

8. Use the BRT model built in step 4 to impute for cases with missing survival outcomes.
9. Use the cut-off threshold p_T computed in step 6 to classify imputed cases as dead or alive.
10. Calculate the CFR using the cases with both imputed and observed survival outcomes.

Confidence interval

Confidence intervals around the mean CFR estimates were calculated by bootstrapping.

Using the optimal tuning parameter set ($tc = 27$, $lr = 0.001$, $bf = 0.75$), the 95% CI of the CFR estimates with imputation were calculated as the 2.5-97.5 percentiles of 1,000 random realisations of the algorithm detailed below (steps 2 – 9 below are the same as steps 1 – 7):

1. Generate a synthetic population by sampling with replacement the 18,644 confirmed, probable, and suspected cases in the Ebola epidemic line list (confirmed and confirmed and probable for sensitivity analysis).
2. Generate the training set by sampling without replacement 65% of cases from the synthetic population sampled in step 1 with known survival outcomes.
3. Generate the out-of-sample validation set – i.e. cases in the synthetic population generated in step 1 with known survival outcomes who are not included in the training set (step 2).
4. Build the BRT model through the 'gbm.step' function in the 'dismo' package, using the training set built in step 3, $tc = 27$, $lr = 0.001$, $bf = 0.75$, 10-fold cross validation and increasing the model in steps of 50 trees at each iteration,

using survival outcome as response predictor and using the predictors in the simplified model.

5. Use the BRT model built in step 4 to predict survival outcomes in the validation set. For each subject i in the validation set, the BRT model calculates the probability p that subject i is dead.
6. Define the cut-off threshold p_T by which cases with $p_i > p_T$ are classified as dead or alive otherwise. I chose cut-off values giving equal sensitivity and specificity using the 'optimal.threshold' function in the 'PresenceAbsence' package.
7. Use the cut-off thresholds p_T computed in step 6 to classify the cases in the validation set generated in step 3 as dead or alive.
8. Compute the sensitivity (proportion of dead cases correctly classified), specificity (proportion of case that survived correctly classified) and percentage of predictions correctly classified in the validation set and the area under the receiver operating characteristic curve (AUC).
9. Use the BRT model built in step 4 to impute the survival outcomes of the cases in the synthetic prediction set (i.e. the cases sampled in step 1 with missing survival outcomes).
10. Calculate the CFR among the cases with complete and imputed survival outcomes.

3.10 Algorithm implemented for CFR adjusted for BRT model bias

Mean Adjusted CFR estimated

The mean CFR estimates adjusted for bias in the BRT model were calculated from 1,000 realisations of steps 1 – 7 (page 81-82), followed by steps 8,9,10 and 11 below:

8. Generate cases with just the imputed death outcomes.
9. Use the dead cases in step 8, the total number of cases in the imputed data and *tp* function in the ‘RSurveillance’ package to estimate the inferred CFR [90].
10. Multiply the number of cases in the imputed data by the inferred CFR to get the number of inferred deaths f_T .
11. Adjusted CFR is the ratio of the sum of f_T and the number of deaths in data with known survival outcomes, divided by the sum of all deaths and survivals in the complete data (i.e. data with both imputed and observed outcomes).

Confidence interval

Although the *tp* function of the “RSurveillance” package provides parametric approaches to estimate confidence intervals [90], in our implementation, confidence intervals around the mean adjusted CFR estimates were calculated by non-parametric bootstrapping (see section 3.7). Using the optimal tuning parameter set ($tc = 27$, $lr = 0.001$, $bf = 0.75$), the 95% CI of the adjusted CFR estimates were calculated as the 2.5-97.5 percentiles of 1,000 random realisations of the algorithm detailed in both steps 1 – 7 (page 81-82) and 8 – 11 (page 84).

3.11 Algorithm implemented to validate sensitivity/specificity adjustment

The mean CFR for simulated missingness were calculated from 1000 realisations of step 1-7 (page 81-82), followed by steps 8 to 17 below:

8. Generate a subset of data d_1 containing only cases with observed deaths.

9. Generate a new predictor for d_1 from a uniform distribution with minimum 0 and maximum 1 and order d_1 by this new predictor.
10. Sample from d_1 (without replacement) cases in the proportion of (inferred CFR) estimated after adjusting for sensitivity and specificity as described previously (page 84).
11. Generate a subset of data d_2 containing only cases with observed survivors.
12. Generate a new predictor for d_2 from a uniform distribution with minimum 0 and maximum 1 and order d_2 by this new predictor.
13. Sample from d_2 (without replacement) cases in the proportion of (1–inferred CFR) estimated after adjusting for sensitivity and specificity as described previously (page 84).
14. Combine d_1 and d_2 to create a full data d_x with simulated missingness.
15. Use the BRT model in step 1-7 in section 3.8 to predict survival probabilities for d_x .
16. Use optimal threshold p_T from the BRT model to convert survival probabilities to binary alive and survival outcomes.
17. Estimate CFR of the full data d_x and compare with CFR of only the imputed outcomes for BRT model in step 1-7.

Confidence interval

The confidence interval around the mean CFR for simulated missingness was calculated by non-parametric bootstrapping (i.e. sample the data with replacement each time I ran the algorithm). The 95% CI is calculated as the 2.5-97.5 percentiles of 1,000 random realisations of the algorithm.”

3.12 Algorithm implemented for spatial CFR prediction and residual estimation

The median CFR predictions were calculated from 1,000 realisations of steps 1 – 7 (page 81-82), followed by steps 8 to 14 below:

8. Use the BRT model built in step 4 to predict for all cases in the dataset.
9. Use the cut-off threshold p_T computed in step 6 to classify imputed cases as dead or alive.
10. Calculate the predicted CFR for individual cases.
11. Adjust the predictions for model imperfection (see section 3.10)
12. Calculate the median prediction for all individuals within a particular district (i).
13. Calculate the residual for each district. Residual (i) = Observed CFR (i) – Median predicted CFR all individuals in district (i). Observed CFR (ia) was estimated without adjusted imputation for cases with unknown survival outcomes and Observed CFR (ib) was estimated with adjusted imputation for cases with unknown survival outcomes. Thus, Residual(ia) and Residual (ib) were estimated.
14. Calculate the semi-variogram for these spatial residuals (i.e. ia and ib).

Confidence interval

Confidence intervals around the median CFR predictions and residuals were calculated by bootstrapping as previously described in section 3.7. Using the optimised hyperparameter set ($tc = 27$, $lr = 0.001$, $bf = 0.75$), the 95% CI of the CFR estimates with imputation for missing outcome were calculated as the 2.5-97.5 percentiles of 1,000 random (i.e. sampling the data each time with replacement) realisations of the algorithm.

3.13 Algorithm implemented for temporal CFR prediction and residual estimation

The median temporal predictions and residuals were calculated following steps 1–11 (page 83-84) followed by steps 1 to 3 below:

1. Calculate the median prediction for all individuals by quarter (i.e. three-month intervals).
2. Calculate the residual for each quarter. $\text{Residual}(q) = \text{Observed CFR}(q) - \text{Median predicted CFR per quarter}(q)$. Observed CFR (qa) was estimated without adjusted imputation for cases with unknown survival outcomes and Observed CFR (qb) was estimated with adjusted imputation for cases with unknown survival outcomes. Thus, Residual(qa) and Residual (qb) were estimated.
3. Calculate the semi-variogram for these temporal residuals (i.e. qa and qb).

Confidence interval

Confidence intervals around the median CFR predictions and residuals were calculated by bootstrapping as previously described in section 3.7. Using the optimised hyperparameter set ($tc = 27$, $lr = 0.001$, $bf = 0.75$), the 95% CI of the CFR estimates with imputation were calculated as the 2.5-97.5 percentiles of 1,000 random realisations of the algorithm.

3.14 Algorithms for data simulation

These algorithms follow closely those implemented in a preceding study [91]. However, the algorithms were adapted to compare the five machine learning models (logistic regression [LR], random forest [RF], boosted regression trees [BRT], Bayesian additive regression trees [BART] and artificial ANN) in a framework that ensured optimal hyperparameterisation of each model given the data.

3.141 Generating Missingness

To investigate the impacts of the type of missingness, I simulated missingness in the survival outcome and the predictors of survival. The simulations for the different types of missingness are described as follows:

Missing completely at random (MCAR)

For $M\%$ data missingness, I drew a number from a uniform distribution u for every value in the dataset, and for values less than $M/100$, I replaced survival outcome with NA (i.e. the missing value). I used this same approach to generate MCAR missingness in the predictors.

Missing at random (MAR)

For $M\%$ data missingness, for every value of u greater than $M/100$ (*days*), I replaced survival outcome with NA (i.e. the missing value).

Missing not at random (MNAR)

For $M\%$ data missingness, I fitted a logistic regression using all predictors and survival outcome as the dependent variable, for every value of u greater than $M/100$, I replaced survival outcome with NA (i.e. the missing value). Thus, making the missingness depend on the survival outcome itself.

I used Little's MCAR test [92] to differentiate between MCAR and non-MCAR missingness. I confirmed missingness was MAR by ensuring that I conditioned missingness in the survival outcome on reporting delay which is simulated as one of the observed predictors of outcome. I confirmed missingness was MNAR by ensuring that the missingness in survival outcome was conditional on the probability of the outcome itself.

The missingness in the predictors is simulated as MCAR in all scenarios investigated.

3.15 Algorithm implemented to investigate the performance of models as a function of outbreak data characteristics

For given values of parameters characterising the simulated outbreak data; type of missingness in the survival outcome (MCAR, MAR, MNAR), proportion of outbreak data size (p), proportion of simulated missingness (*missingness*) and the proportion of cases in training data, logistic regression (LR), random forest (RF), boosted regression trees (BRT), BART and ANN models were built and validated using the following algorithm:

1. Generate the proportion of outbreak data (p) from the complete simulated data.
2. Estimate the true CFR for this simulated outbreak dataset
3. Generate the specific type of missingness (MCAR, MAR or MNAR) (as described in section 1.1) and proportion of missingness (*missingness*) of the survival outcome but generate MCAR missingness in the predictor each time.
4. Generate the training data by randomly sampling without replacement the specified proportions of cases with death and recovery outcomes.
5. Generate the out-of-sample validation data, consisting of the cases with observed survival outcomes excluded from the training data (step 4).
6. Build the LR, RF, BRT, BART and ANN models using functions in the 'mlr' package, using the training data built in step 1, the specified type of missingness in the survival outcome (MCAR, MAR, MNAR), proportion of outbreak data (p), proportion of simulated missingness (*missingness*) and the proportion of cases in training data, 5-fold cross validation and 50 random search iterations of the hyperparameters of each model using survival outcome (i.e. death or recovery) as response predictor and

unexplained bleeding, fever, jaundice, fatigue, anorexia, vomiting, diarrhoea, headache, muscle pain, joint pain, chest pain, difficult breathing, conjunctivitis, confused, case classification, quarter, age, reporting delay, current hospitalisation status, country of origin as predictors. Optimal models minimising the holdout (cross validation) deviance are computed.

7. Use the optimal LR, RF, BRT, BART and ANN models built in step 6 to predict the survival outcomes of the cases in the validation data. For each subject i in the validation data, the model gives the probability p_i that subject i is dead.
8. Define the cut-off threshold p_T by which cases with $p_i > p_T$ are classified as dead or alive otherwise. I chose cut-off values giving equal sensitivity and specificity using the 'optimal.threshold' function in the 'PresenceAbsence' package [82].
9. Use the cut-off threshold p_T computed in step 8 to classify cases in the validation data as dead or alive.
10. Compute the sensitivity (proportion of deaths correctly classified), specificity (proportion of survival correctly classified) and percentage of percentage correctly classified in the validation data (PCC) and the area under the receiver operating characteristic curve (AUC).
11. Use the optimal models built in step 6 to calculate the survival probability of each case.
12. Use the cut-off threshold p_T computed in step 8 to convert the survival probabilities to binary outcomes (i.e. death or recovery)

13. Calculate the CFR estimates and corresponding 95% confidence intervals using the cases with both imputed and observed survival outcomes.

The choice of optimising the cut-off thresholds to have equal sensitivity and specificity in step 5 was made to avoid introducing bias in CFR estimated with imputation.

3.16 Algorithm implemented for CFR adjusted for bias in the models

CFR estimates and confidence intervals adjusted for bias in the models (i.e. LR, RF, BRT, BART and ANN) were calculated from 100 realisations of steps 1 – 9, followed by steps 14-17 below:

14. Generate cases with just the imputed death outcomes.
15. Use the dead cases in step 9, the total number of cases in the imputed data and tp function in the 'RSurveillance' package to estimate the CFR adjusted for model sensitivity and specificity [90].
16. Multiply the number of cases in the imputed data by the inferred CFR to get the number of inferred deaths f_T .
17. The adjusted CFR is the ratio of the sum of f_T and the number of deaths in data with known survival outcomes, divided by the sum of all deaths and survivals in the complete data (i.e. data with both imputed and observed outcomes).

Chapter 4: Case fatality ratio estimates for the 2013 – 2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation

Work in this chapter forms the basis of **Forna A**, Nouvellet P, Dorigatti I, Donnelly CA. Case fatality ratio estimates for the 2013 – 2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation. Clinical Infectious Diseases 2019.

In this chapter, I used a relatively simple boosted regression tree (BRT) model to impute for unreported survival outcomes, re-estimate CFRs from the dataset with the imputations and estimate adjusted CFRs using the estimated sensitivity and specificity of the BRT model.

4.1 Introduction

Ebola virus disease (EVD), a haemorrhagic fever disease of zoonotic origin, has been detected in several outbreaks in African countries since it was first identified in 1976 [3]. While outbreaks have typically been of relatively limited size and confined to rural Central Africa, the 2013-2016 mostly urbanised epidemic in West Africa took the world by surprise and showed the potentially devastating impact of the Ebola virus [3].

The case fatality ratio (CFR), the proportion of EVD cases that died [26], remains a challenging quantity to estimate due to undetected cases, incomplete data on detected cases and the lack of laboratory confirmation of probable and suspected cases. CFR estimates may vary because of unreported survival outcomes, time of case diagnosis and the statistical methods used for estimation [93, 94]. For instance, 14,694 (44%) of the survival outcome data for confirmed, probable and suspected cases are unreported in the dataset analysed by the WHO Ebola Response Team [21]. Using clinical, laboratory and field surveillance data, I imputed unreported outcomes based on models fitted to observed outcome data. Imputed outcomes have previously been used in other contexts to generate

baseline estimates of CFR necessary in evaluating interventions and designing randomised control trials [95].

Regression modelling has been used widely to predict EVD survival outcomes [17, 78, 96].

For instance, after the first nine months of this epidemic, logistic regression was used to identify predictors of CFR [17]. It was found that age, sex, country, and fever were correlated with CFR [17]. However, much of the observed heterogeneity remained unaccounted for. Thus, there is a need to use methods which can account for multiple predictors and their interactions as well as limitations of the data available.

Machine learning (MaLe) techniques are powerful alternatives to conventional statistical regression models [58]. In MaLe, less restrictive algorithms are used to learn the relationship between outcome and proposed predictors [58]. Because of this flexibility, MaLe techniques are generally very effective in predicting outcomes at the individual level. Boosted Regression Trees (BRT), a well-characterised MaLe technique, have been used in infectious disease modelling for data imputation [65]. The predictive performance of BRT, its ability to identify influential predictors and interactions and its ability to model non-linear relationships, makes it ideal to answer key questions in infectious disease epidemiology. For instance, Bhatt *et al.* used BRT to map the global distribution and burden of the dengue virus [65]. Similarly, Dorigatti *et al.* used BRT to refine the efficacy estimate of the Sanofi Pasteur dengue vaccine CYD–TDV [97] illustrating the effectiveness of BRT in estimating the uncertainty around missing or unrecorded data. However, despite these attempts, MaLe models are still seldomly used in the analysis and modelling of infectious diseases.

Here, I used BRT to identify, using a non-parametric approach, the predictors of CFR for the West African epidemic, imputed unreported survival outcomes and combined the imputed

and reported survival outcomes to re-estimate CFRs for the epidemic. Finally, I compared the CFR estimates obtained without data imputation with estimates obtained including data imputation and estimates adjusted for model imperfection.

4.2 Methods

4.2.1 Data sources

I used case-level data that were reported to the WHO using the viral haemorrhage fever (VHF) case reporting forms as published in the supplementary data of Garske *et al.* [21]. On the case reporting forms, cases were defined as confirmed, probable and suspected, using the WHO EVD case classification system described in chapter 2, section 2.4 [98].

Our main analyses present CFR inference based on confirmed, probable and suspected cases, while analyses using either ‘confirmed’ cases or ‘confirmed and probable’ cases are presented in the sensitivity analysis (section 4.6). A wealth of information was recorded for each case including demographic (e.g. age, country) and clinical (e.g. fever, diarrhoea, bleeding) data (see Chapter 2 for details).

The survival outcome for this dataset was reported as the “final status” of cases. Final status was defined as alive, Dead or Missing. The fate of cases with missing final status was unknown e.g. cases could have been alive in the communities or later died in the communities but were not reported to, or recorded by, the appropriate authorities. Cases with known final outcomes accounted for 18,644 (56%) of the cases in the dataset.

As mentioned in Chapter 2, I selected 43 candidate predictors of CFR. This selection was informed by the literature [17, 99] as well as judgement on the influence of potential predictors on CFR.

4.2.2 Statistical methods

BRT relies on decision trees and gradient boosting. With regression trees, a recursive binary split of data is carried out until a stopping criterion is reached [58]. Used alone for analysis, regression trees are susceptible to bias and inaccuracy. The gradient boosting applies randomness into the stage-wise fitting which, by introducing stochasticity, helps to avoid overfitting of the model to the data (See Chapter 3, section 3.6 for details).

To fit the BRT model, I used the survival outcomes as the dependent variable and the potential predictors as the explanatory variable.

I evaluated the model performance in terms of multiple measures of predictive accuracy, i.e. sensitivity, specificity, proportion of predictions correctly classified and the area under the receiver operating characteristic (ROC) curve (AUC). The performance measures quantify the ability of the BRT model to discriminate between outcomes (death/recovery) for the given predictors [100]. For AUC, performance values of 0.5 (50%) indicate random discrimination of outcomes, and performance values of 1 (100%) indicate perfect discrimination of outcomes. These performance measures were used to determine the optimal hyperparameters.

To identify the model hyperparameters [tree complexity (tc), learning rate (lr) and bag fraction (bf)], multiple BRT models were fitted using different parameter combinations. In total, I explored 72 hyperparameter combinations. Figure 4.1 presents a summary schematic of the analysis.

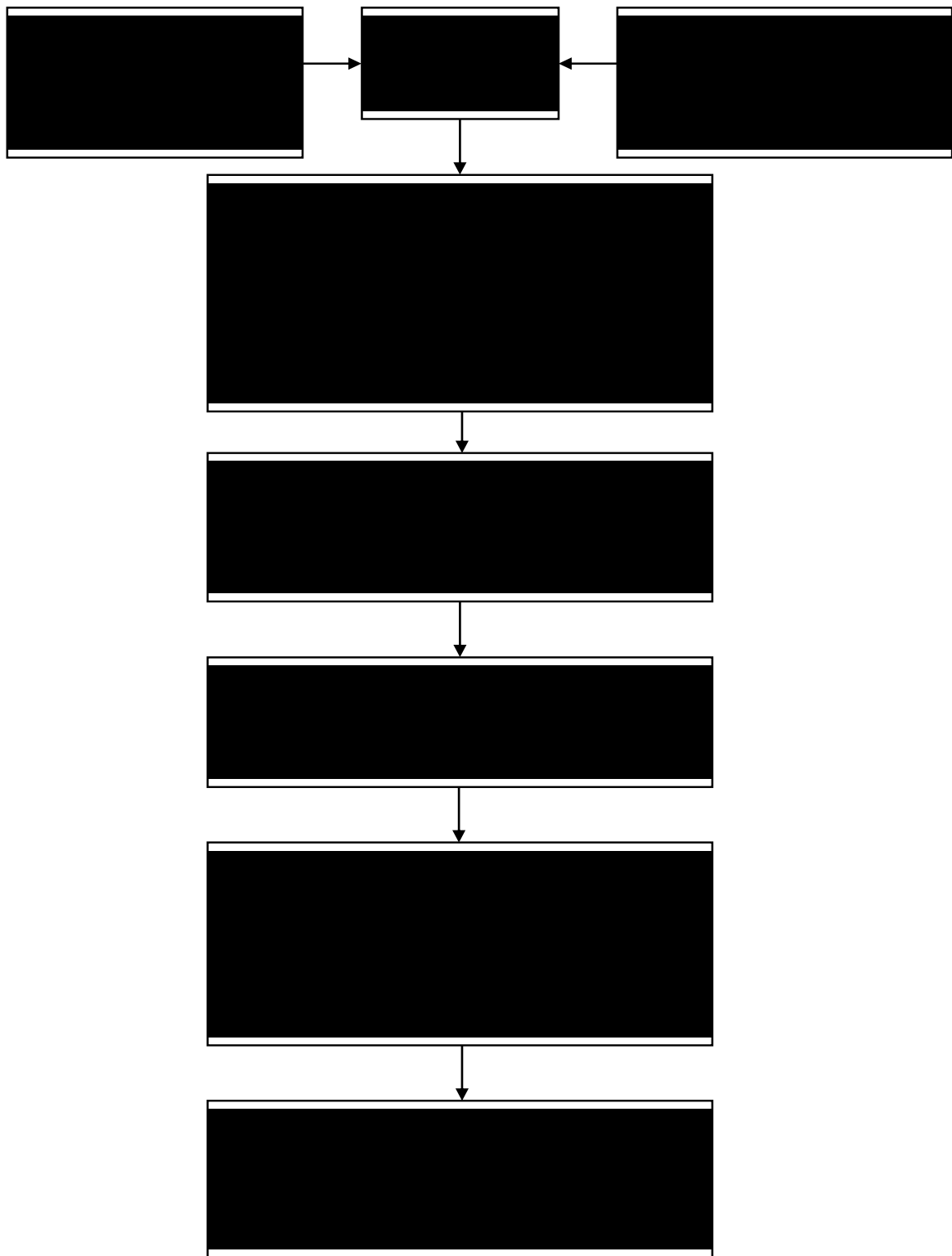


Figure 4. 1: Schematic summary of the analysis steps used in this chapter.

BRT models return probabilities for each outcome. Converting these probabilities into binary values requires a threshold probability [101]. I selected our threshold to obtain equal sensitivity and specificity. It allowed us make predictions without trading off death and survival outcomes.

I simplified the BRT model by excluding non-informative predictors with minimal effect on prediction [58]. The relative influence (RI) of a predictor is the number of times the predictor is selected to discriminate outcomes (death/recovery) and the improvement to the model for each selection, which is based on the number of residuals explained at each selection [102]. These RIs are scaled to sum to 100 and higher influences indicate greater effects on survival outcome. The model was repeatedly refitted, each time removing the predictor with the lowest RI. I applied a default rule that continued to remove predictors with the lowest RIs until the average change in predictive deviance exceeded the standard error of the full model (i.e. model using all candidate predictors) [58].

The simplified model was used to predict the survival outcomes for the 44% of cases in the dataset with missing outcome and the CFRs were re-estimated for the full dataset. Using nonparametric bootstrapping with replacement for 1,000 realisations, I calculated the means and 95% confidence intervals of the re-estimated CFRs (see Chapter 3, section 3.9).

CFR estimates with imputation ignore the bias arising from the imperfect sensitivity and specificity of the imputation algorithm. I, therefore, corrected our CFR estimate using the estimated sensitivity and specificity of the BRT model to obtain unbiased estimates [90]. The inferred CFR was calculated for each subset of interest according to the formula,

$$\text{Inferred CFR} = \frac{(q + \text{specificity} - 1)}{(\text{sensitivity} + \text{specificity} - 1)}, \quad (2)$$

where q is the CFR for the imputed data (i.e. data with unknown outcomes) using the estimated sensitivity and specificity of the BRT model [103, 104].

We can derive equation (2) as follows:

Let q denote the observed CFR for the imputed data and c denote the inferred CFR.

Let Se be the sensitivity and Sp be the specificity.

So, the observed CFR for the imputed data is a combination of truly detected deaths ($Se * c$) and falsely detected deaths $(1-Sp) * (1-c)$:

$$q = Se * c + (1-Sp) * (1 - c)$$

Multiplying through we get:

$$q = Se * c + 1 - c - Sp + Sp * c$$

Moving all terms not involving c to the left-hand side and keeping all the terms involving c on the right-hand side we get:

$$q - 1 + Sp = c (Se + Sp - 1)$$

Then dividing by $(Se + Sp - 1)$ we get:

$$c = (q + Sp - 1) / (Se + Sp - 1)$$

Thus, the final CFR was calculated after adjusting the CFR for missing outcomes for imputation sensitivity and specificity (Chapter 3, section 3.10).

Equation (2) is unstable when the sum of sensitivity + specificity is close to 1; and the denominator in (2) is therefore close to zero. To account for this in my analysis, inferred CFR estimates <0 and >1 were excluded from the analysis. Another approach to account for the

instability is to only retain bootstrap estimates which result in the sum of sensitivity + specificity being substantially greater than 1 (with a count kept of the retained number). While this approach was not implemented in this thesis, it could be investigated in future work.

4.3 Results

In the dataset, a total of 18,644 (confirmed, probable and suspected) cases with known outcomes were reported out of 33,338 cases total (Table 4.1). The proportions of cases with dead, alive, and unknown survival outcomes vary across age classes and reporting delay categories (Figure 4.2) with statistically significant differences indicating missingness departed from random (Table 2.2).

Additional results (Figure 4.4) show the out-of-sample predictive performance of the full BRT model (i.e. fitted using all available predictors) using the 72 tested hyperparameter configurations. Further analyses were based on the following optimal hyperparameters: data partitioning ratio ($p=0.65$), tree complexity ($tc=27$), learning rate ($lr=0.001$) and bag fraction ($bf=0.75$). In the simplified model, I retained 24 predictors (out of the 43 originally included in the full model), accounting for 96% of the RI of the full model (i.e. the model with all candidate predictors).

District, hospitalisation status, age, case classification and quarter were the most important predictors with average RIs of 41.7%, 28.9%, 10.7%, 10.1%, and 2.2%, respectively (Table 4.2). Unadjusted CFR estimates (i.e. CFR estimates obtained with imputation not corrected for imperfect sensitivity and specificity) and CFR estimates without imputation initially decreased with age, with the lowest CFR estimates found in the 15-19 years age group with

CFR of 57.3% (95% CI: 39.4%-68.2%) and 57.2% (95% CI: 48.6%-65.4%) unadjusted and without imputation, respectively (Figure 4.3). Thereafter, CFR increased steadily with the highest rates in 75+ years age group with CFR of 90.9% (95% CI 75.8%-96.3%) and 93.4% (95% CI 88.8%-97.2%) unadjusted and without imputation, respectively (Figure 4.3).

CFR estimates increased with increasing reporting delays. The lowest rates were at 0 to 5 days, with CFR of 71.5% (95% CI 54.1%-79.2%) and 74.4% (95% CI 72.3%-76.5%) unadjusted and without imputation, respectively. The highest rates were at 16 to 21 days, with CFR of 78.5% (95% CI 63.6%-88.0%) and 82.0% (95% CI 70.6%-90.3%) unadjusted and without imputation, respectively.

Figure 4.3 shows the overall and country level estimates. Overall CFR estimates were 71.9% (95% CI 56.1%-79.8%) and 75.1% (95% CI 73.5%-76.9%) unadjusted and without imputation, respectively (Table 4.1, Figure 4.3). Guinea had the lowest rates, with CFR of 65.6% (95% CI 61.3%-69.6%) and 65.8% (95% CI 61.6%-69.9%) unadjusted and without imputation, respectively. The Liberian rates were higher, with CFR of 69.7% (95% CI 55.6%-78.7%) and 71.7% (95% CI 67.2%-75.6%) unadjusted and without imputation respectively. Sierra Leone had the highest rates with CFR of 74.6% (95% CI 52.2%-84.3%) and 81.3% (95% CI 79.3%-83.3%) unadjusted and without imputation, respectively (Table 4.1, Figure 4.3).

Fever, anorexia, difficult breathing and fatigue are the clinical predictors in the minimal model (Table 4.2). The unadjusted CFR estimates were 69.3% (95% CI 50.7%-78.2%) and 82.1% (95% CI 67.1%- 88.7%) for cases with fever and cases without fever, respectively (Figure 4.3). The CFR estimates without imputation were 70.7% (95% CI 68.6%-72.7%) and 87.5% (95% CI 84.6%-90.3%) for cases with fever and cases without fever, respectively (Figure 4.3). Fever, anorexia, difficult breathing and fatigue together accounted for 4.7% of

the RI of the minimal model, defined as the model using the ten most influential predictors (district, hospitalisation status, reporting delay, age, quarter, case classification, anorexia, difficult breathing, fatigue, and fever, see Table 4.2).

Adjusted CFRs (i.e. corrected for imperfect sensitivity and specificity) with imputation for all predictors were broadly higher compared to unadjusted estimates and estimates without imputation (Figure 4.3). For instance, adjusted CFR was 82.8% (95% CI 45.6%-85.6%) overall and 65.6% (95% CI 61.3%-69.6%), 79.2% (95% CI 45.4%-84.1%) and 89.1% (95% CI 40.8%-91.6%) for, Guinea, Liberia, and Sierra Leone, respectively (Table 4.1).

In Table 4.3, I found that the full (i.e. using 43 predictors), simplified (i.e. using 24 predictors) and the minimal (i.e. using 10 predictors) models produced comparable performances, for instance, the AUC of the full model was 76.7% (95% CI 61.0%-82.7%), that of the simplified model was 76.0% (95% CI 56.8%-82.1%) and that of the minimal model was 75.7% (95% CI 56.1%-82.1%).

Table 4. 1: CFR estimates without imputation, unadjusted with imputation and adjusted with imputation for confirmed, probable and suspected cases. All CFR estimates and corresponding CIs were calculated using a non-parametric bootstrap of the BRT model. * Adjusted CFR estimated using tp function to correct for bias in BRT model performance.

	Cases in the dataset without imputation	Cases in the dataset including those with imputation	CFR without imputation (95% CI)	Unadjusted CFR with imputation (95% CI)	Adjusted CFR with imputation (95% CI)
Guinea	3,740	3,757	65.8 (61.6-69.9)	65.6 (61.3-69.6)	65.6 (61.3-69.6)
Liberia	4,624	8,130	71.7 (67.2-75.6)	69.7 (55.6-78.7)	79.2 (45.4-84.1)
Sierra Leone	10,280	21,451	81.3 (79.3-83.3)	74.6 (52.2-84.3)	89.1 (40.8-91.6)
Overall*	18,644	33,338	75.1 (73.5-76.6)	71.9 (56.1-79.8)	82.8 (45.6-85.6)

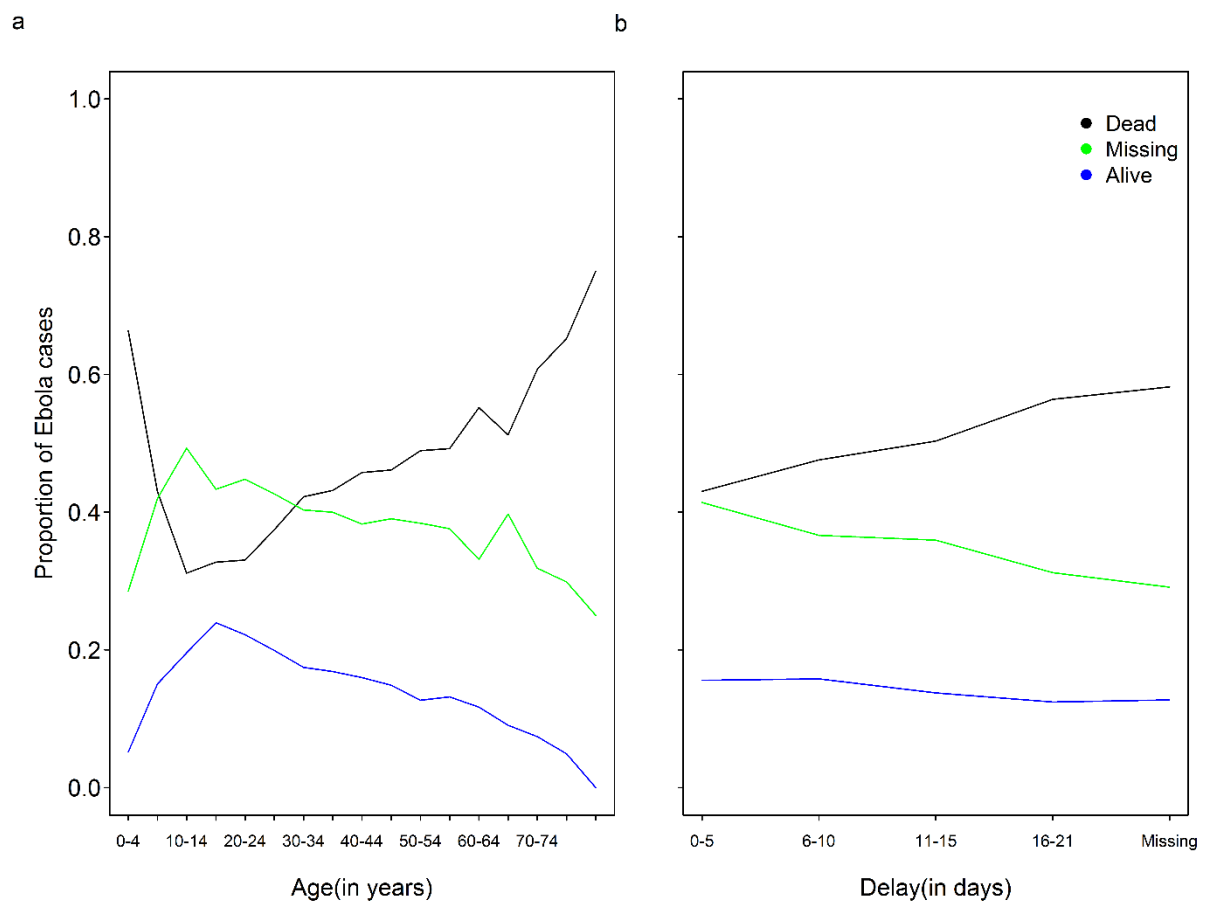


Figure 4. 2: Proportion of known survival outcomes (i.e. Dead and Alive) and unknown survival outcomes (i.e. Missing) for 'confirmed, probable and suspected' cases. a: Proportion of deaths, survivals, and entries with unknown outcome by age group (in years). b: Proportion of deaths, survivals, and entries with unknown outcome by reporting delay (in days).

Table 4. 2: Relative contributions (%) of the predictors in the minimal BRT model using 10-fold cross validation, $tc = 27$, $lr = 0.001$ and $bf = 0.75$ and trained on 1000 training sets generated by randomly sampling 65% cases with known survival outcomes without bootstrapping. The minimal model used these 10 predictors.

Relative Contribution	
Predictors	(%)
District of origin	41.7
Hospitalisation status	28.9
Age	10.7
Case Classification	10.1
Quarter	2.2
Delay	1.8
Anorexia	1.6
Difficult Breathing	1.6
Fever	1.0
Fatigue	0.5

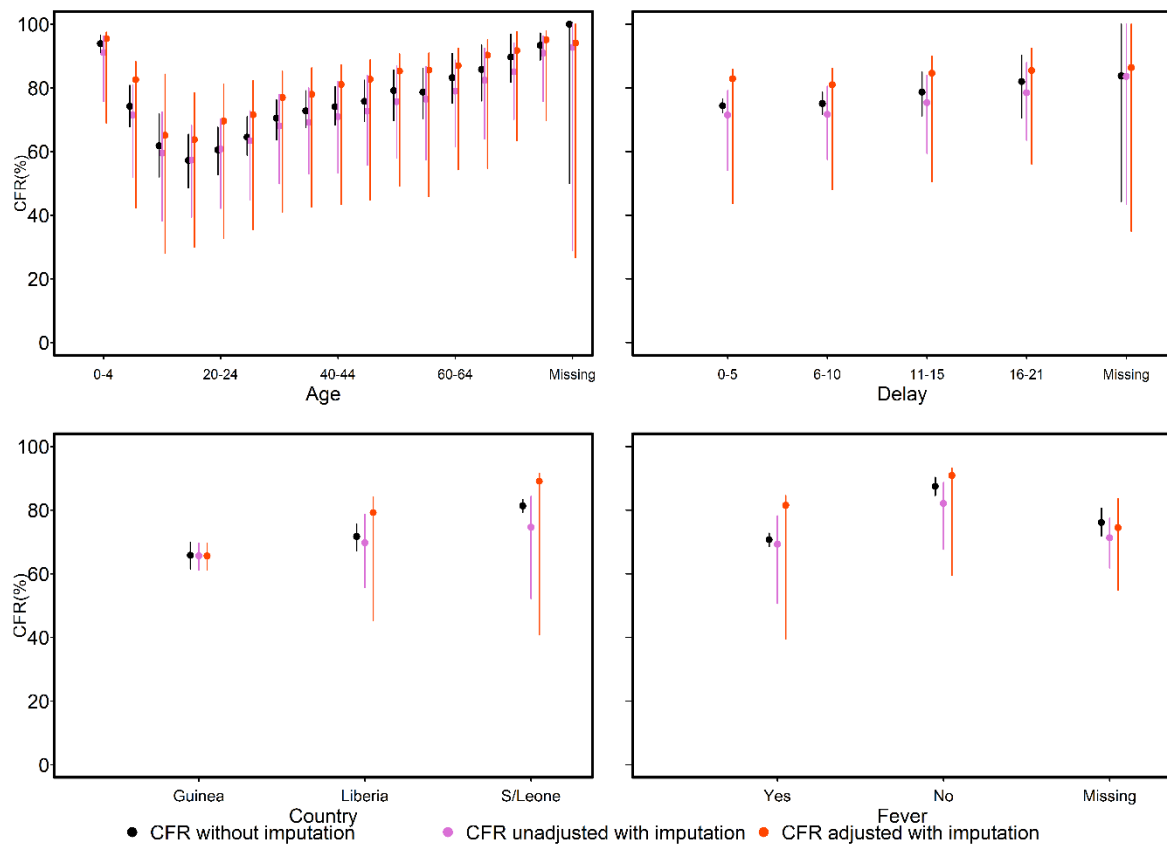


Figure 4. 3: CFR by age, delay, country, and fever. Median and 95% confidence intervals (CI) are plotted (based on 1000 bootstrap realisations for ‘confirmed, probable and suspected cases’).

Table 4. 3: BRT model performance. ^a the full BRT model used all 43 candidate predictors. ^b the simplified BRT model used the 24 predictors retained after model simplification. ^c the minimal BRT model used the ten most important predictors. Median and 95% confidence intervals (CI) reported, based on 1,000 bootstrap realisations using confirmed, probable, and suspected cases.

Performance		Modal Performance	
Measures	Full Model ^a with	Simplified Model ^b	Minimal Model ^c with
	bootstrap	with bootstrap	bootstrap
	Median (%)	Median (%) (95% CI)	Median (%) (95% CI)
	(95% CI)		
Sensitivity	70.5 (56.0-76.5)	69.7 (52.5-75.6)	69.7 (51.7-75.7)
Specificity	70.5 (56.8-75.9)	69.8 (54.1-75.6)	69.8 (51.2-75.6)
PCC	70.5 (56.6-75.9)	69.9 (53.7-75.5)	69.7 (51.0-75.4)
AUC	76.7 (61.0-82.7)	76.0 (56.8-82.1)	75.7 (56.1-82.1)

4.4 Discussion

I have used BRT to impute for unreported outcomes and produced improved (i.e. representative of all cases and adjusted for BRT imputation performance) CFR estimates for the 2013–2016 West African Ebola epidemic. To our knowledge, the dataset used in this study (33,338 cases) is the largest case-level data used to characterise the CFR of EVD.

The effects of BRT's hyperparameters on the survival outcome was consistent with those in the published literature [58], with AUC >70% and <90% indicating appreciable predictive performance [86]. The data to which I fitted the BRT model were large relative to typical epidemiological datasets in outbreak settings. However, our method proved robust to changes in sample sizes, with comparable mean imputation performances reported even after data down-sampling (additional results Figure 4.21).

Model simplification suggested that district, hospitalisation status, age, case classification, quarter, reporting delay, anorexia, difficult breathing, fever, and fatigue were the most important predictors retained in the model. Most bleeding predictors (except for unexplained bleeding) were removed, which corroborates the existing literature suggesting that external bleeding was not a prominent characteristic of the West African Ebola epidemic [105-107]. Consequently, the predictive value of bleeding predictors was limited compared to the other (e.g. age, reporting delay, district, and fever) predictors of CFR. Simplification removed some clinical predictors, but I found that the significance of clinical predictors varied with sample size and the subset of data included in the analysis. For example, hiccups occurrence was excluded from the final (simplified) model in our analysis but was found to have a predictive value in an earlier study conducted in Sierra Leone [105], although larger sample sizes would be needed to validate the significance of those results

[105]. Earlier analyses of the West African epidemic showed that fever was not a significant predictor of the CFR [17, 21], while, in contrast, our study identified fever as the ninth most important predictor of the CFR. Further investigations revealed that the occurrence of fever only became a significant predictor in analyses using the datasets compiled after 25th November 2014 (additional results Table 4.5). From this finding, I cannot determine whether the emergence of fever occurrence as an important predictor of the CFR is related to changing clinical presentation of the EVD over time or was caused by changes in reporting practices throughout the epidemic. The high CFR in patients without fever might also suggest immune evasion [108, 109]. Interestingly, survival outcomes were better for patients with anorexia (see additional results), perhaps because these patients were better medically attended.

The predictors retained in the simplified model were broadly similar to those reported in the literature and the identified clinical predictors reflected the symptomatic profile of EVD cases during the West African epidemic and in previous outbreaks [17, 105, 110].

Demographic predictors of age, case classification, reporting delay, quarter and district were important predictors of CFR, which broadly agrees with discussions in the literature suggesting that notwithstanding the clinical presentations of EVD, demographic factors were predictors of EVD fatalities [17, 111, 112]. Age-dependent CFR variation, consistent with previous studies [78, 95], suggests a biological effect. Potential biological reasons include variation in the immunocompetency of children and young adults, and the increasing likelihood of comorbidities in the aged [108, 113]. CFR increased with increasing reporting delay, suggesting that early case reporting could have improved survival outcomes, although previous findings have suggested that early reporting was insufficient to increase survival probability [78]. A clinical consideration is that people who were reported

late may have been more likely to die because their symptoms persisted while those with milder symptoms would have recovered earlier and were less likely to report as Ebola patients.

While adjusted CFRs were broadly higher than uncorrected CFRs, they corroborate with estimates of the Ebola Zaire CFR, from 1979 to 2014, compiled by Van Kerkhove *et al.* [3]. However, adjusted CFRs must be interpreted with appropriate caution, as confidence intervals are wide. Results validating the sensitivity and specificity adjustments are reported in the additional results of this chapter.

Country level CFR estimates were consistent for Guinea, which had the highest proportion of complete outcome data (99.5% i.e. 3,740 out of 3,757 cases, see Table 4.1) among the three countries. The relatively high adjusted CFR estimates obtained in Liberia and Sierra Leone compared to estimates without and with imputation, and the associated relatively wide confidence interval could reflect the true CFR in these two countries but perhaps also the changing data collection practices over time. For instance, Liberia switched from an aggregate system of reporting to an individual-based system of reporting that could have affected data curation at the country level [77]. Potential biases may arise because definitive outcomes are more frequently reported for hospitalised cases, who this study shows were less likely to die from EVD infection [26]. Adjusted CFRs account for the estimated sensitivity and specificity of the BRT imputation model and the dependence of missingness on characteristics of the cases (e.g. age and reporting delay).

These robust CFR estimates in conjunction with other resources, can be used to inform public health contingency planning for future Ebola epidemics, for instance, to allocate resources and evaluate the effectiveness of interventions [114]. In epidemic settings with

limited resources, as was the case in West Africa, reliable CFR estimates are essential for planning hospital care and managing cases [115].

While previous studies show clinical characteristics indicative of death to have a profile that could be used in subsequent case triaging [18, 116], the model built in this study cannot be used in such manner as it lacked clinically significant predictors, such as viral load, which were not captured in the dataset. Rather the output of our analyses should be used to inform resource allocation and care planning. Also, our analysis provides a baseline CFR against which the CFR observed with interventions (such as a new treatment) could be assessed. As a known predictor of EVD CFR [95, 117, 118], viral load data, where available, should be used in future attempts to further refine CFR estimates of EVD. If some survival outcomes were incorrectly recorded during data collection and/or some cases were missing entirely from the database, then in the absence of at least some gold standard data, no estimation strategy can fully account for such limitations and this may be in part responsible for the imperfect nature of the imputation process.

There are a few methodological approaches used in this work that could be modified in future work to further improve the precision of the CFR estimates. While I chose to convert BRT probabilities to binary outcomes, BRT probabilities could also be used to randomly sample outcomes, as part of the bootstrap process. To account for the instability in estimating inferred CFR, I excluded estimates <0 and >1 . While this pragmatic approach was reasonable, only retaining bootstrap estimates which resulted in the sum of sensitivity + specificity being substantially greater than 1 could have further improve the precision of the CFR estimates. Finally, even though BRT “discovers” the relationship between CFR and

predictors like age, using prior subject-matter knowledge, simple methods like logistic regression could be parametrised to characterise such relationships.

In conclusion, I have demonstrated that BRT modelling can be used to obtain improved estimates of CFR where outcome data are unreported. This BRT modelling framework could be implemented to further investigate EVD CFR heterogeneity and adapted to other infectious disease epidemiological analyses based on datasets with missing data.

4.5 Additional Results

4.5.1 Identification of hyperparameters

I created 72 hyperparameter combinations from tree complexities varying from 5 to 40 in steps of 5, learning rates of 0.01, 0.001 and 0.0005 and bag fractions from 0.5 to 1 in steps of 0.25. Additionally, to optimise the data partitioning for model training and validation, partitioning ratios of 0.5, 0.65 and 0.8.

Figure 4.4 shows prediction accuracy on the validation data which is a measure of out-of-sample predictive performance and Figure 4.5 shows the prediction accuracy on the training data which is a measure of the

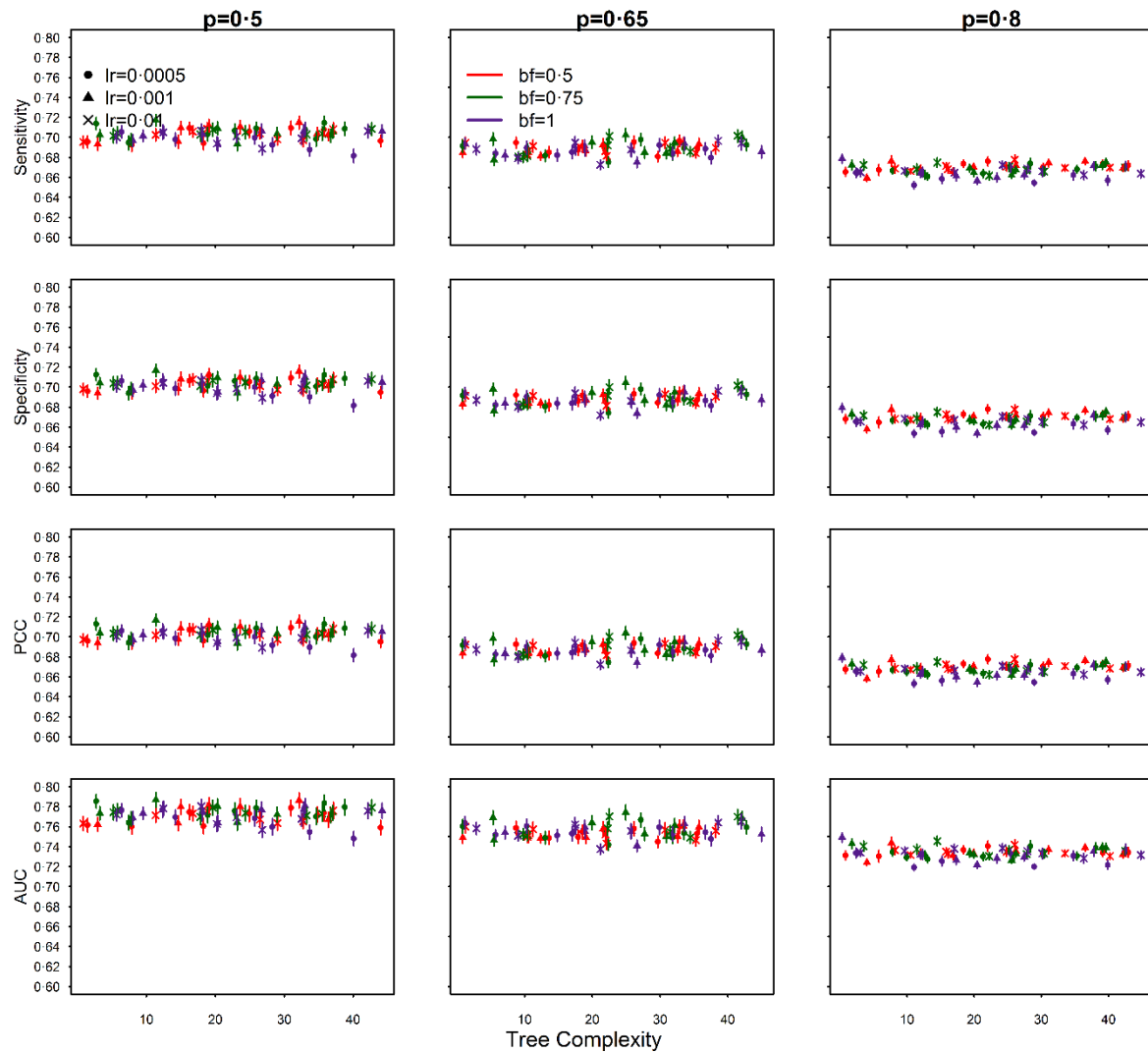


Figure 4. 4: Out-of-sample predictive accuracy measures (sensitivity, specificity, proportion of percentage correctly classified (PCC) and area under the receiver operating curve (AUC)) on the validation data obtained using 72 different hyperparameterisations of the BRT model. p represents the proportion of data to train the model with $p=0.5$, $p=0.65$ and $p=0.8$ explored.

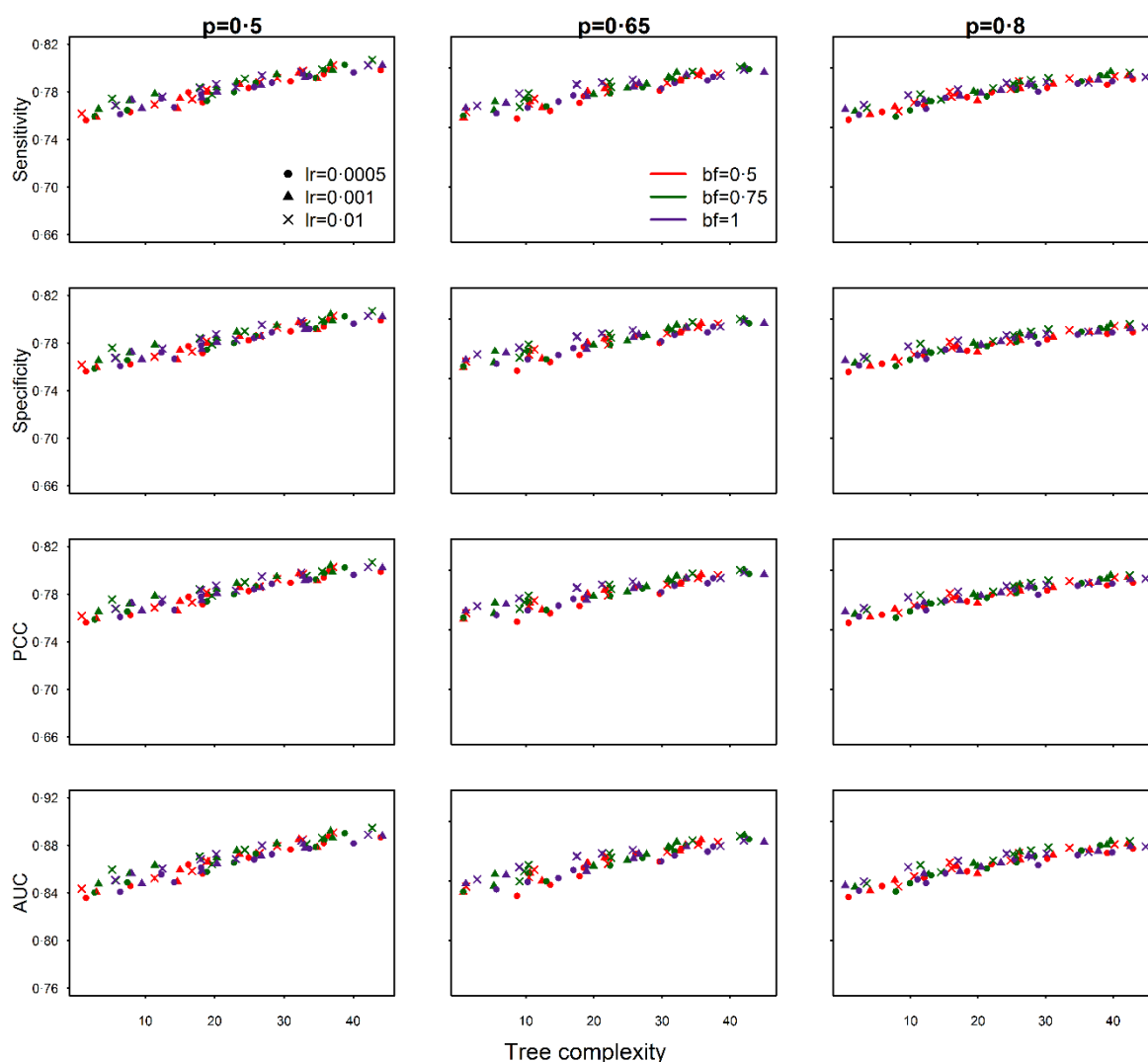


Figure 4. 5: Within-sample accuracy measures (sensitivity, specificity, proportion of percentage correctly classified (PCC) and area under the receiver operating curve (AUC)) on the training data obtained using 72 different hyperparameterisations of the BRT model. p represents the proportion of data to train the model with $p=0.5$, $p=0.65$ and $p=0.8$ explored.

4.5.2 Comparison of CFR estimates obtained on the training set, without and with imputation

The goodness-of-fit performance (i.e. making a prediction using the data on which the model was built) on the training data was very good for all predictors and all case classification combinations as confidence intervals neatly overlap (Figures 4.6—4.11).

4.5.2.1 Confirmed, probable, and suspected cases

CFR for age, delay, country, and fever

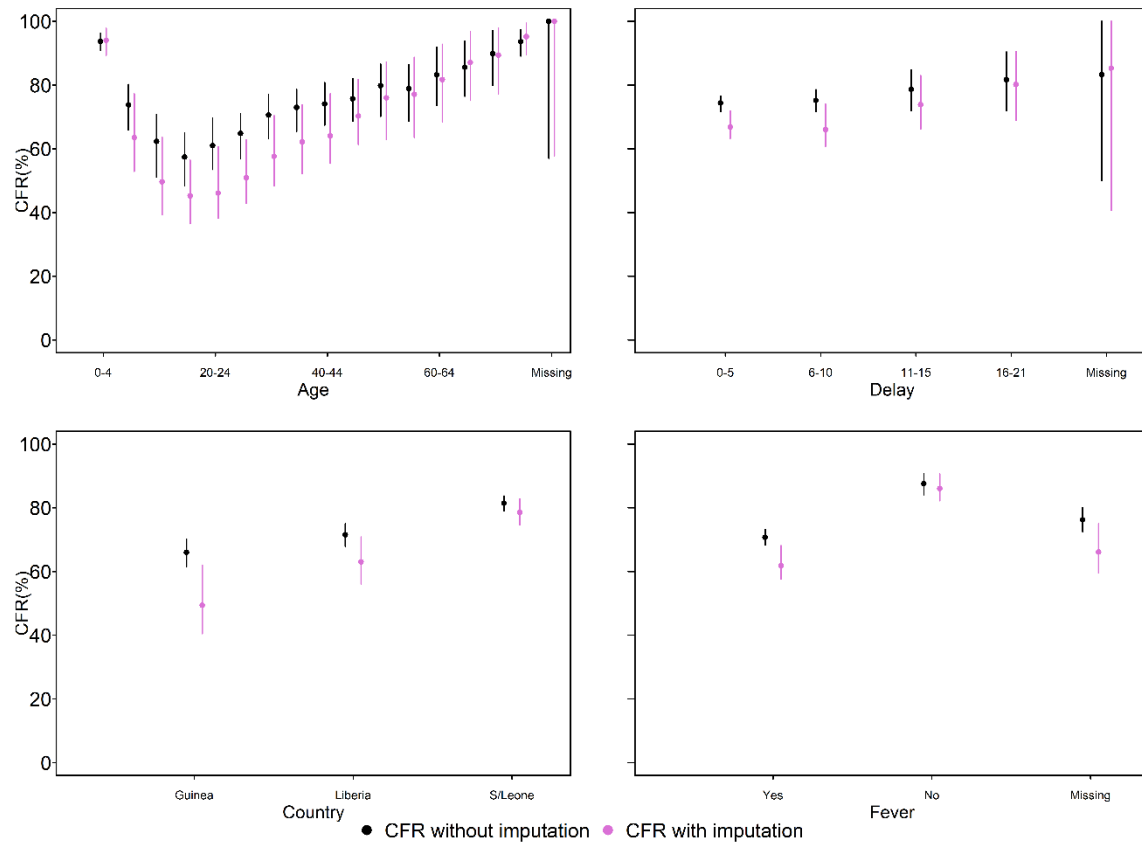


Figure 4. 6: CFR estimates without (black) and with (orchid) imputation for age, delay, country and fever, for confirmed, probable and suspected cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

CFR for other Predictors

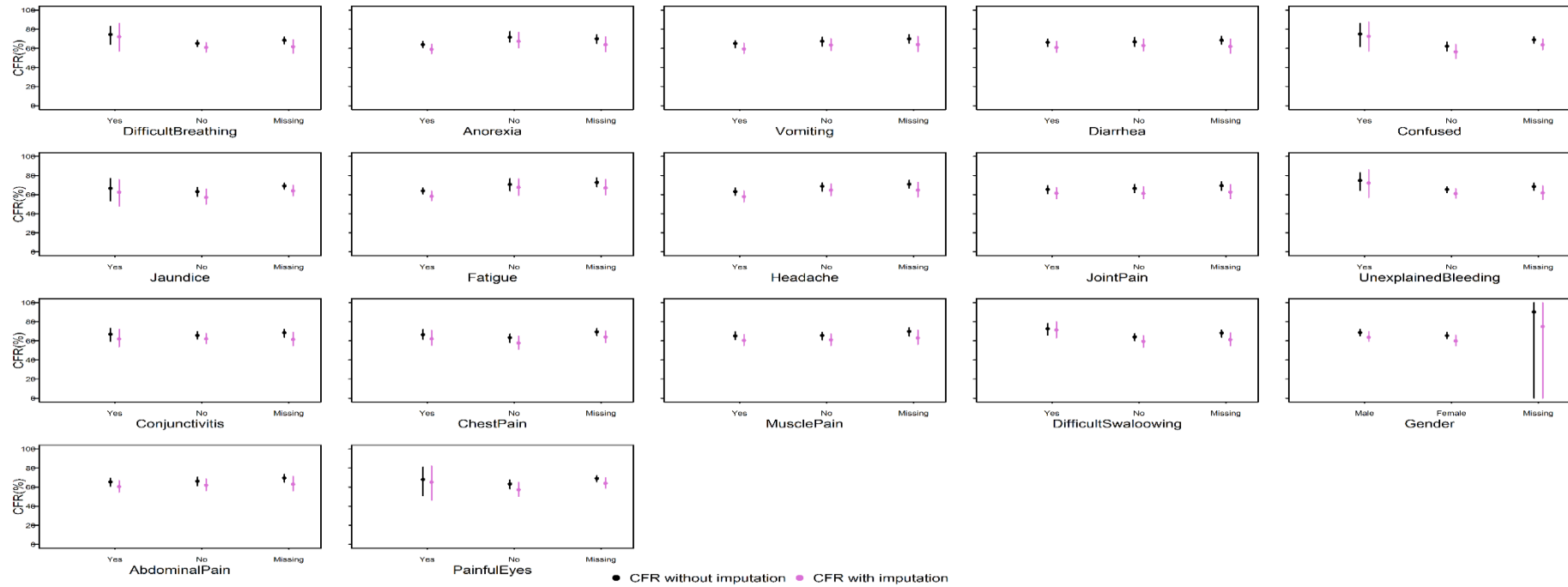


Figure 4. 7: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed and probable cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

4.5.2.2 Confirmed and probable cases.

CFR for age, delay, country, and fever

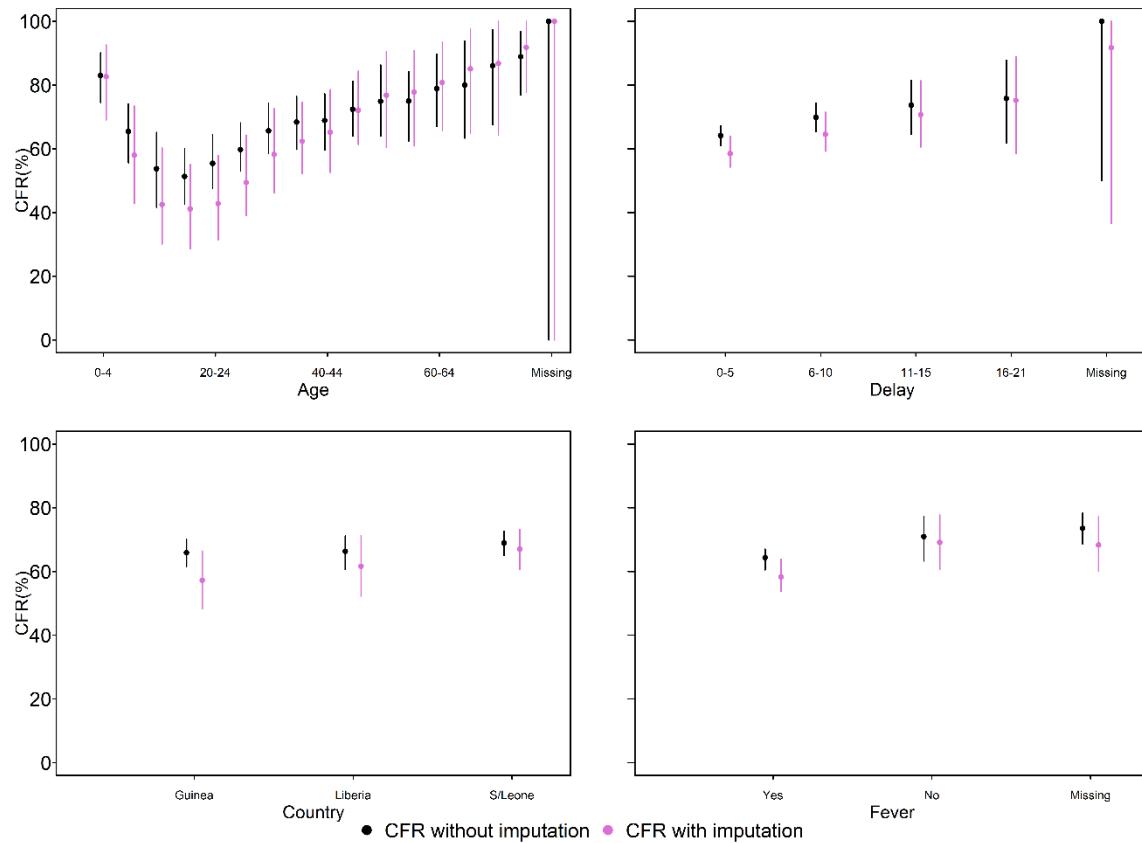


Figure 4. 8: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed and probable cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

CFR for other Predictors

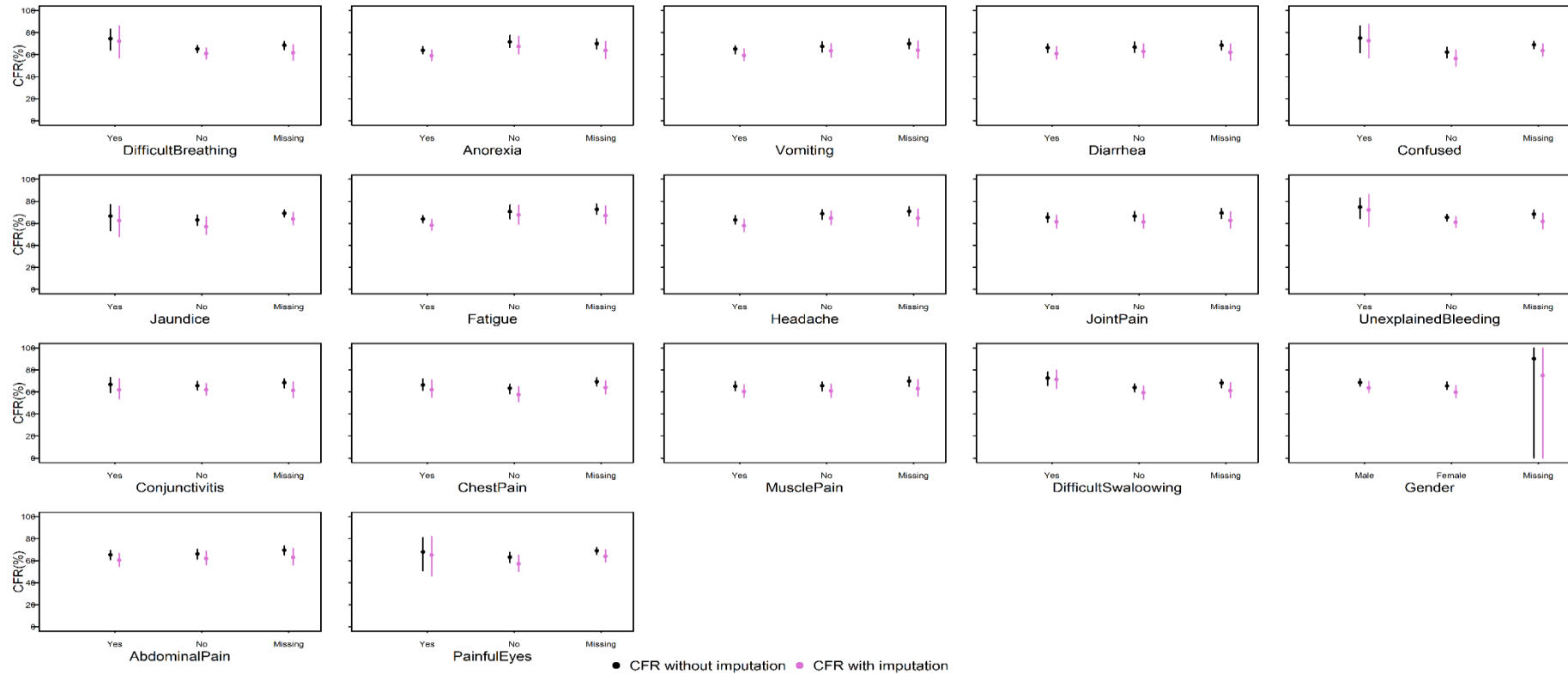


Figure 4. 9: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed and probable cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

4.5.2.3 Confirmed cases.

CFR for age, delay, country, and fever

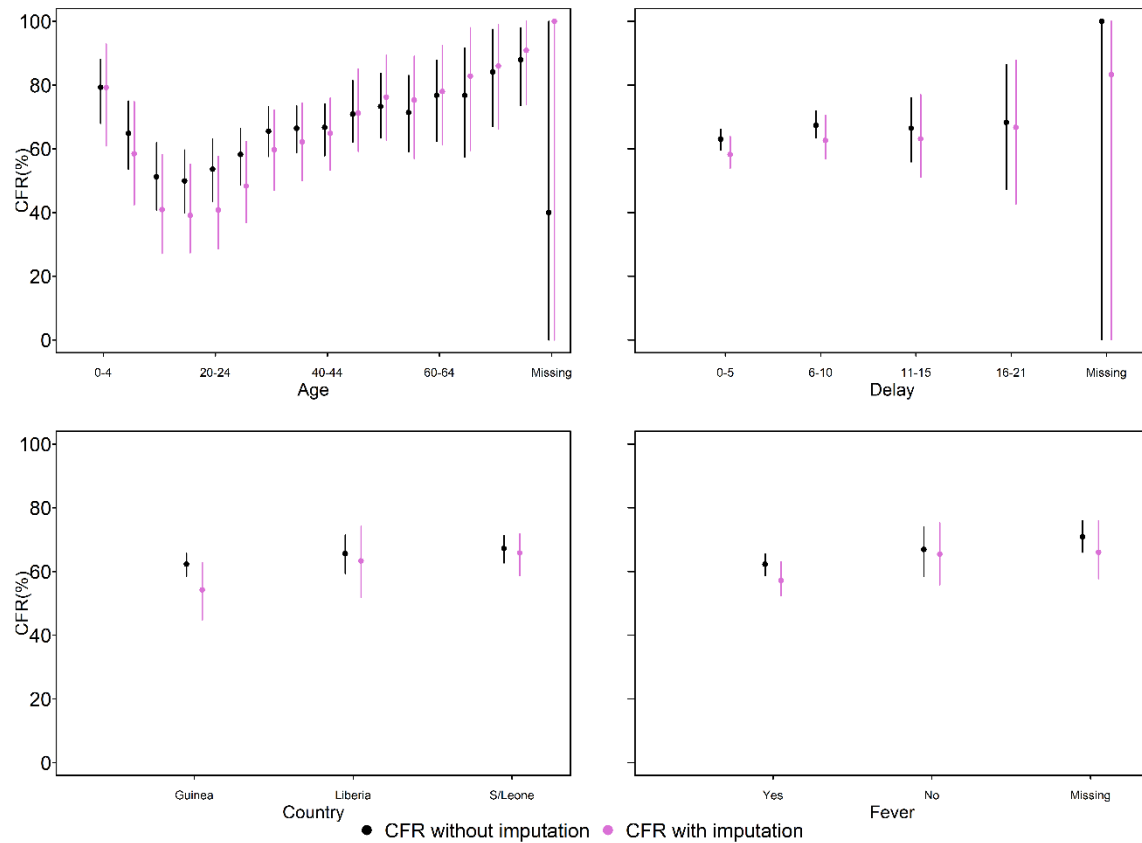


Figure 4. 10: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

CFR for other Predictors

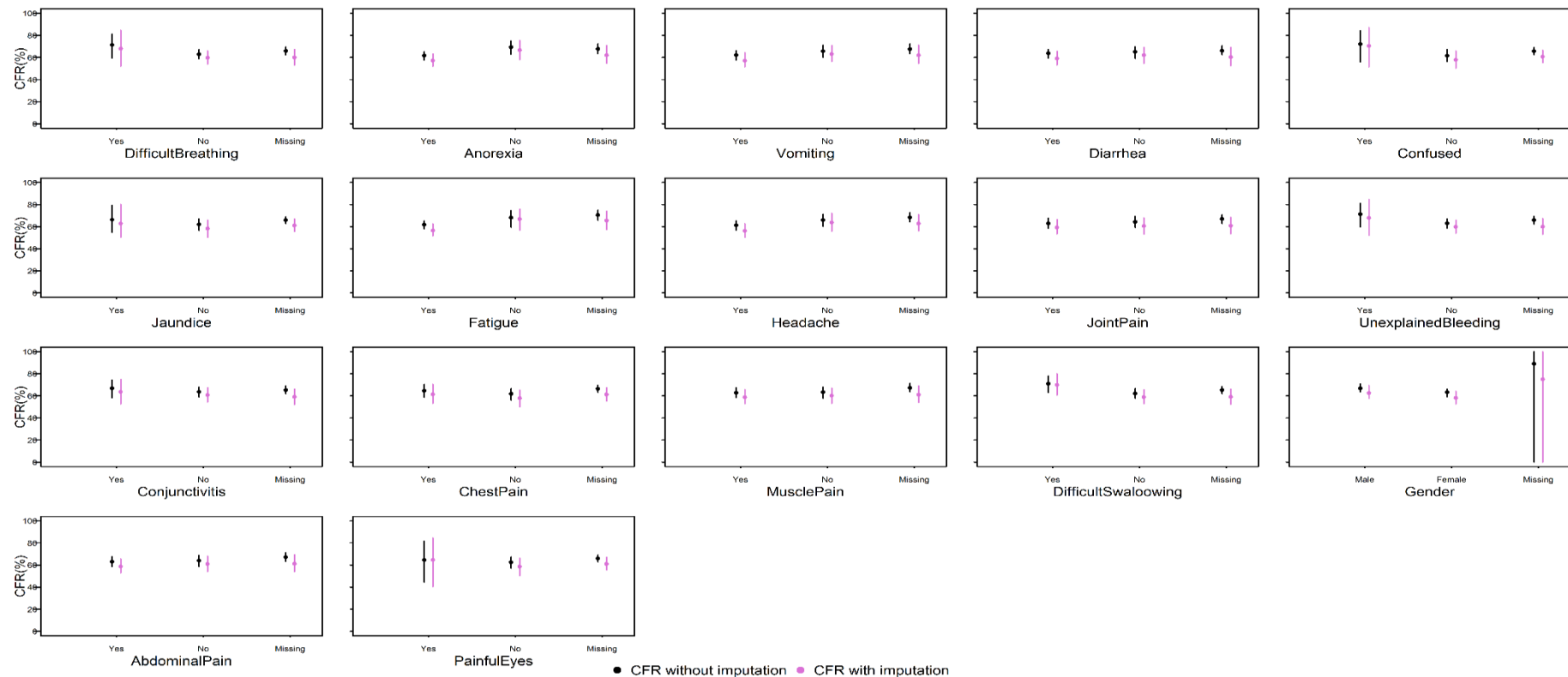


Figure 4. 11: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed cases obtained for the training data using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

4.5.3 Comparison of CFR estimates obtained on the validation data, with and without imputation

4.5.3.1 Confirmed, probable, and suspected cases

This is the performance on the validation data (without combining the observed and imputed survival outcomes). I applied the model on the out-of-sample data to evaluate performance. The validation performance was appreciable for all predictors and all case classification combinations as confidence intervals with and without imputation show some overlap (Figures 4.12— 4.17).

CFR for age, delay, country, and fever

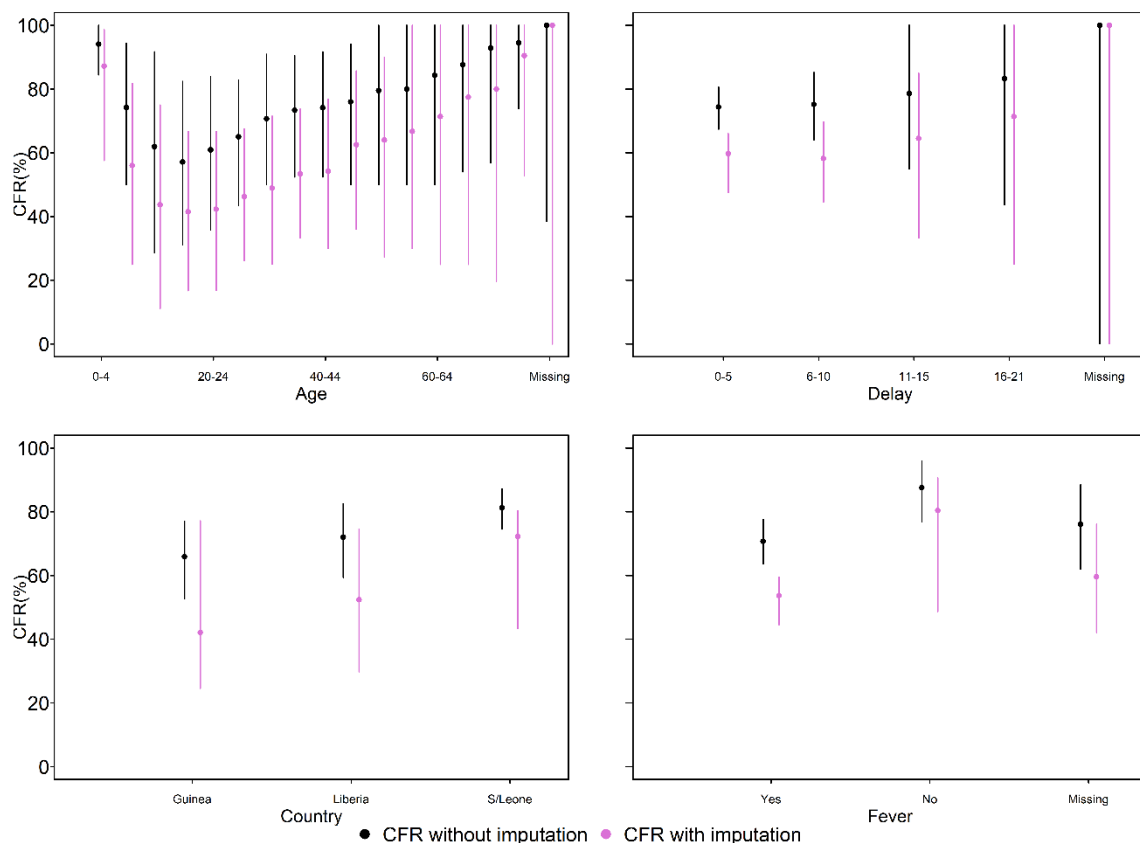


Figure 4. 12: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed, probable and suspected cases for the validation data

obtained using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

CFR for other Predictors

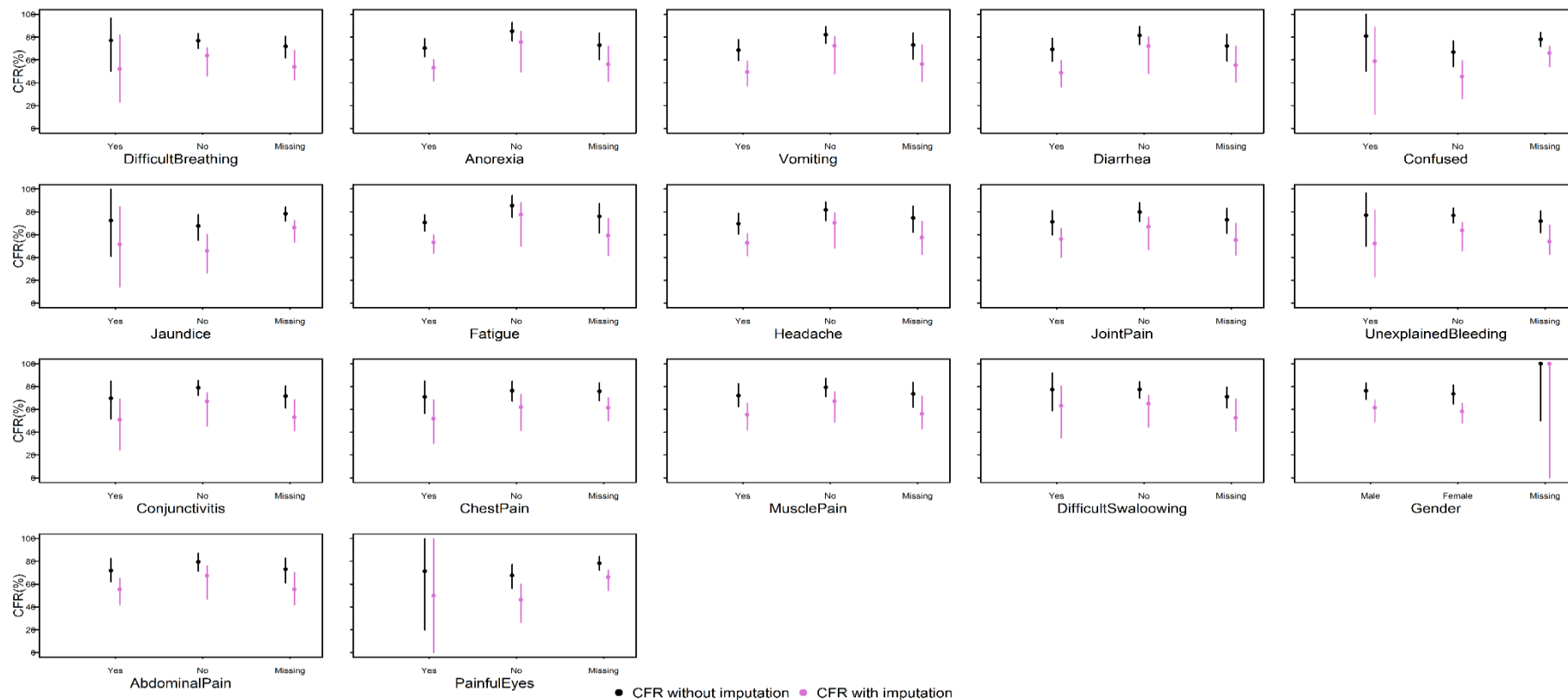


Figure 4. 13: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed, probable and suspected cases obtained for the validation data using the (simplified) BRT model with tree complexity =10, learning rate=0.01, bag fraction=0.5, data partitioning ratio=0.8. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

4.5.3.2 Confirmed and probable cases

CFR for age, delay, country, and fever

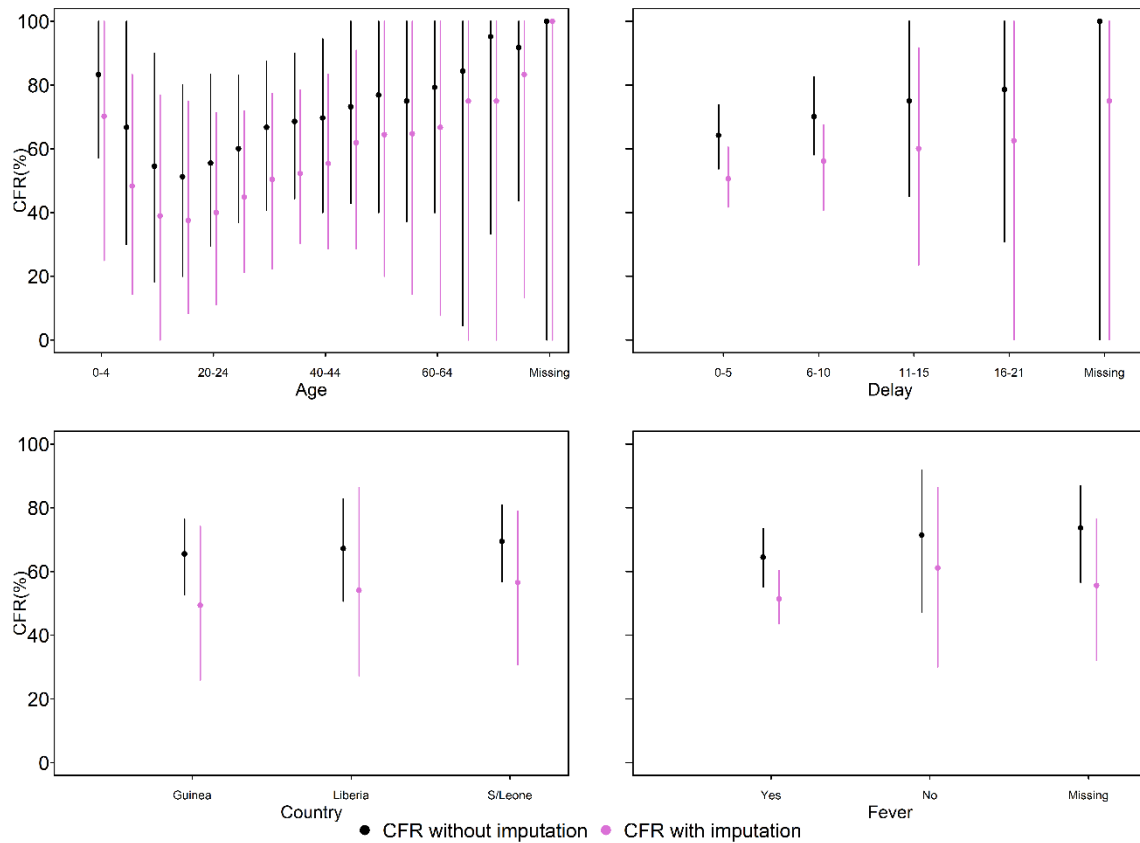


Figure 4. 14: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed and probable cases for the validation data obtained using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

CFR for other Predictors

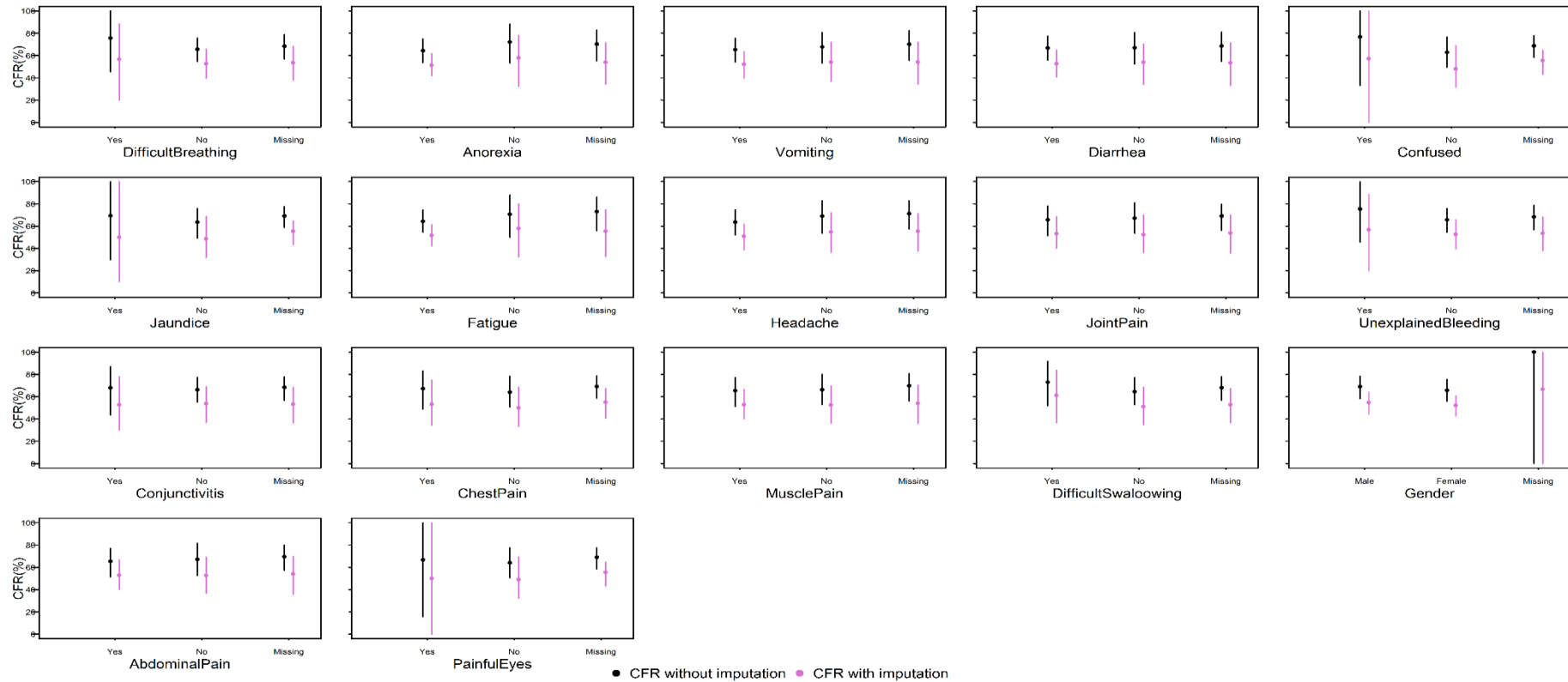


Figure 4. 15: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed and probable cases obtained for the validation data using the (simplified) BRT model with tree complexity =10, learning rate=0.01, bag fraction=0.5, data partitioning ratio=0.8. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

4.5.3.3 Confirmed cases

CFR for age, delay, country, and fever

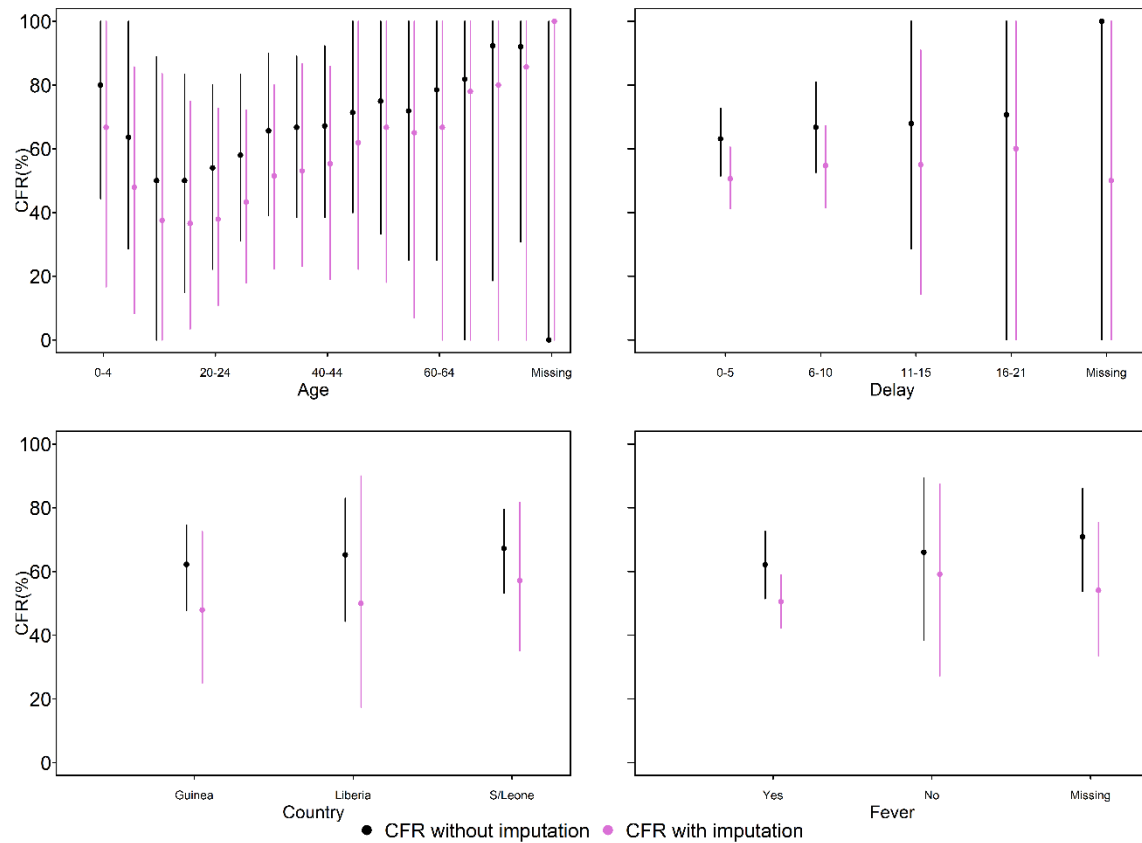


Figure 4. 16: CFR estimates without (black) and with (orchid) imputation for age, delay, country, and fever, for confirmed cases for the validation data obtained using the (simplified) BRT model with tree complexity =27, learning rate=0.001, bag fraction=0.75, data partitioning ratio=0.65. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

CFR for other Predictors

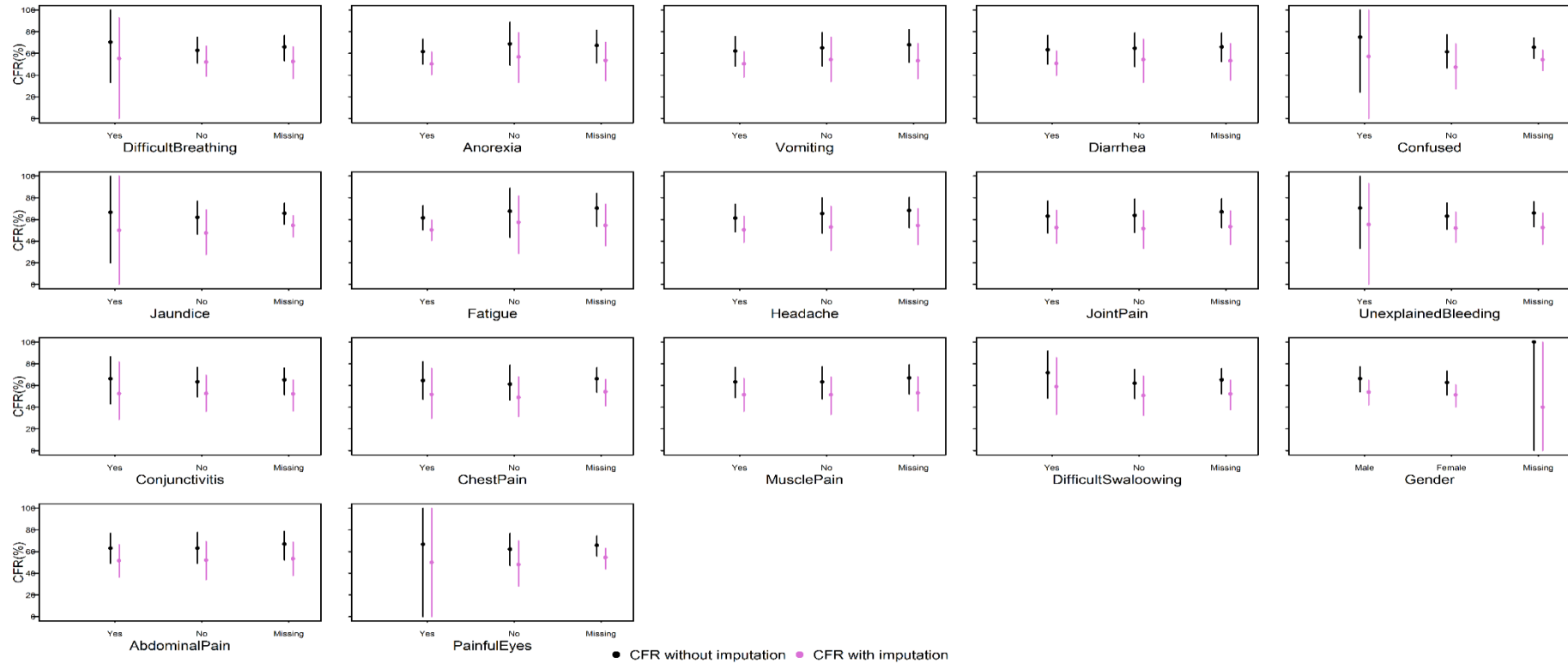


Figure 4. 17: CFR estimates without (black) and with (orchid) imputation for other predictors, for confirmed probable cases obtained for the validation data using the (simplified) BRT model with tree complexity=10, learning rate=0.01, bag fraction=0.5, data partitioning ratio=0.8. Median and 95% confidence intervals (CI) plotted (based on 1,000 bootstrap realisations).

For example, for confirmed, probable and suspected cases, the performance of the model on the training data (e.g. AUC= 93.2 (95% CI: 87.1-98.3)) is higher than that on the out-of-sample data (e.g. AUC= 76.2 (95% CI: 56.3-82.3)). This is demonstrated by the comparison of the observed CFR with the CFR of the training data and the imputed CFR. To improve the CFR estimates and make the estimates more representative, I used two approaches to reduce the bias. These were (i) I combined the imputed CFR with the observed CFR to estimate CFR with imputation and (ii) considering the out-of-sample performance (i.e. sensitivity and specificity) of the (BRT) imputation model, I estimated the adjusted CFR.

4.5.4 CFR adjusted with imputation by gender and age (categories)

Gender, fever, and age are common predictors of epidemiological outcomes [119]. Gender, fever and age were significant predictors of EVD survival; however, this contradicts an earlier analysis of fever [120]. This contradiction could be explained by differences in the sample size and by a selection of the subset of data included in the analysis. Furthermore, this could also be indicative of an overwhelmed system earlier in the outbreak where many people were hospitalised as a precaution. Estimates adjusted with imputation are presented in Table 4.4.

Table 4. 4: CFR estimates with imputation and confidence intervals are calculated from 1000 bootstraps of a BRT model. The BRT parameters used for imputation are $tc=27$, $data\ partitioning=0.65$, $lr=0.001$ and $bf=0.75$.

Predictors	CFR adjusted with imputation
------------	------------------------------

	Confirmed cases Median (95% CI)	Confirmed and probable cases Median (95% CI)	Confirmed, probable and suspected cases Median (95% CI)
Gender			
Male	67.6 (43.4-79.8)	72.6 (43.1-81.8)	84.8 (46.9-86.9)
Female	59.2 (40.1-76.5)	65.1 (40.5-80.2)	79.8 (44.7- 84.9)
Age categories (years)			
0_4	83.3 (49.8-91.5)	88.3 (52.3-93.8)	95.5 (69.0-97.5)
5_9	67.9 (34.4 82.7)	72.2 (33.5-84.9)	82.6 (42.3-88.3)
10_14	48.4 (23.7-78.2)	53.3 (22.9-81.9)	65.1 (28.1-84.3)
15_19	41.3 (26.0-71.6)	47.8 (26.9-74.5)	63.7 (30.1-78.4)
20_24	42.6 (28.8-74.1)	47.0 (30.7-76.1)	69.6 (32.9-81.1)
25_29	46.7 (33.9-75.5)	54.9 (34.5-77.9)	71.6 (35.6-82.2)
30_34	61.2 (41.2-80.8)	65.6 (40.2-81.9)	77.0 (41.0-85.2)
35_39	57.1 (40.3-82.3)	61.8 (41.3-83.5)	78.0 (42.6-86.3)
40_44	66.5 (40.3 81.7)	72.4 (40.6-84.4)	81.1 (43.4-87.3)
45_49	72.8 (43.9-85.2)	77.1 (44.3-86.5)	82.7 (44.8-88.8)
50_54	71.8 (45.0-85.9)	79.0 (46.8-88.0)	85.3 (49.2-90.7)

55_59	76.8 (45.2-87.8)	81.2 (46.4-90.5)	85.6 (45.9-91.0)
60_64	78.5 (48.8-89.5)	81.9 (48.9-91.3)	87.0 (54.3-92.5)
65_69	82.2 (42.7-94.1)	86.7 (43.4-95.1)	90.3 (54.8-95.1)
70_74	83.9 (55.2-95.8)	87.9 (55.3-96.3)	91.7 (63.4-97.7)
75+	89.6 (58.6-98.0)	91.0 (58.4-98.1)	95.1 (69.8-97.8)

4.5.5 CFR estimates without, unadjusted and adjusted with imputation for other predictors in the simplified model

For predictors not included in the main results, figures 4.18 –4.20 show the CFR estimates obtained without, unadjusted with imputation and adjusted with imputation ‘confirmed, probable and suspected’, ‘confirmed and probable’ and ‘confirmed’ cases, respectively. I found that the confidence intervals overlap and the confidence intervals for the CFR unadjusted with imputation and the adjusted CFR are wider. However, because imputed CFR includes all cases and the adjusted CFR account for sensitivity and specificity of the BRT model, the CFR estimated using imputation is more representative and should be closer to the ‘true’ CFR during the epidemic.

4.5.5.1 Confirmed, probable and suspected cases

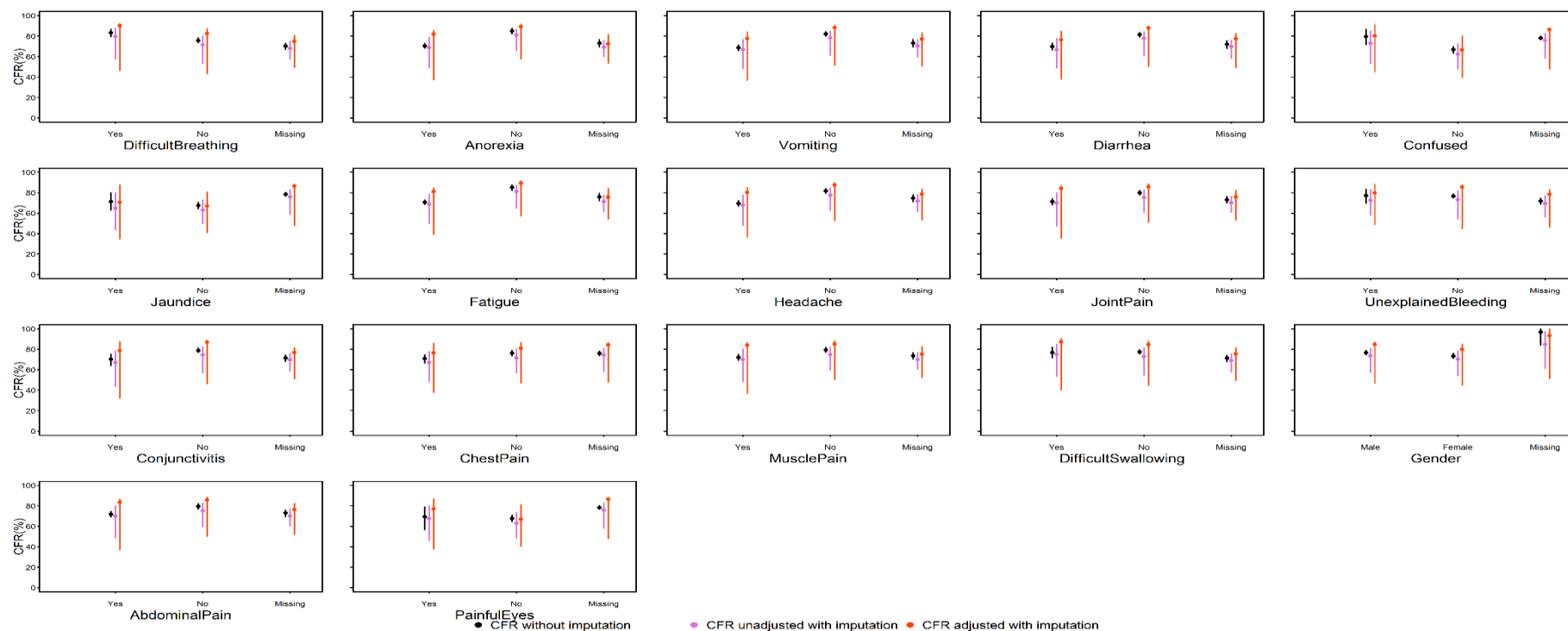


Figure 4. 18: CFR for the other predictors in the simplified model without imputation (black), with imputation unadjusted(orchid) and with imputation adjusted (orange) for estimated BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for ‘confirmed, probable and suspected’ cases).

4.5.5.2 Confirmed and probable cases

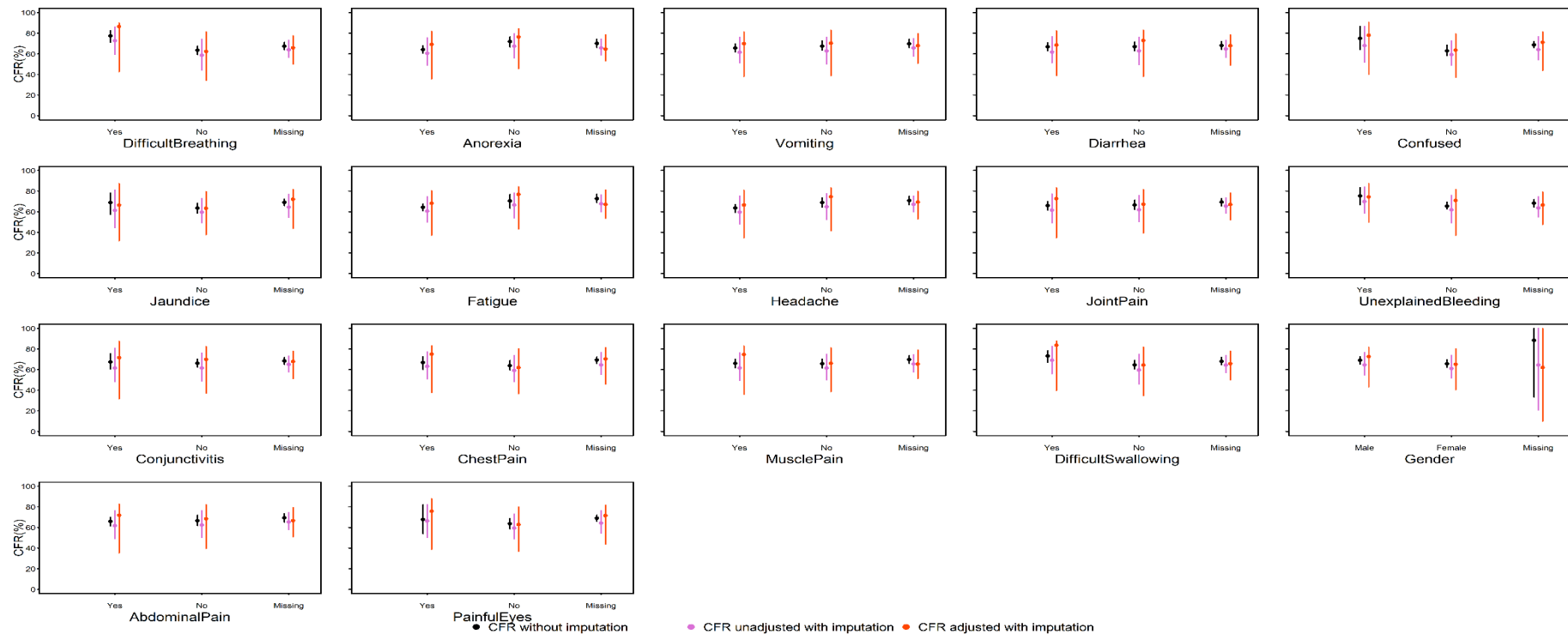


Figure 4. 19: CFR for the other predictors in the simplified model without imputation (black), with imputation unadjusted (orchid) and with imputation adjusted (orange) for estimated BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed and probable' cases).

4.5.5.3 Confirmed cases

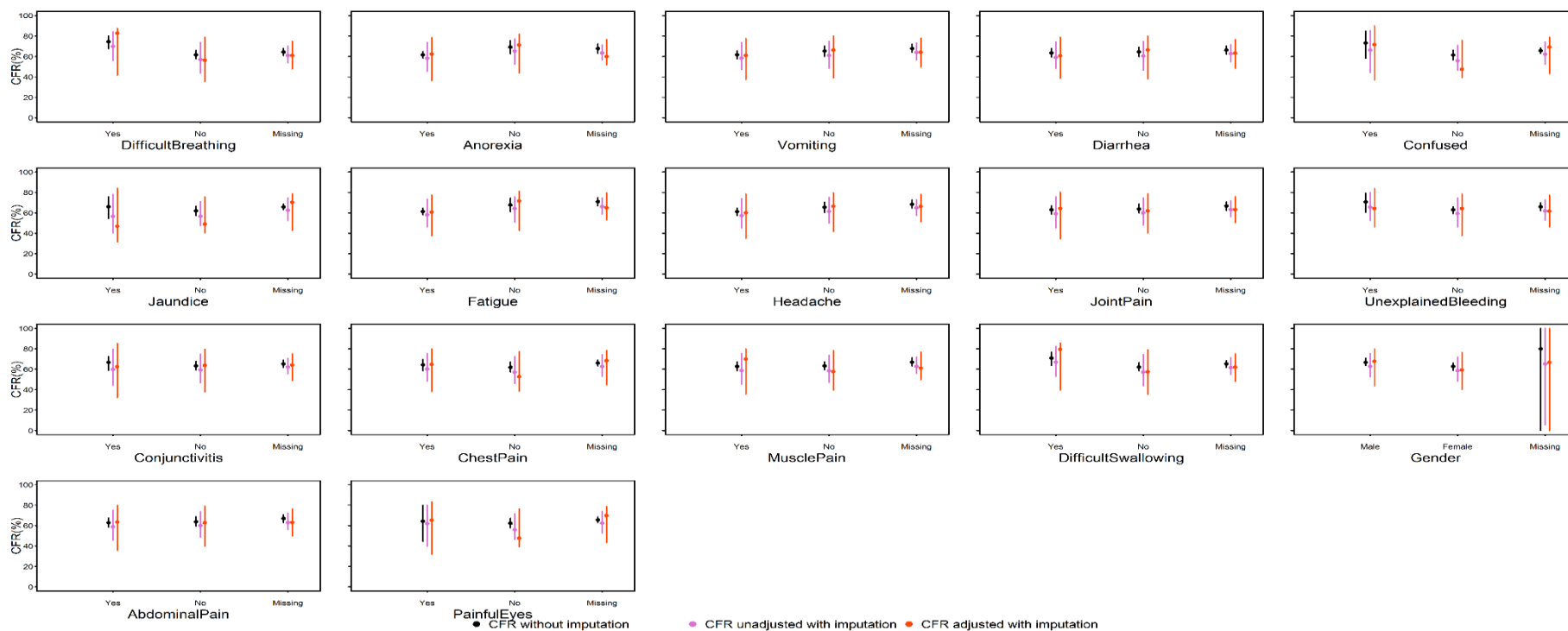


Figure 4. 20: CFR for the other predictors in the simplified model without imputation (black), with imputation unadjusted (orchid) and with imputation adjusted (orange) for estimated BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed').

4.5.6 CFR by fever occurrence and date of case reporting

The earliest and latest dates of case reporting in the dataset are 2nd January 2014 and 25th September 2015, respectively. Informed by the latest date (25th November 2014) from the analyses in West African Ebola Epidemic After One Year: Slowing But Not Yet Under Control [21], and starting from the earliest date 2nd January 2014, I used a cut-off date of 25th November 2014 to estimate CFR for before and after the cut-off. Table 4.5 shows the CFR estimates obtained in the time windows 2nd January to 25th November 2014 and 26th November 2014 to 25th September 2015. The CFR estimates by fever occurrence in the time window 2nd January 2014 to 25th November 2014 show that the estimates obtained in this study are consistent with the estimates obtained in previous publications [17, 21]. Furthermore, these results show that both before and after the cut-off date CFR was higher in patients with no occurrence of fever. Perhaps, this is due to fever not being measured in patients found dead in the community or arriving dead at the ETC.

Table 4. 5: CFR by fever occurrence and date of reporting. The cut-off date (25th November 2014) was informed by previous work using the same data but for an earlier stage of the epidemic.

Case classification	CFR On or before 25 th November 2014		CFR After 25th November 2014	
	Fever Median (%)	No Fever Median (%)	Fever Median (%)	No Fever Median (%)
	(95% CI)	(95% CI)	(95% CI)	(95% CI)
confirmed, probable and suspected cases	70.2 (69.1-71.4)	75.6 (73.0-78.0)	76.1 (74.6-77.5)	95.6 (95.0-96.6)
confirmed and probable cases	66.4 (65.0-67.7)	66.6 (63.0-70.0)	62.7 (60.3-64.9)	84.2 (80.2-87.7)
confirmed cases	63.9 (62.3-65.4)	63.7 (59.8-67.6)	58.6 (56.1-61.1)	78.0 (72.4-82.9)

4.5.7 Robustness of BRT imputation to data down-sampling

In our main analysis, I used the complete line list data to impute unknown survival outcomes. By the end of the West African Ebola epidemic, the data to which I fitted the BRT model were many relative to typical epidemiological datasets in outbreak settings.

Subsequent outbreaks will provide smaller datasets, especially at the initial stages. To investigate the robustness of BRT modelling to different outbreak sample sizes, I down sampled our dataset from 100% of cases (as presented in the main analyses) to 75%, 50%, 25% and 10% of cases and estimated the imputation performance in each scenario (Figure 4.21). It is reassuring that BRT mean performance varied only slightly with decreased sample size. However, as expected the uncertainty surrounding measures of performance increased as sample size decreased. It is important to be able to communicate these types of analyses to policy makers and public health responders in the frontlines of outbreaks as our study show which predictors would be important to obtain reliable estimates in the future.

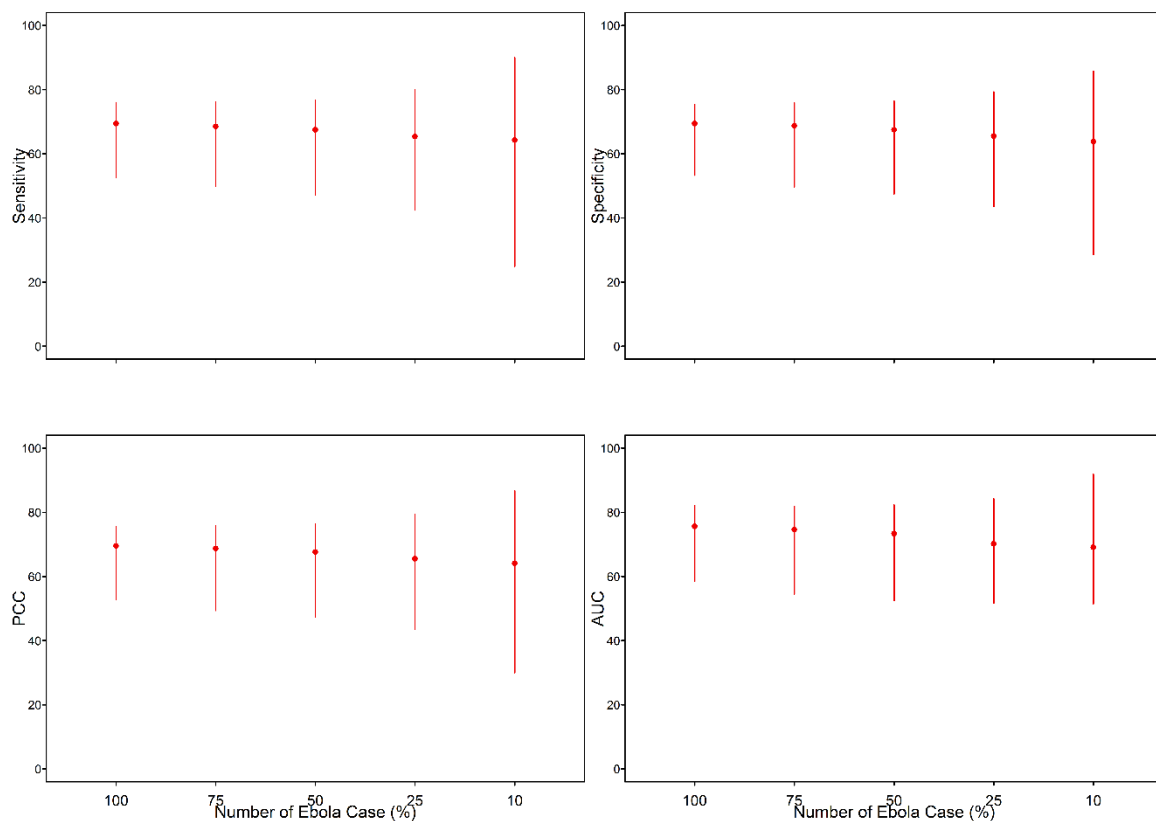


Figure 4. 21: Performance (i.e. sensitivity, specificity, PCC, and AUC) of BRT model after down-sampling of data for all Ebola cases (i.e. confirmed, probable and suspected).

4.5.8 Sampling to explore simulated missingness

Although the adjusted CFR estimates with imputation corroborate with estimates reported in the literature, they were relatively higher than those estimated unadjusted with imputation. While CFR unadjusted with imputation reflects the bias in the BRT model, adjusted CFR estimates which account for this bias are therefore worth reporting. To validate further these adjusted CFR estimates, I carried out a simulated missingness analysis (detailed algorithm described in Chapter 3, section 3.11). Briefly, if I know the CFR of a dataset but pretend as if this CFR were unknown, using the sensitivity and specificity of the BRT model and the tp function, the re-estimated adjusted CFR should be similar to the known CFR. Figure 4.22 shows that for different case combinations, CFR for imputed cases is similar to that for simulated missingness. For instance, for confirmed, probable and suspected cases, CFR for imputed cases was 66.1% (95% CI: 22.7%-88.2%) and CFR for simulated missingness was 65.0% (95% CI: 14.9%-89.2%). Perhaps also, the inclusion of probable and suspected cases that lack laboratory confirmation of EVD is driving the higher adjusted CFR estimates. For instance, for laboratory confirmed cases, the overall adjusted CFR estimate is very similar to the CFR estimated by Garske *et al.* [78]. The adjusted overall CFR I estimated was 62.0% (95% CI: 51.6%-72.5%) while that estimated by Garske *et al.* was 62.9% (95% CI: 61.9%-64.0%) [78].

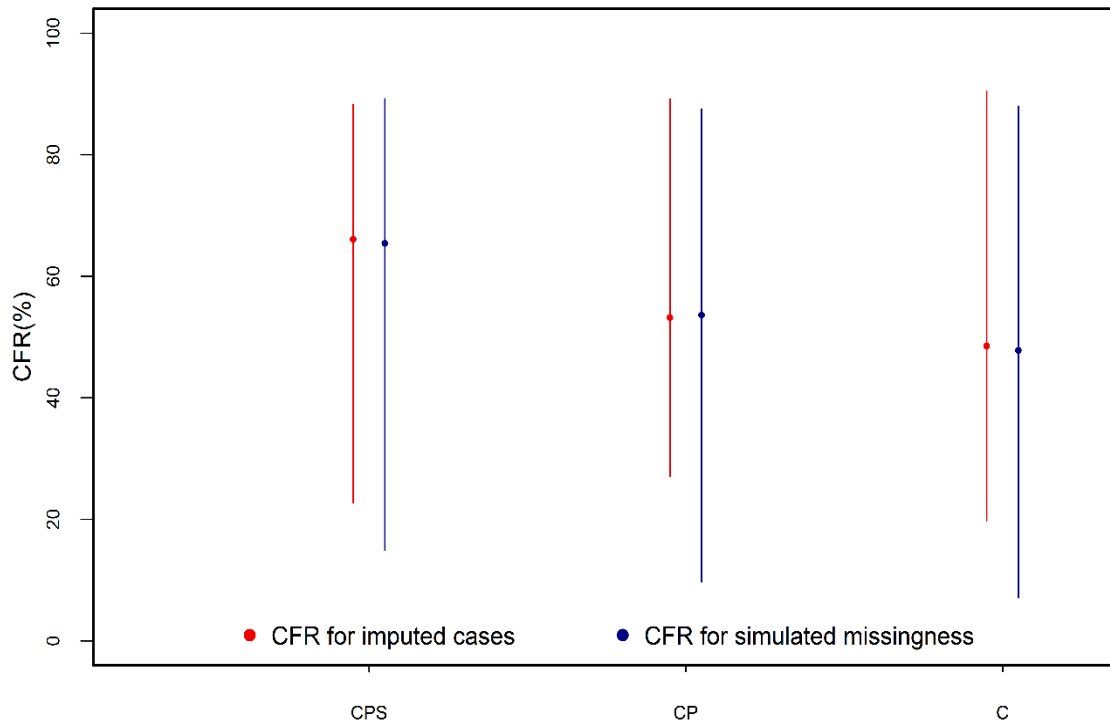


Figure 4. 22: Imputed CFR and simulated missingness CFR for CPS (i.e. confirmed, probable and suspected) cases, CP (i.e. confirmed and probable) cases and C (i.e. confirmed) cases.

4.6 Sensitivity Analysis

In the section, sensitivity analysis is carried out for case combinations (i.e. ‘confirmed’ and confirmed and probable’) not included in the main analysis.

4.6.1 Comparison of known and unknown outcome data for ‘confirmed’ and ‘confirmed and probable’ cases

Figure 4.23 shows the proportion (i.e. dead, alive and missing) of EVD cases by age and reporting delay for ‘confirmed and probable’ and ‘confirmed’ cases.

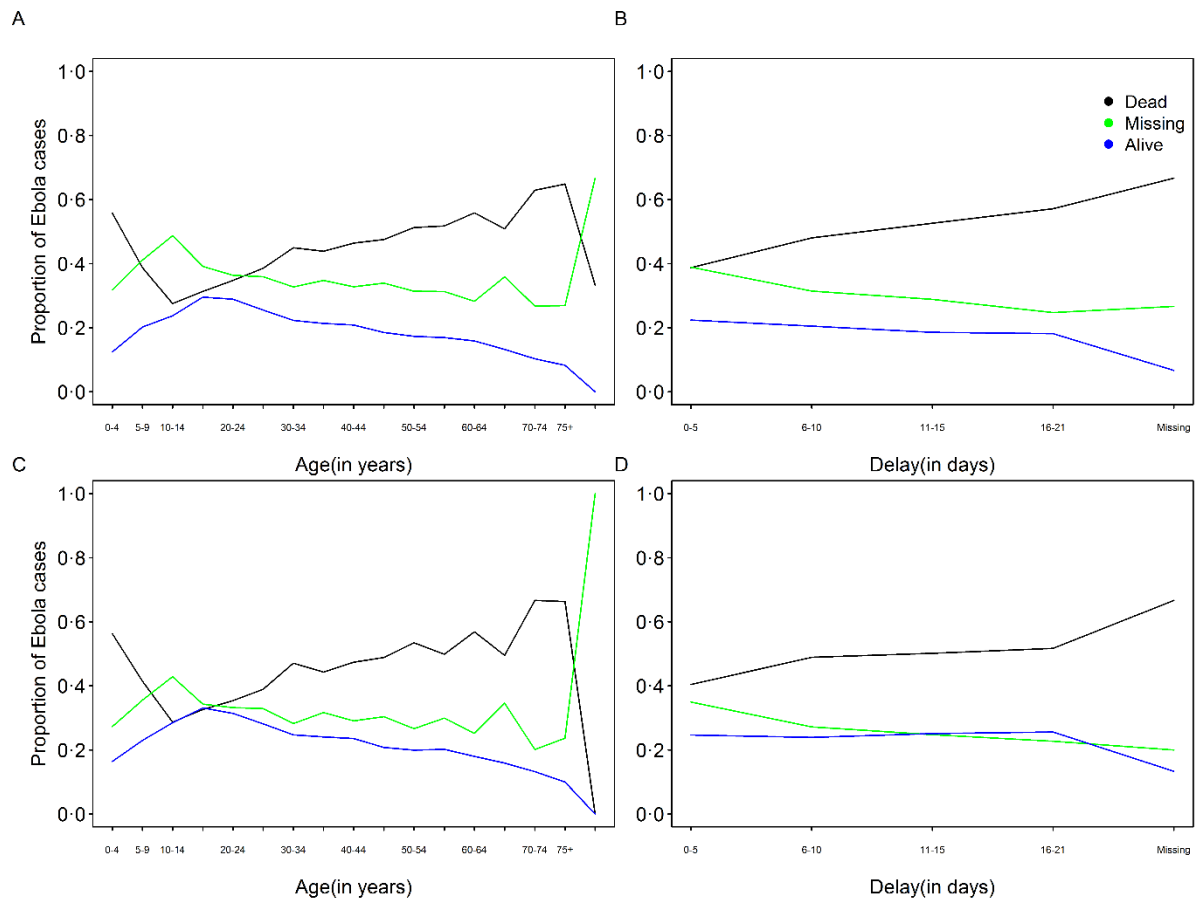


Figure 4.23: A, B: EVD case proportions for confirmed and probable cases by age and reporting delay. C, D: EVD case proportion for confirmed cases by age and reporting delay.

4.6.2 BRT model parameterisation for confirmed case data

Using non-bootstrap resampling (i.e. 50 realisations without replacement for the 72 hyperparameter combinations), re-parameterising the model for 'confirmed' case resulted in similar out-of-sample predictive performance (sensitivity=70% and AUC=78%) (Figure 4.24) compared to that for confirmed, probable and suspected EVD cases (sensitivity=70% and AUC=78%) (Figure 4.4) and the hyperparameters at which this performance was achieved remained unchanged: $tc=27$, $lr=0.001$, $bf=0.75$ and data partitioning ratio=0.65.

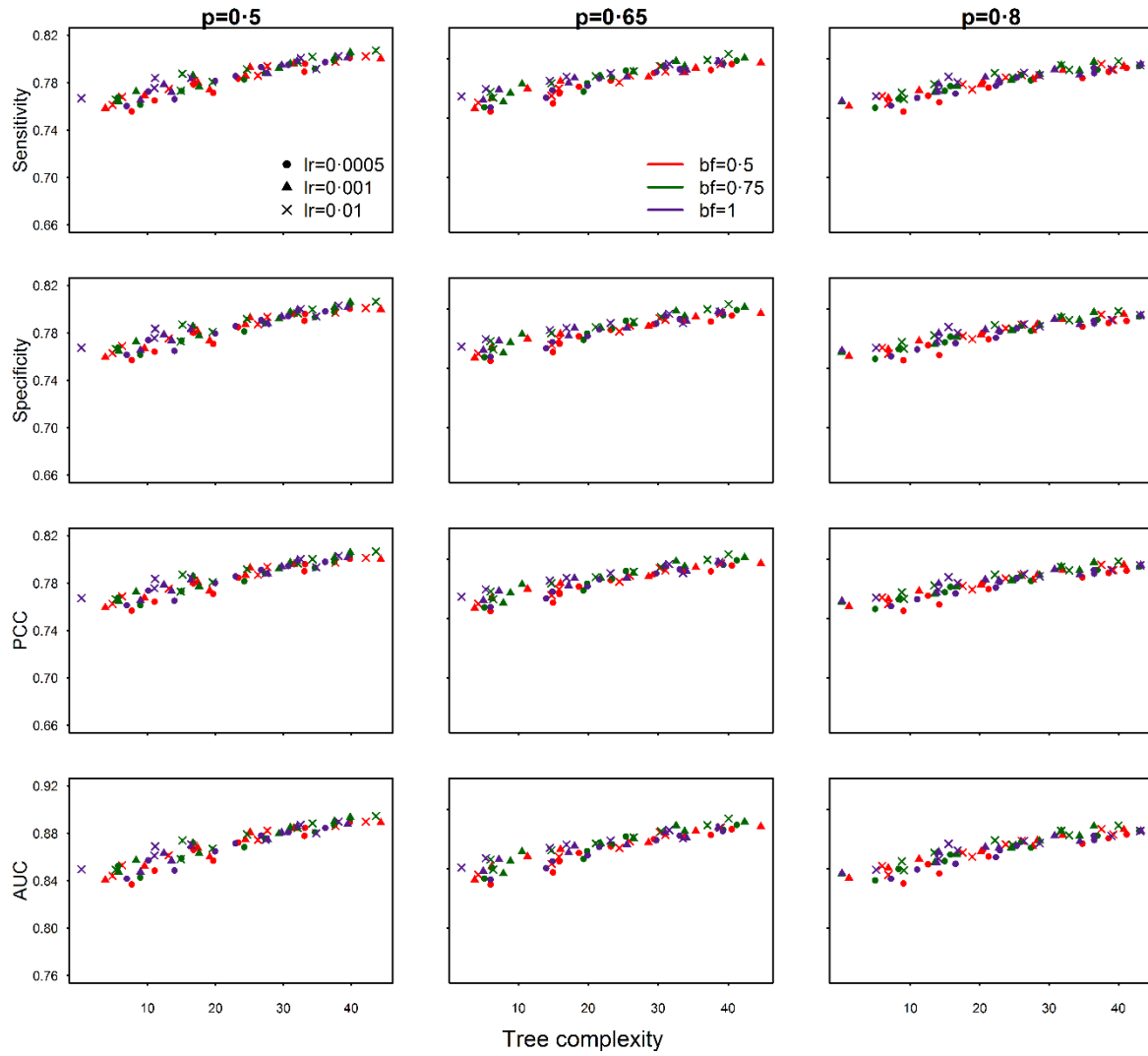


Figure 4. 24: Within-sample accuracy for confirmed cases. p is the proportion of data used as training set. $p=0.5$, $p=0.65$ and $p=0.8$, respectively. PCC=Percentage Correctly Classified, AUC area under the receiver operating characteristic (ROC) curve. This shows an optimistic training performance but also indicates that I am not overfitting the model.

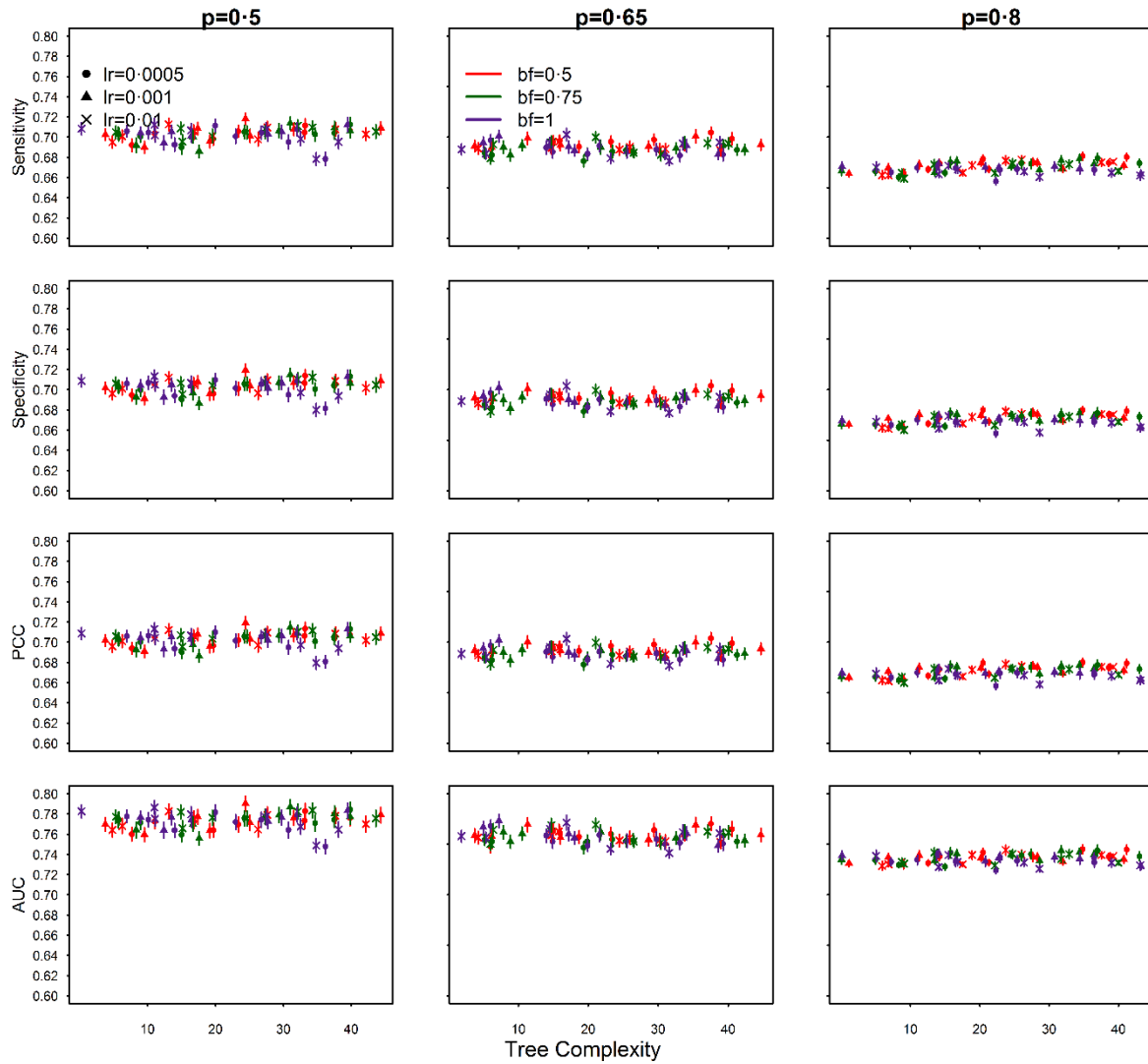


Figure 4. 25: Out-of-sample accuracy for confirmed and probable cases. p is the proportion of data used as training set. $p=0.5$, $p=0.65$ and $p=0.8$, respectively. PCC=Percentage Correctly Classified, AUC area under the receiver operating characteristic (ROC) curve. This shows, overall lower performance for confirmed only data compared to confirmed, probable and suspected data. The hyperparameters broadly remain the same as for confirmed, probable and suspected case data.

4.6.3 Minimal model predictors for variation in case classification

Confirmed and probable cases

I report the important predictors in the model simplified for confirmed and probable cases and confirmed cases.

Table 4. 6: Relative contributions (%) of the important predictors used in BRT models using 10-fold cross validation, $tc = 27$, $lr = 0.001$ and $bf = 0.75$, for confirmed and probable cases, and trained on 1000 training sets generated by randomly sampling 65% cases with known survival outcomes. The minimal model used 10 predictors.

Predictors	Relative Contribution (%)
District	42.2
Hospitalisation status	34.3
Age	16.1
Difficult Breathing	2.5
Delay	1.4
Quarter	1.4
Anorexia	0.8
Case classification	0.6
Fever	0.3

Fatigue	0.3
---------	-----

Confirmed cases

Table 4. 7: Relative contributions (%) of the important predictors used in BRT models using 10-fold cross validation, $tc = 27$, $lr = 0.001$ and $bf = 0.75$, for confirmed cases, and trained on 1000 training sets generated by randomly sampling 65% cases with known survival outcomes. The minimal model used 10 predictors.

Predictors	Relative Contribution (%)
District	43.8
Hospitalisation status	37.2
Age	14.8
Quarter	1.4
Difficult Breathing	1.1
Anorexia	0.5
Fever	0.5
Delay	0.4

Fatigue 0.3

Case classification -

4.6.4 CFRs for variations in case classification

In Tables 4.8 and 4.9, I show the CFR estimates, without imputation, unadjusted with imputation, and adjusted with imputation for 'confirmed and probable cases' and 'confirmed cases'.

Table 4. 8: CFR for confirmed cases. Overall and country level estimates. All CFR estimates and corresponding CIs were calculated using a non-parametric bootstrap of the BRT model.

	Cases for dataset without imputation	Cases for the dataset with imputation	CFR without imputation (95% CI)	Unadjusted CFR with imputation (95% CI)	Adjusted CFR with imputation (95% CI)
Guinea	3,297	3,304	62.1 (58.4-66.0)	62.0 (58.4-66.1)	62.0 (58.7-66.0)
Liberia	2,384	3,743	65.1 (59.0-70.2)	63.2 (53.9-72.4)	67.0 (47.0-75.6)
Sierra Leone	3,790	9,397	67.0 (63.6 -71.4)	58.8 (43.4-80.2)	61.1

(30.8-86.1)

Overall*	9,471	16,444	64.6	60.4	62.5
			(62.4-67.2)	(50.5-73.4)	(42.6-77.7)

Table 4. 9: CFR for confirmed and probable cases. Overall and country level estimates. All CFR estimates and corresponding CIs were calculated using a non-parametric bootstrap of the BRT model.

	Cases for dataset without imputation	Cases for the dataset with imputation	CFR without imputation (95% CI)	Unadjusted CFR with imputation (95% CI)	Adjusted CFR with imputation (95% CI)
Guinea	3,740	3,747	65.9 (62.3-69.7)	65.8 (62.4-69.7)	65.7 (62.3-69.7)
Liberia	3,131	5,343	66.7 (62.2-71.9)	66.4 (56.2-76.1)	77.2 (42.7-80.9)
Sierra Leone	4,421	11,264	69.0 (65.1-73.1)	59.4 (44.7-80.5)	62.5 (30.2-87.1)
Overall*	11,292	20,354	67.3	62.6	68.5

(64.8-70.1) (53.5-75.4) (42.2-80.5)

4.6.5 BRT imputation performance for variations in case classification

Out-of-sample performance obtained using the final (simplified) BRT model for 'confirmed and probable' and 'confirmed' cases (Table 4.10).

Table 4. 10: BRT Imputation performance (simplified model) for confirmed cases and confirmed and probable cases.

Performance Measures	Imputation Performance	
	Confirmed Cases Model	Confirmed and Probable
	Performance	Cases Model Performance
	Median (%) (95% CI)	Median (%) (95% CI)
Sensitivity	58.8 (47.4-67.4)	60.0 (48.0-69.2)
Specificity	58.8 (48.3-67.4)	59.9 (48.1-68.8)
PCC	58.8 (48.3-67.0)	59.9 (48.3-68.5)
AUC	63.2 (51.6-72.5)	64.5 (52.6-74.7)

4.6.6 CFR estimates by age group, delay, country and fever for other case classifications

I show the CFR estimates without, unadjusted with imputation and adjusted with imputation obtained with the final (simplified) BRT model for 'confirmed and probable' and

'confirmed' cases (Figures 4.26, 4.27). The adjusted CFR with imputation for confirmed and probable cases is consistent with CFR estimates for confirmed and probable cases in the present outbreak in the Democratic Republic of the Congo, whereas, the adjusted CFR with imputation for confirmed cases is consistent with CFR cases for the West African Ebola epidemic [19]. For instance, Garske *et al.* estimated CFR for confirmed cases as 62.9% (95% CI: 61.9%-64.0%) while the adjusted CFR I estimated for confirmed cases was 62.5% (95% CI: 42.6%-77.7%). As of 26 December 2018, the CFR for confirmed and probable cases in the Democratic Republic of the Congo was 60.4% (95% CI: 56.3%-64.4%) while the overall adjusted CFR with imputation for West Africa I estimated was 68.5% (95% CI: 42.2%-80.0%) [78]. However, on a note of caution, the CFR for the Democratic Republic of the Congo outbreak is subject to change due to ongoing reclassification, retrospective investigation, and the ongoing availability of confirmed laboratory results.

4.6.6.1 Confirmed and probable cases

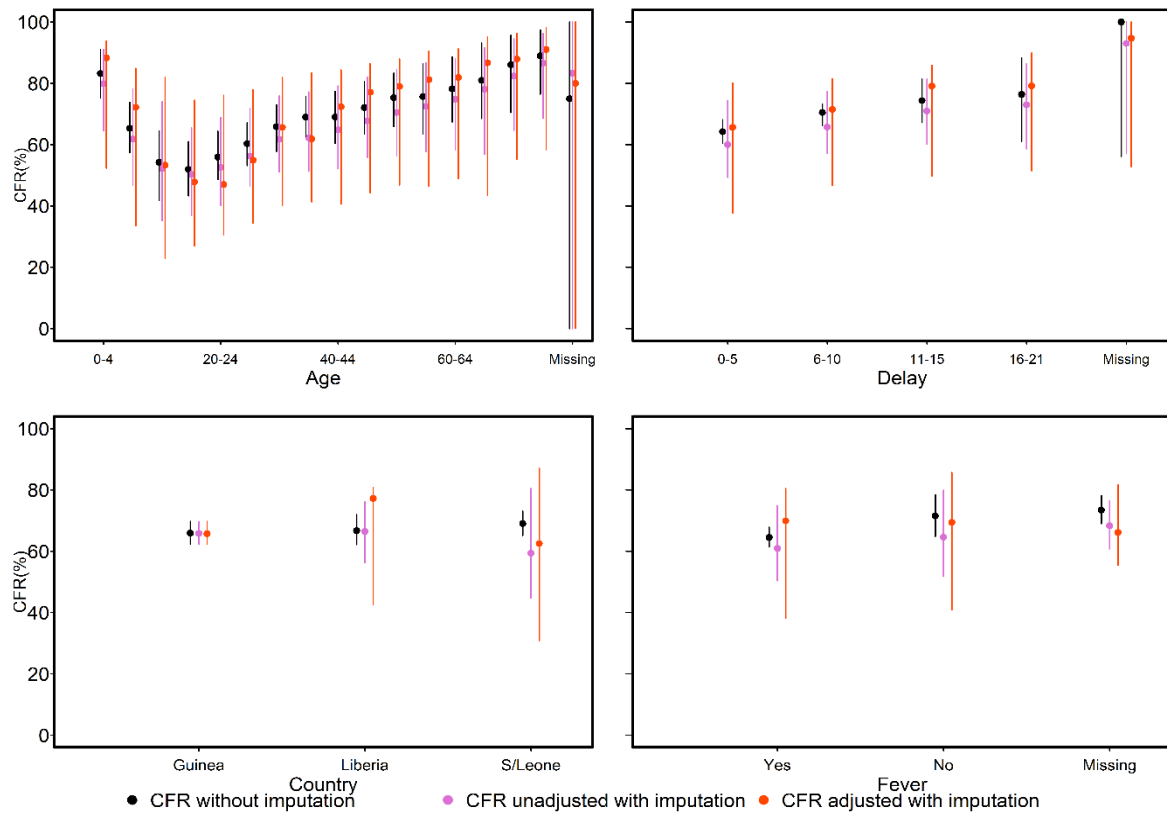


Figure 4. 26: CFR by age group, delay, country, and fever without imputation (black), with imputation unadjusted (orchid) and with imputation adjusted (orange) for BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed and probable' cases).

4.6.6.2 Confirmed cases

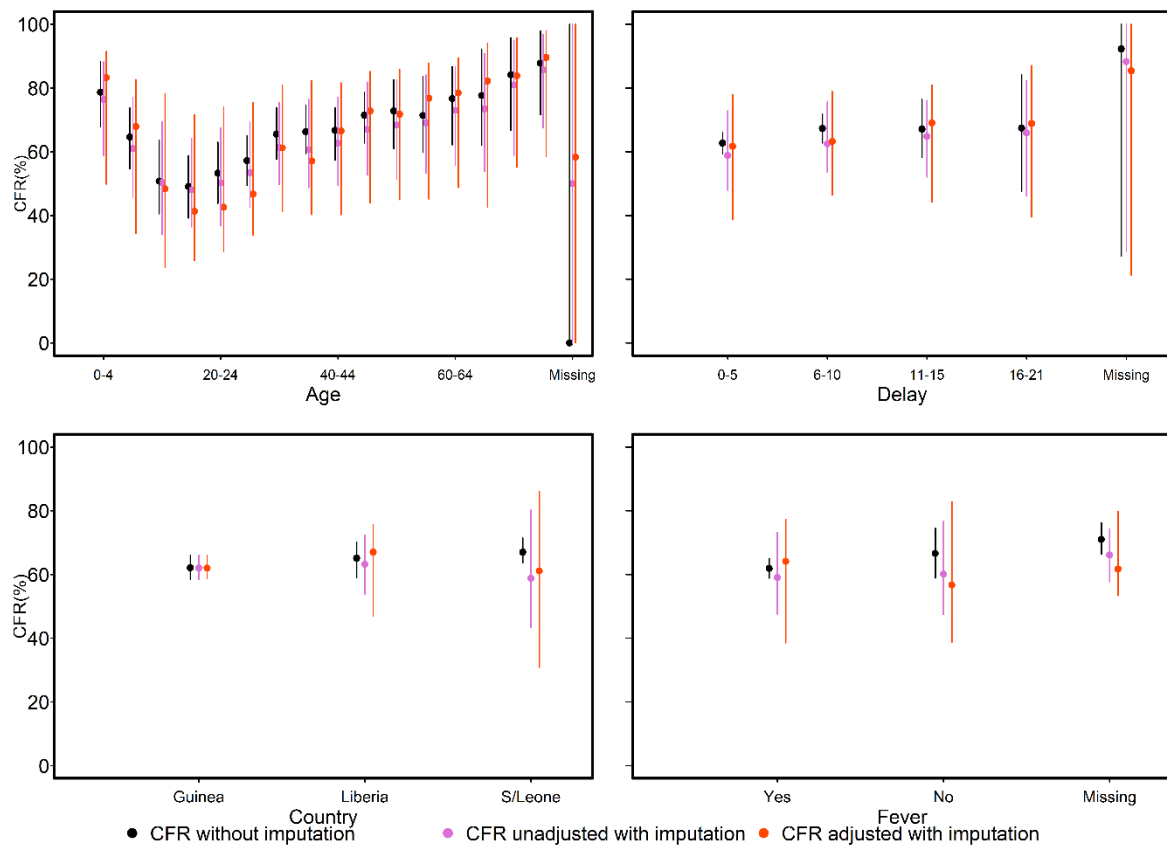


Figure 4. 27: CFR by age group, delay, country, and fever without imputation (black), with imputation (orchid) unadjusted and with imputation adjusted (orange) for BRT sensitivity and specificity. Median and 95% confidence intervals (CI) plotted (based on 1000 bootstrap realisations for 'confirmed and probable' cases).

Chapter 5: Spatiotemporal variability in Case Fatality Ratios for 2013–2016 Ebola epidemic in West Africa

Work in this chapter forms the basis of: **Forna A**, Dorigatti I, Nouvellet P, Donnelly CA.

Spatiotemporal Variability in Case Fatality Ratios for 2013–2016 Ebola epidemic in West Africa. International Journal of Infectious Diseases 2020.

In this chapter, building on the algorithmic framework in Chapter 4, I used semivariograms and kriging to investigate the patient mix in detail to characterise the subnational spatiotemporal heterogeneities in EVD CFR.

5.1 Introduction

Ebola virus disease (EVD) is associated with a very high fatality, with 10,884 deaths (among 26,277 cases) reported for the West African epidemic [3, 108]. Although previously limited to rural communities, recent outbreaks have occurred in urbanised towns and cities in West Africa and more recently in the Democratic Republic of the Congo, further increasing the possibility of fatalities across densely populated communities with limited material and human resources [17, 19, 121, 122]. The case fatality ratio (CFR), defined as the proportion of patients dying from EVD, generally varied between 60% to 90% for known outbreaks [3, 123]. Robust CFR estimates for entire epidemics and subgroups not only inform the situational awareness during ongoing outbreaks but are essential for future public health contingency planning, resource allocation and for evaluating the impact of interventions subsequently implemented during outbreaks [78].

While published West African CFR estimates are consistent with those reported for other outbreaks, including the most recent outbreaks in the Democratic Republic of Congo (typically 60% to 70% for confirmed and probable cases), CFRs have shown variability over

space and time [19, 78]. Some analyses have aimed to quantify these variations [17, 78], for example, Garske *et al.* detected outliers in CFR estimates between districts of residence and treatment centres during the West African epidemic based on individual level observed survival outcomes [78]. Spatiotemporal characterisation of other epidemiological parameters, such as transmission rates and mutation rates, has also been informative [73, 124]. For example, Kramer *et al.* used network models to estimate the spatial process of EVD transmission and found that the probability of transmission between locations depended on distance, population density and the international cross-border importations of EVD cases [73]. A strong case has been made that simple spatiotemporal characterisations of such heterogeneities, available in real time, could have enhanced the national and global response as the West African epidemic unfolded [67].

In chapter 4, a methodological framework built around Boosted Regression Trees (BRT) with adjustments for imperfect model sensitivity and specificity, generated more representative CFR estimates for the West African Ebola epidemic than the crude CFR estimates by imputing survival outcomes for the 44% of the individual cases with missing survival outcomes [91]. Here, I use a similar BRT imputation framework to investigate the spatiotemporal heterogeneities in CFR for the 2013-2016 West African Ebola epidemic.

5.2 Methods

5.2.1 Geographical characteristics of the study setting

In this Chapter, the West African region refers to the three most affected countries during the 2013–2016 EVD epidemic, Sierra Leone, Guinea, and Liberia with a combined area covering 429,110 km². The region has been divided into 63 administrative units in total, comprising 14 districts in Sierra Leone, 34 prefectures in Guinea and 15 counties in Liberia

(See Figure 2.1 in Chapter 2). For this study, districts, prefectures, and counties for Sierra Leone, Guinea and Liberia, respectively, are referred to as districts. Districts were then used as the spatial unit of analysis.

I used GADM (i.e. Database of Global Administrative Areas) data to map the districts of the West African region [125]. District-level CFR estimates were georeferenced, and all 63 districts were included in the analysis. There were seven districts in Guinea without any EVD cases according to the case database. The Western Urban and Western Rural districts in Sierra Leone were aggregated into a single western district as used in previous analyses [78].

5.2.2 Data sources

EVD cases were reported to the World Health Organisation (WHO) using a viral haemorrhagic fever (VHF) case reporting forms as published in the supplementary information of Garske *et al.* [21]. In the WHO case reporting system, cases were classified as confirmed, probable and suspected. The epidemiological details of the case classification system have been described in detail elsewhere [98] and a brief description of the case classification is provided in Chapter 2. Here, I report CFR estimates based on all reported cases (confirmed, probable, and suspected). Case classification (confirmed, probable, and suspected) was an important predictor (explained 10.1% of observed CFR variance) of EVD CFR for the West African epidemic and was accounted for in the BRT models used in this chapter.

5.2.3 Spatiotemporal Analysis

I used BRT models [58] to impute missing survival outcomes and correct for sensitivity and specificity in imputed outcomes in the calibrated BRT model, thus estimating more representative CFRs (i.e. estimates that account for the missingness in survival outcomes

unlike the crude estimates that do not account for missingness in survival outcomes) at national and district-level. Briefly, BRT models were calibrated to predict the survival outcomes using the 24 most important predictors of CFR [91]. Hyperparameters (tree complexity ($tc=27$), learning rate ($lr=0.001$), bag fraction ($bf=0.75$) and data partitioning= 0.65) were optimised to maximise out-of-sample accuracy, allowing us to generate district-level and quarterly CFR estimates, along with the relative uncertainty, using a non-parametric bootstrap approach described in Chapter 3. Residuals from the adjusted BRT model were also calculated. These residuals were defined as the difference between the observed CFR (i.e. *ib* for spatial and *qb* for temporal observed CFR estimates as defined in Chapter 3) and predicted CFR estimates adjusted for imputation. Details of the algorithms used for CFR prediction and residual estimation are included in Chapter 3.

Choropleth maps of district-specific median CFRs and their standard deviations were produced for the West African region. Spatial correlation in CFR residuals in the whole West African region was tested for using spatial autocorrelation global Moran's I index [126]. I used local Moran's I index as a local indicator of spatial association (LISA) statistic. The LISA statistic for each district indicates the extent of spatial clustering of similar CFR values around that district and the sum of LISA statistics for all districts is proportional to the global Moran's I index [127, 128]. For the LISA statistic, the spatial relationship of the districts was specified using a spatial weight matrix in which neighbours were defined using Delaunay triangulation (i.e. natural neighbours) whereby centroid per district defines a set of finite points on a plane. For each district centroid, there is a corresponding region (i.e. Voronoi cell) consisting of all points on the plane closer to that specific district than to all the districts on the plane [129].

Semivariograms [130] were used to represent graphically the autocorrelation in the estimated CFR at the district-level and in time, at a quarter-level both for the West African region as a whole and for each country (Sierra Leone, Guinea and Liberia) (Brooker *et al.*, 2004) [55]. Mathematically, the spatial autocorrelation was measured as the semivariance $[\gamma(h)]$ which is half the mean squared difference between pairs of district-level CFR values y_i and y_j , separated by lag distance (in km) h :

$$\gamma(h) = \frac{1}{2N(h)} \sum_i^{N(h)} (y_i - y_j)^2, (3)$$

where $N(h)$ is the number of district-level CFR pairs at each lag distance h [70]. In this analysis, the largest lag h was defined as half the maximum distance among all pairs of district-level CFR estimates. A functional form was then fitted to the semivariogram defining three parameters; (i) the sill, which reflects the spatial component of the semivariance (i.e. the level of spatial autocorrelation between district-level CFR pairs), (ii) the range, which is distance at which district pairs are considered to be spatially independent and (iii) the nugget, which represents the small-scale spatial variability or measurement errors during data collection [131]. Typically, the semivariance between neighbouring districts is expected to be low if there is spatial autocorrelation and gradually increase with the distance between pairs of districts, levelling off at the sill (on the y-axis) and range (on the x-axis) [132].

To provide a continuous description of the covariance structure between district pairs, spatial models were fitted to each empirical semivariogram [133]. Five spatial models were used to examine the spatial relationships including linear, Cauchy, exponential, Gaussian and Matern functions [134]. I used the maximum likelihood to identify the spatial model that yielded the smallest Akaike's Information Criterion (AIC), fitted the overall empirical

variogram to each spatial model and used this spatial model for all subsequent analysis (see additional results). Where significant spatial correlation was found, I produced continuous maps of the predicted CFR added to the residuals. The 95% confidence intervals for the sill, range and nugget were obtained using the 'gboot_variogram' and the 'gboot_CI' functions in the 'geotoolsR' R package, refitting the semivariogram to 1,000 non-parametric bootstrap samples of the errors of each fitted semivariogram [135].

Ordinary kriging is a spatial interpolation procedure that assumes that the outcome, in this case, the CFR, can be modelled as a continuous spatial process [136]. Using the semivariogram model, kriging allowed us to estimate residuals and predicted CFR values at unrecorded districts from the neighbouring districts with recorded estimates using the global neighbourhood approach. The kriging generating a near continuous grid of georeferenced points (26,208 geospatial points on 0.05x0.05km grid) within the West African region. This grid unit-size was chosen to fit within the limit of the computational resource. All 26,208 geospatial points are considered in estimating residuals and predicted CFR for each geospatial point covering districts with both recorded and unrecorded estimates [137]. The kriged residuals were then added to the kriged predicted CFR adjusted for imputation from the BRT model to produce a best linear unbiased predicted isopleth (i.e. continuous) map of CFR.

As a sensitivity analysis, I investigated the effect of direction on spatial autocorrelation in the CFR estimates. Directional semivariograms were estimated and these were compared to the estimated omnidirectional semivariogram (see sensitivity analysis).

Furthermore, to investigate whether the district effects explained spatial variations, I refitted the BRT model excluding district from the predictors and re-estimated the global

and local autocorrelation coefficients (i.e. Moran's I indexes and their corresponding p-values) for the resulting residuals.

To investigate temporal autocorrelation, I aggregated the CFR estimates and residuals by temporal quarter, estimated temporal semivariograms and fitted models treating differences in time as differences in space. In the temporal semivariogram model, I substituted the lag distance h with quarterly time lag t and $N(t)$ is the number of temporal-level (i.e. quarter) CFR pairs at each lag time t . Thus, temporal autocorrelation was estimated as $\gamma(t)$. I used the Durbin-Watson (D-W) test to further quantify and test for temporal autocorrelation [138].

5.3 Results

Georeferenced district-level predicted CFRs adjusted for imputation and their corresponding standard deviations showed substantial heterogeneity across the districts of Guinea, Sierra Leone, and Liberia (Figure 5.1a) with district-level residuals showing moderate regional spatial autocorrelation (Global Moran's I Index=0.146, $p=0.033$). Kailahun (Local Moran's I Index=-0.685, $p=0.019$), Kenema (Local Moran's I Index=0.772, $p=0.003$) [in Sierra Leone], Beyla (Local Moran's I Index=-0.908, $p<0.001$), and Nzerekore (Local Moran's I Index=0.885, $p=0.002$) [in Guinea], and Bomi (Local Moran's I Index=0.729, $p=0.01$) and Grand Cape Mount (Local Moran's I Index=0.852, $p=0.003$) [in Liberia] were the districts that showed significant local spatial autocorrelation (additional results Table 5.2).

Excluding district from the predictors in the BRT model produced district-level residuals with substantially greater and highly significant global spatial autocorrelation (Global Maron's I index=0.348, $p<0.001$) and significant local spatial autocorrelation in ten districts (additional

results Table 5.2). One district; Nzerekore shows a significant correlation with/without district as a predictor in the BRT model (additional Results Table 5.2).

In Guinea, the predicted CFR adjusted for imputation was highest in Dinguiraye and Dabola and lowest in Tougue (Figure 5.1a). In Sierra Leone, the predicted CFR adjusted for imputation was highest in a few districts from Freetown in the West Urban area to Kono to the east of the country, and the predicted CFR adjusted for imputation was lowest in Kailahun. In Liberia, the predicted CFR adjusted for imputation was highest in Grand Kru and Maryland and lowest in Sinoe (Figure 5.1a). Figure 5.1b shows the observed CFR (observed CFR ib, as described in Chapter 3 section 3.12) and Figure 5.1c shows the corresponding district-level CFR residuals.

Compared to other spatial models (i.e. linear, Cauchy, exponential and Matern), the Gaussian model yielding the smallest AIC (additional Results Table 5.3) and the more precise range parameter was subsequently fitted to the empirical semivariograms. Semivariogram analysis showed significant spatial autocorrelation overall (all three countries combined) and for Sierra Leone, Guinea and Liberia individually (Figure 5.2, Table 5.1). The range parameters were 89.6 km (95% CI, 33.3–99.69 km), 91.1 km (95% CI, 40.9–95.0 km), 82.9 km (95% CI, 30.4–131.8 km) and 74.3 km (95% CI, 34.9–74.3 km) for the region as a whole, Sierra Leone, Guinea and Liberia, respectively (Table 5.1).

Because spatial structure was identified across the region (Figure 5.3a & 5.3b), an isopleth map combining the predicted CFR adjusted for imputation and the kriged residuals was used to account for unexplained spatial structure in the data (Figure 5.3c). On this kriged map, CFR was highest in coastal West African region, from Conakry in Guinea to the Western

district (including the capital Freetown) in Sierra Leone and towards Montserrado county (including the capital Monrovia) in Liberia (Figure 5.3c).

Quarter-specific results were reported for 2014 and 2015, this was because there were too few cases in late 2013 and early 2016 to be informative. The quarter-specific predicted and observed (observed CFR qb, as described in Chapter 3, section 3.13) CFR also showed substantial heterogeneity and overall, both CFRs were lowest in the second quarter of 2014 (Figure 5.4a & 5.4b). The quarterly residuals of CFR showed no significant temporal clustering (D-W autocorrelation coefficient=-0.417, $p=0.454$) (Figure 5.4c) as demonstrated by the flat temporal semivariograms of the residuals for both the whole region and each country (Figure 5.5).

The total number cases per district are presented in Figure 5.6 For confirmed, probable and suspected cases, the Western district in Sierra Leone, that includes Freetown (the capital city) and sub-urban areas, reported the highest number of cases (8813).

Table 5. 1: Parameter estimates for the best-fit Gaussian semivariogram model fitted to the spatial residuals (observed CFR minus predicted CFR adjusted for imputation) from the BRT model.

Spatial Parameters	Overall with bootstrap median, (95% CI)	Sierra Leone*	Guinea with bootstrap median, (95% CI)	Liberia with bootstrap median, (95% CI)
Range (km)	89.6 (33.3-99.69)	91.1 (40.9-95.0)	82.9 (30.4-131.8)	78.4 (34.3-89.6)
Sill (semivariance)	0.0158 (0.0151-0.0166)	0.0218 (0.0206-0.0233)	0.0173 (0.0161-0.0187)	0.0111 (0.0104-0.0121)
Nugget (semivariance)	0.0042 (0-0.0105)	0.0062 (0.0004-0.0120)	0.0031 (0-0.0122)	0.0005 (0-0.0068)

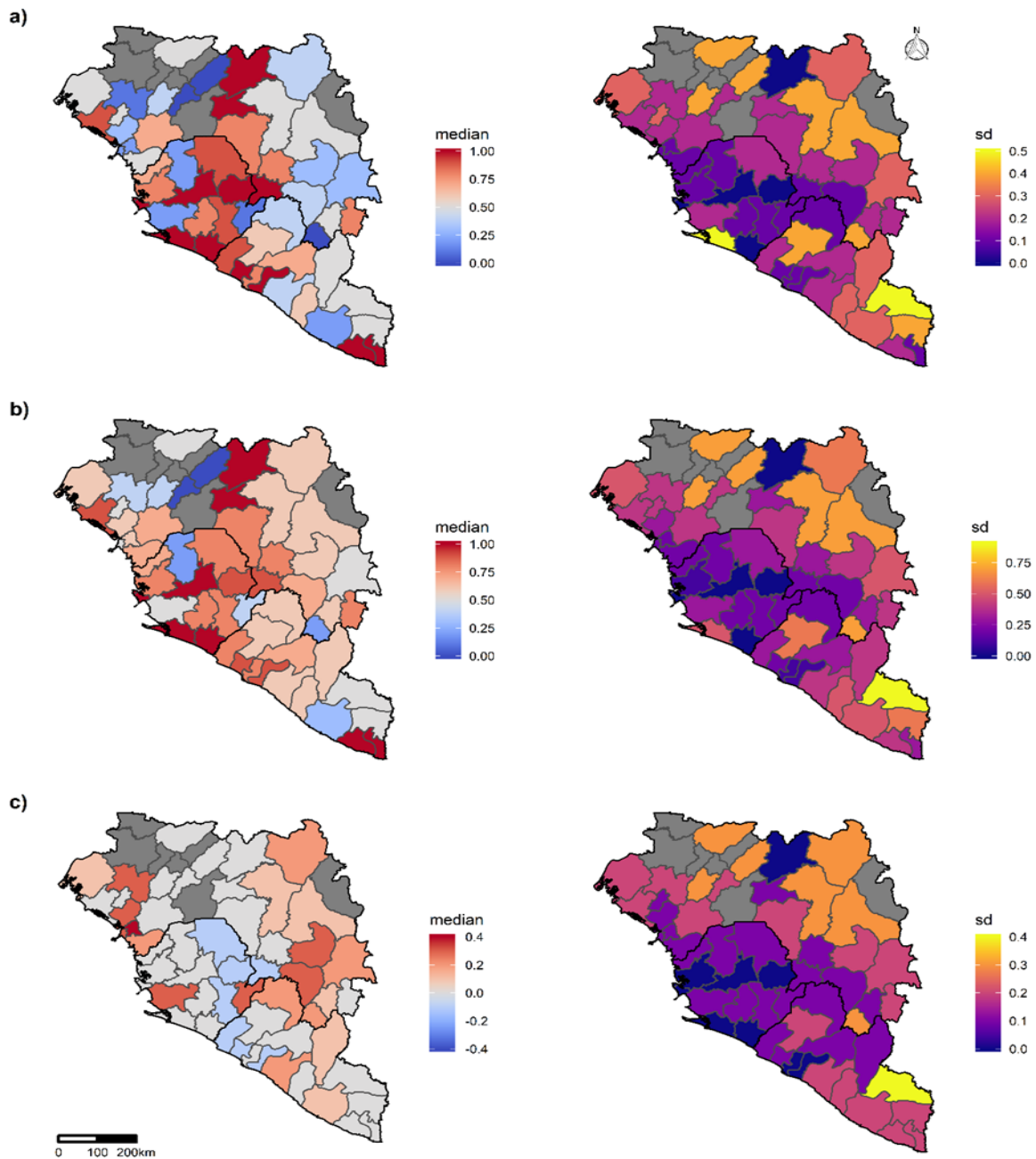


Figure 5. 1: Choropleth spatial distribution of case fatality ratio (CFR) for the 2013-2016 Ebola epidemic in West Africa. a) Median predicted CFR adjusted for imputation (and standard deviation) for each district. b) The median observed district-level CFR. Dark grey shading denotes districts for which data were unavailable. c) Median district-level residuals (i.e. observed CFR minus predicted CFR adjusted for imputation). In each case, the median is on the left and the standard deviation (sd) is on the right.

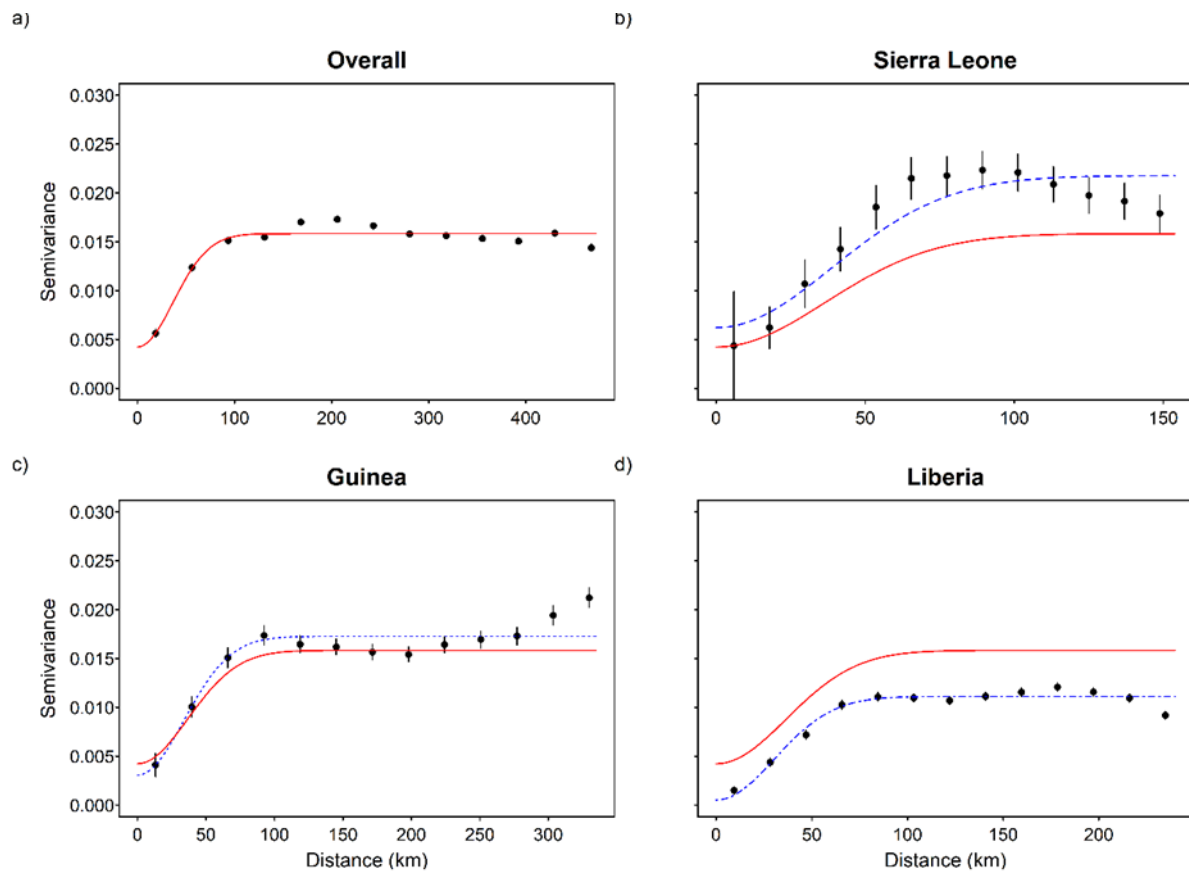


Figure 5. 2: Semivariograms fitted with a Gaussian model to the residuals of district-level CFR adjusted for imputation based on the Boosted Regression Tree model (BRT). The red line is the fitted model for the region as a whole and the blue lines are fitted models for individual countries (Sierra Leone, Guinea and Liberia). Note that the x-axes vary.

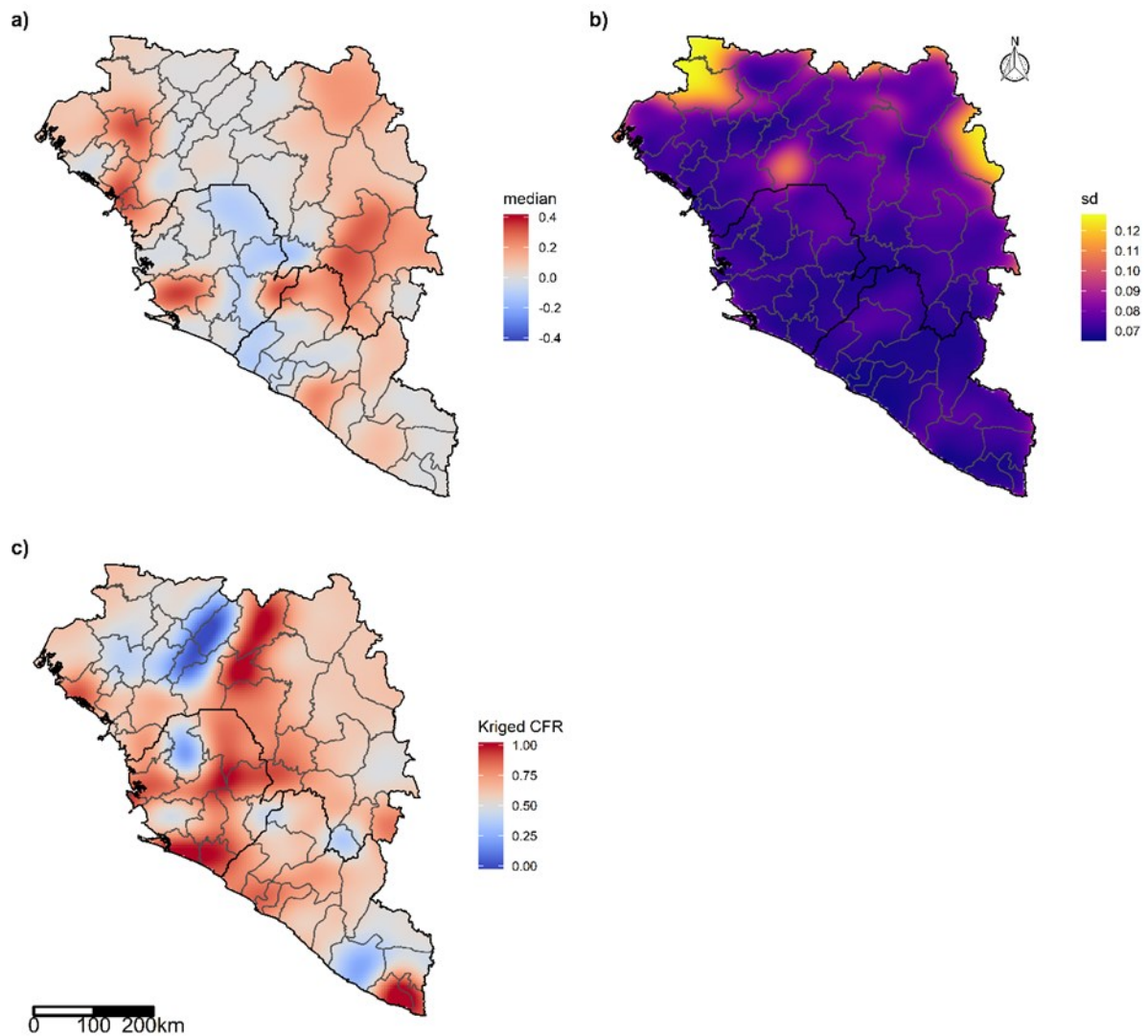


Figure 5. 3: Isopleth case fatality ratio (CFR) map for West Africa of a) median and b) standard deviation (sd) of kriged residuals (i.e. observed CFR minus predicted CFR adjusted for imputation). c) Isopleth map for kriged CFR (i.e. predicted CFR adjusted for imputation plus the kriged residuals).

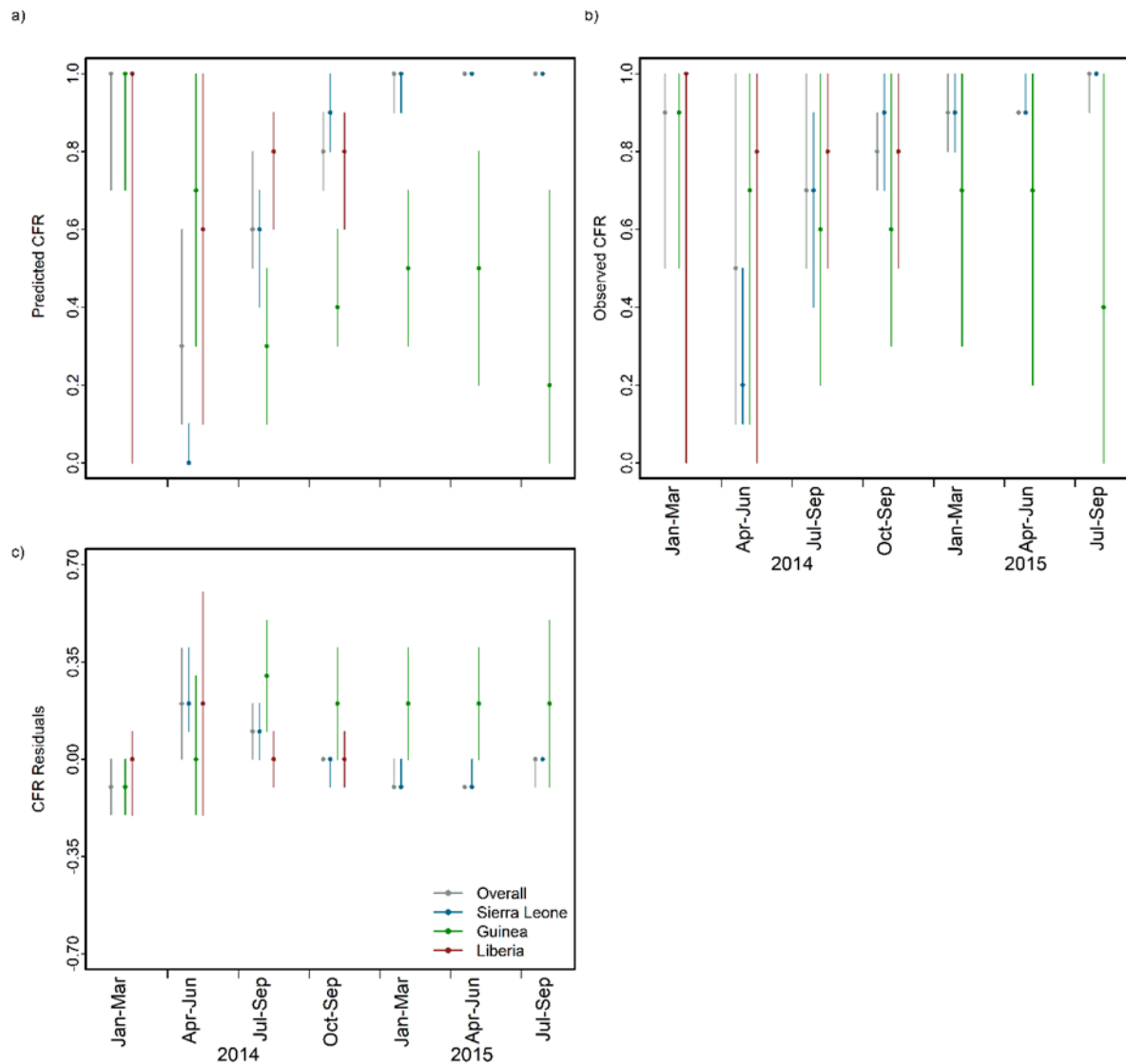


Figure 5. 4: Temporal distribution of case fatality ratio (CFR) for the 2013-2016 Ebola epidemic in West Africa. a) Median predicted CFR adjusted for imputation (and 95% CI) for each quarter. b) The median observed quarterly CFR. c) Quarterly residuals (i.e. observed CFR minus predicted CFR adjusted for imputation). Note that there were too few cases in late 2013 and early 2016 to be meaningfully aggregated into quarter.

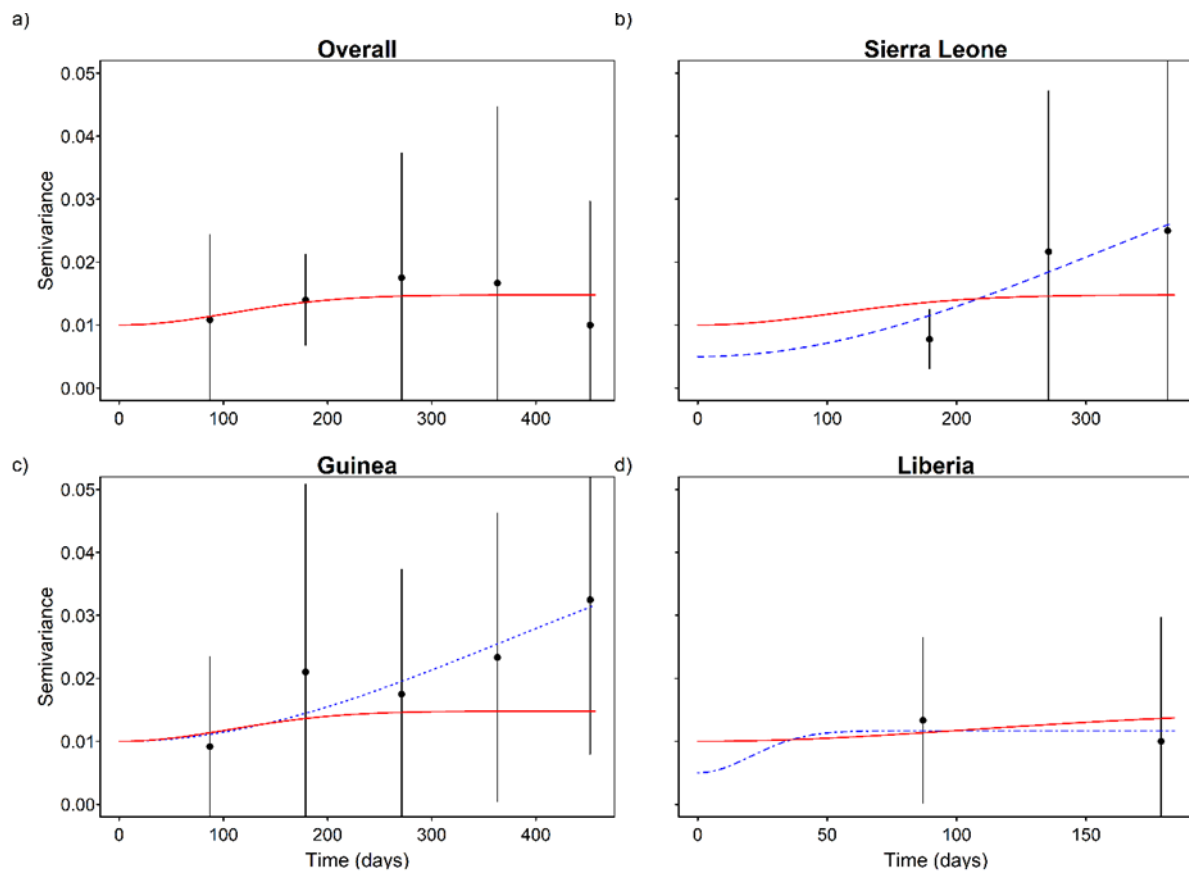


Figure 5. 5: Semivariograms fitted with a Gaussian model to the residuals of quarter-specific CFR adjusted for imputations based on the Boosted Regression Tree model (BRT). The red line is the fitted model for the region as a whole and the blue lines are fitted models for individual countries (Sierra Leone, Guinea and Liberia). Note that the x-axes vary.

5.4 Discussion

Accounting for the entire EVD patient population including those with missing survival outcomes, this chapter represents the first detailed investigation of the spatiotemporal heterogeneity of CFR for the 2013–2016 Ebola epidemic in West Africa. This analysis characterises the magnitude of heterogeneity arising in a large epidemic and may be used to inform public health preparedness for future epidemics. In the future, a similar analytical framework might be used to improve situational awareness during ongoing outbreaks, such

as that currently underway in the Democratic Republic of the Congo. This analytical framework could be extremely informative in real time if case referral routes are also recorded during outbreaks. Case referrals routes will help disentangle whether survival outcome is reported as promptly for cases who survive to be discharged from care as for deaths in communities.

As observed previously, district-level CFR varied extensively across Sierra Leone, Guinea and Liberia [17, 78]. Indeed, Forna *et al.* identified district of origin as the most important predictor in explaining the observed variation in CFR [91]. Here I found that even after accounting for district-level variation in the models, there remained significant (Moran's I Index=0.146, $p=0.03$) regional and country level (except for Sierra Leone) clustering of the CFR residuals and spatial independence was achieved only beyond 89.6 km (95% CI, 33.3–99.69 km). Thus, the distribution of CFR residuals was unlikely to have resulted from chance alone (regional Moran's I Index $p=0.04$). Other sources of spatial autocorrelation could be linked to the omission of potentially important predictors [139, 140]. However, because the residuals of the BRT model excluding district as a predictor showed even more significant autocorrelation (Moran's I Index=0.340, $p<0.001$) and BRT is a non-parametric approach (i.e. captures both linear and non-linear relationships in data), I conclude that, at least for this dataset, the effects of the district explain only some of the spatial variation in Ebola CFR.

In resource-constrained settings, as was the case in West Africa, the considerable spatial heterogeneity in district-level CFR could potentially be further explained by variations in case reporting practices across Sierra Leone, Guinea and Liberia. For instance, Liberia switched from an aggregate system of reporting to an individual-based system of reporting that could have led to preferential reporting of survival outcomes over time [77]. While in

Sierra Leone, survival outcome reporting rates varied considerably throughout the epidemic [78]. Inconsistent case reporting appears to be contributing to spatial heterogeneity in EVD CFR underscoring the need for the strengthening and harmonising of outbreak surveillance systems. Effective surveillance systems would improve data quality and subsequently, help us disentangle the contributions from the different mechanisms that may give rise to genuine differences in CFR estimates. During case reporting, a greater emphasis on survival outcome reporting would serve better future surveillance.

The kriging procedure yielded two advantages, as previously discussed by Carrat and Valleron [136]. Firstly, CFR estimates were not constrained by the borders between administrative units, and sudden transitions in CFR between two neighbouring districts were avoided. Thus, these maps are robust to future boundary changes. Both the BRT model and kriging allowed us to interpolate CFR values for districts without observed survival data (e.g. Koubia, Lelouma, Labe and Mamou in Guinea) giving more complete CFR maps for the epidemic. In these districts, predicted CFR adjusted for imputation reflected the patterns observed in areas for which data were available.

Future work could compare these CFR estimates for every point on the map with additional localised epidemiological and demographic parameters. The identified spatial patterns can serve as a baseline for future preparedness and evaluation of control measures in the event of another Ebola outbreak in West Africa. The variability in the CFR across districts could inform the allocation of resources for control activities, with the possibility of focussing additional efforts in areas with higher CFRs [141]. Furthermore, future work could also consider providing MaLe algorithms with predictors that could identify spatial neighbours.

Results from this algorithmic approach could be compared with those from the kriging procedure to investigate the extent to which both approaches capture spatial correlation.

The findings from this study show no temporal pattern in CFR over the seven quarters of the epidemic, from January 2014 to September 2015, after accounting for the parameters in BRT model. This indicates that the BRT model captures all of the temporal correlation.

There are some limitations to this study that need to be considered when interpreting our results. Firstly, our findings do not explicitly reflect the variation in smaller administrative units such as chiefdoms, districts (the sub-administrative level in Liberia) and sub-prefectures in Sierra Leone, Liberia and Guinea, respectively. With more detailed data on patient residence, it would be interesting to compare our current results with those obtained based on the analysis of smaller administrative units. Secondly, unrecorded (including asymptomatic) EVD [6] cases were not accounted for in our analysis. However, Forna *et al.* discussed the imperfect nature of the imputation process such that the uncertainty in CFR is reflected throughout our analysis [91]. Using kriging based on estimated semivariograms for interpolation of CFR estimates for districts without survival data is not assumption-free. After including known predictors and adjusting the BRT model for imperfect sensitivity and specificity, in interpolating CFR, I assume that there are not unknown structural reasons for the districts without survival outcomes to be importantly different from districts with observed CFRs.

In this study, the georeferenced grid unit-size was chosen to fit within the limit of the available computational resource. However, grid unit-size could be chosen to meet other computational and/or epidemiological criteria. For example, for this study, the grid unit size could have been selected to ensure a minimum number of cases by amalgamating districts.

In conclusion, these results demonstrate that although BRT modelling accounted for most of the spatiotemporal variations and interactions in CFR, residual spatial autocorrelation remained. In contrast, residual temporal autocorrelation did not. Kriging was used to produce a more complete CFR map of the epidemic. This study adds to the literature that analysed the epidemiology of the 2013–2016 West African Ebola epidemic to inform future public health contingency planning, resource allocation and impact assessment. The study also provides an analytical framework, which when used with other resources and key domain knowledge, could provide real time support to the response to ongoing outbreaks, such as that currently underway in the Democratic Republic of the Congo.

5.5 Additional Results

5.5.1 Moran's I indices for CFR residuals

Table 5.2 shows the local autocorrelation for districts in Sierra Leone, Guinea and Liberia. I consider districts with p-values < 0.05 (i.e. p-values shaded red) as districts for which CFR is spatially correlated to that of neighbouring districts. Here I fitted the BRT model with and without the district predictor in the model.

Table 5. 2: Spatial autocorrelation analysis for case fatality ratio (CFR) residuals in Sierra Leone, Guinea and Liberia from 2013 to 2016, with/without district predictor in the BRT model. The global Moran's I is for the three countries combined. The p-values in red are statistically significant, <0.05.

District	With district as a predictor in the BRT model		Without district as a predictor in the BRT model	
	Moran's I	p-value	Moran's I	p-value
Global	0.146	0.033	0.348	<0.001
Guinea				
BOFFA	-0.019	0.969	0.011	0.933
BOKE	-0.016	0.978	0.011	0.933
FRIA	-0.019	0.969	-0.055	0.850
CONAKRY	-0.084	0.784	0.475	0.067
DABOLA	0.158	0.588	0.115	0.682
DINGUIRAYE	-0.133	0.647	-0.147	0.610
FARANAH	0.013	0.925	-0.036	0.925
KISSIDOUGOU	0.088	0.735	0.420	0.157
KEROUANE	0.536	0.056	0.812	0.004
KANKAN	0.207	0.441	0.032	0.874

KOUROUSSA	0.273	0.283	0.222	0.380
SIGUIRI	0.095	0.663	0.177	0.438
COYAH	0.536	0.056	0.520	0.064
DUBREKA	0.007	0.941	0.217	0.423
FORECARIAH	-0.101	0.762	-0.084	0.808
KINDIA	0.187	0.484	-0.004	0.974
TELIMELE	-0.181	0.521	-0.147	0.610
MALI	-0.118	0.741	0.004	0.949
TOUGUE	0.062	0.781	0.078	0.737
DALABA	-0.084	0.784	0.078	0.737
PITA	0.207	0.441	-0.034	0.945
BEYLA	-0.908	<0.001	0.361	0.122
GUECKEDOU	-0.165	0.516	0.411	0.079
LOLA	0.013	0.926	0.627	0.026
MACENTA	0.173	0.486	0.904	0.001
NZEREKORE	0.885	0.002	0.910	0.001
YAMOU	0.173	0.486	0.222	0.380

Sierra Leone

KAILAHUN	-0.685	0.019	0.003	0.952
KENEMA	0.772	0.003	0.437	0.091
KONO	0.403	0.118	0.344	0.180
BOMBALI	0.094	0.708	0.479	0.087
KAMBIA	0.013	0.925	0.698	0.008
KOINADUGU	0.441	0.137	0.785	0.009
PORTLOKO	0.045	0.837	0.709	0.012
TONKOLILI	0.289	0.292	0.930	<0.001
BO	0.174	0.535	0.704	0.019
BONTHE	0.013	0.925	0.212	0.399
MOYAMBA	-0.457	0.061	-0.015	0.981
PUJEHUN	0.257	0.311	0.212	0.399
WESTERN	0.013	0.925	-0.102	0.733
Liberia	-			
BOMI	0.729	0.010	-0.012	0.995
BONG	0.077	0.762	0.057	0.812
GBAPOLU	0.012	0.926	0.072	0.781
GRANDCAPEMOUNT	0.852	0.003	0.210	0.437

GRANDBASSA	-0.362	0.140	0.215	0.351
GRANDGEDEH	-0.052	0.894	-0.388	0.194
GRANDKRU	0.143	0.586	0.299	0.276
LOFA	0.408	0.142	0.812	0.004
MARGIBI	0.199	0.384	0.170	0.455
MARYLAND	-0.019	0.969	-0.144	0.573
MONTSERRADO	0.223	0.436	0.252	0.383
NIMBA	0.105	0.680	-0.329	0.274
RIVERCESS	0.028	0.885	0.162	0.562
RIVERGEE	0.079	0.713	-0.020	0.964
SINOE	-0.067	0.808	0.035	0.853

5.5.2 Selecting a spatial model for the analysis

To identify the best theoretical spatial model to fit the empirical semivariogram, I estimated the AIC for five model types (i.e. Linear, Cauchy, Exponential, Matern and Gaussian). Table 5.3 shows the AIC values for each spatial model and because the Gaussian model yielded the smallest AIC, I used the Gaussian model for subsequent analysis.

Figure 5.6 shows the theoretical semivariograms (i.e. linear, Cauchy, exponential and Matern) fitted to the empirical semivariogram of Sierra Leone, Guinea and Liberia combined. Except for the linear model, the other three models also fit well to the empirical

semivariogram. However, the “gboot_variogram” function in R estimates the uncertainty for only the Gaussian and the exponential model. The range [89.6 km (95% CI, 33.3–99.69 km)] of the Gaussian model is more precise than that of the exponential model [112.0 km (95% CI, 0–112.0 km)]. These results together with the smaller AIC for the Gaussian model, affirms our choice of using the Gaussian model for subsequent analysis.

Table 5. 3: Spatial semivariogram models fitted to overall district-level CFR residuals.

Spatial models	AIC
Linear	-46.8112
Cauchy	-62.0444
Exponential	-62.9812
Matern	-62.9812
Gaussian	-64.2103

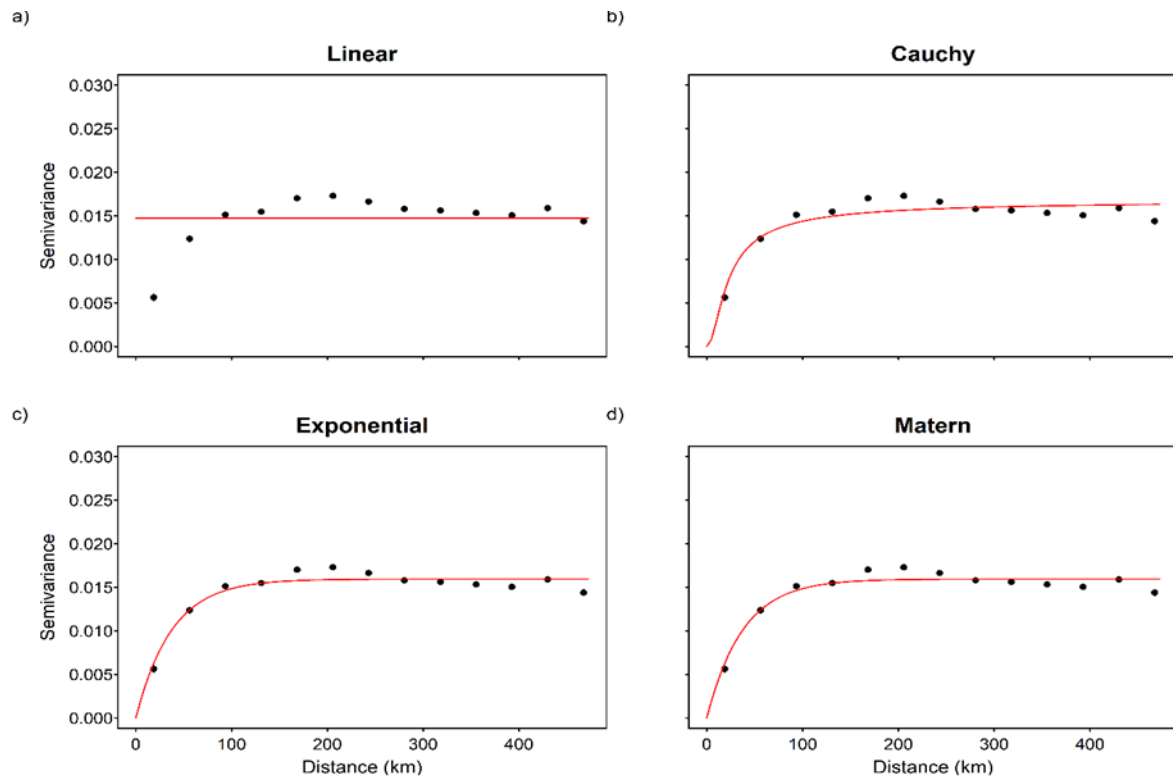


Figure 5. 6: Theoretical semivariograms fitted to the residual empirical semivariogram for Sierra Leone, Guinea and Liberia combined.

5.5.3 Spatial semivariogram for observed CFR residuals

Figure 5.7 shows the spatial semivariogram for observed CFR residuals (iq) without the adjusted imputation for cases with unknown outcomes. These residuals except for Sierra Leone (where semivariance tended to decrease at longer distances), also show spatial autocorrelation. However, because I excluded cases with unknown outcomes, the observed residuals are not fully representative of the patient mix. Thus, BRT imputation makes the CFR estimates more representative and allows for better characterisation of the spatiotemporal heterogeneity in district-level CFR. We did not proceed to generate kriged maps for these residuals.

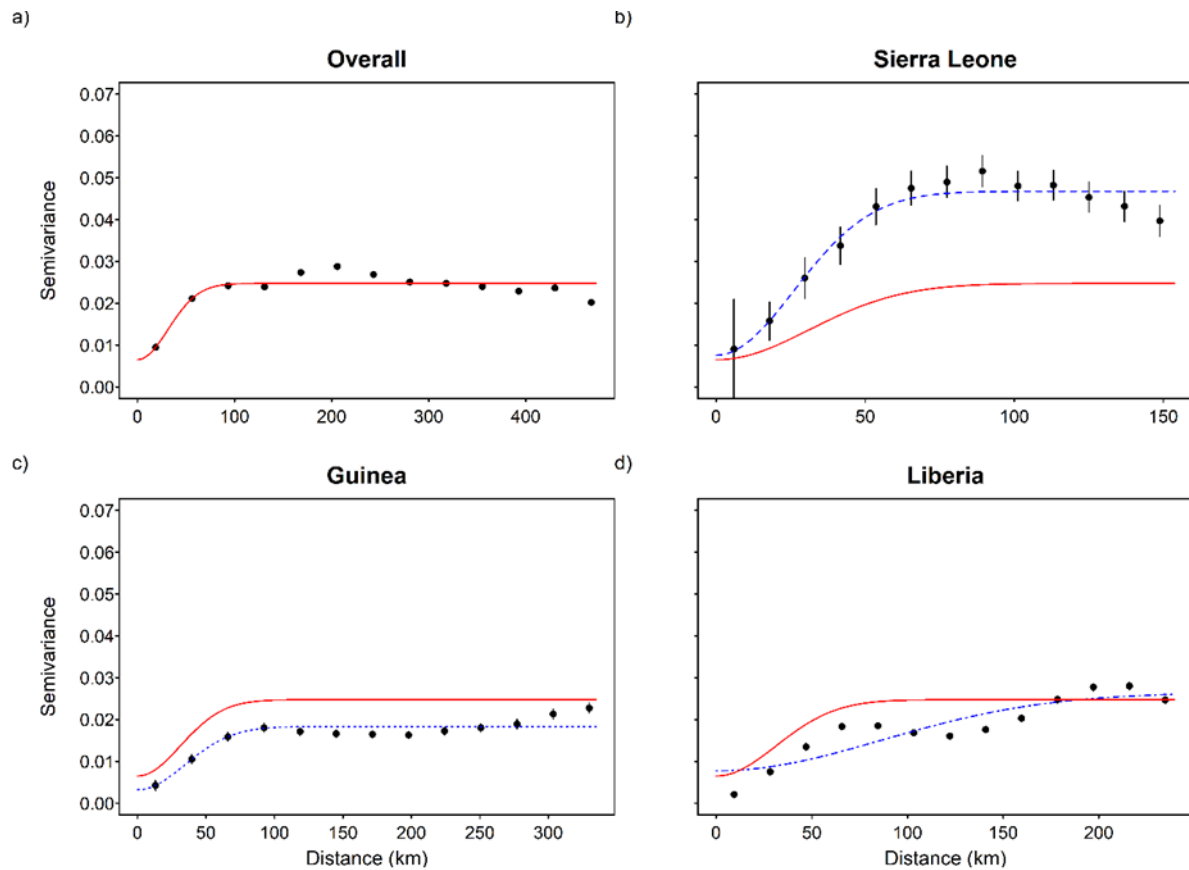


Figure 5. 7. Semivariograms fitted to the residuals of Observed CFR. Red line fitted model for overall, blue lines fitted models for individual countries.

5.5.4 Temporal semivariogram for observed CFR residuals

Figure 5.8 shows the temporal semivariogram for observed CFR residuals (qq) without the adjusted imputation for cases with unknown outcomes. Similar to the residuals for observed CFR with adjusted imputations, these residuals do not show temporal autocorrelation.

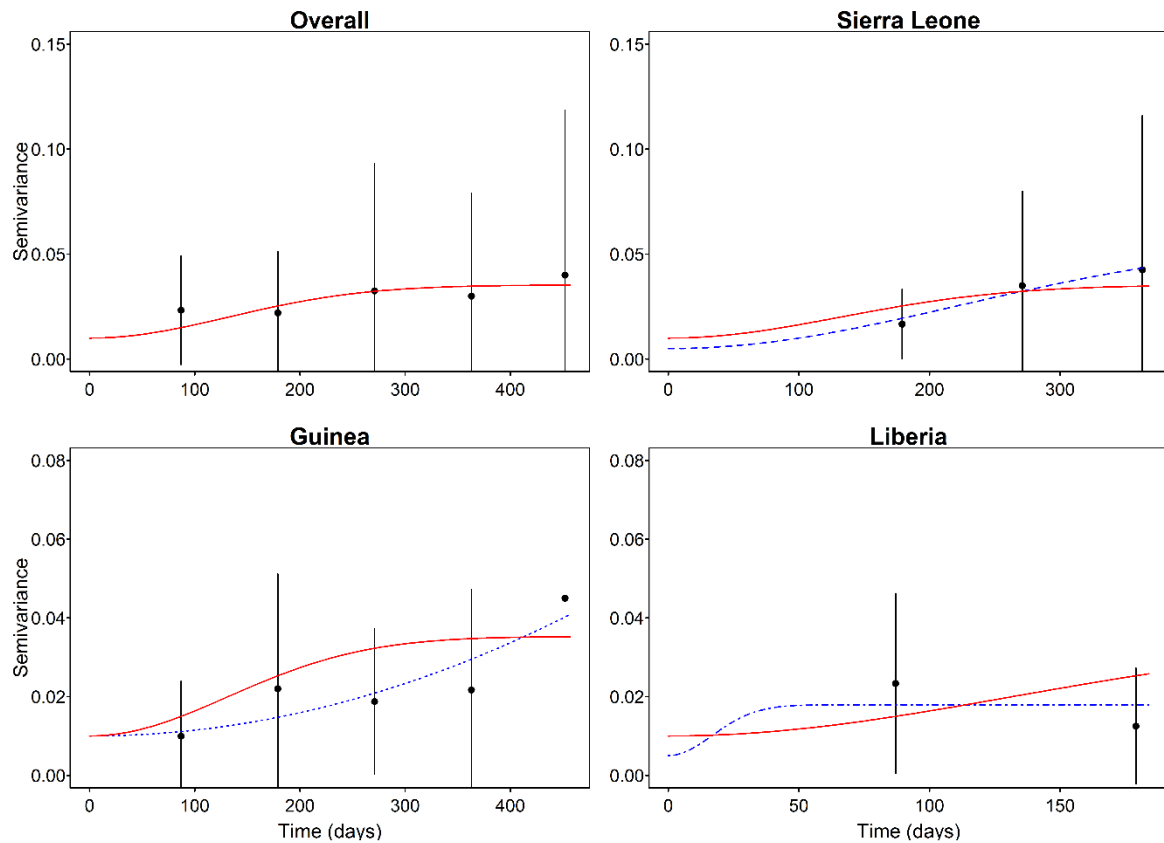


Figure 5. 8. Semivariograms fitted to the residual of observed CFRs. Red line fitted model for overall, blue lines fitted models for individual countries. Note that the x-axes vary.

5.5.5 Complete map of predicted CFR

Figure 5.9a shows the isopleth (i.e. continuous grid) of the predicted CFR adjusted for imputation. The kriged residuals were added to these predictions to produce the complete unbiased map presented in Figure 3c in the main manuscript. Regions with relatively high uncertainty (standard deviation) show districts for which data were unavailable (Figure 5.9b).

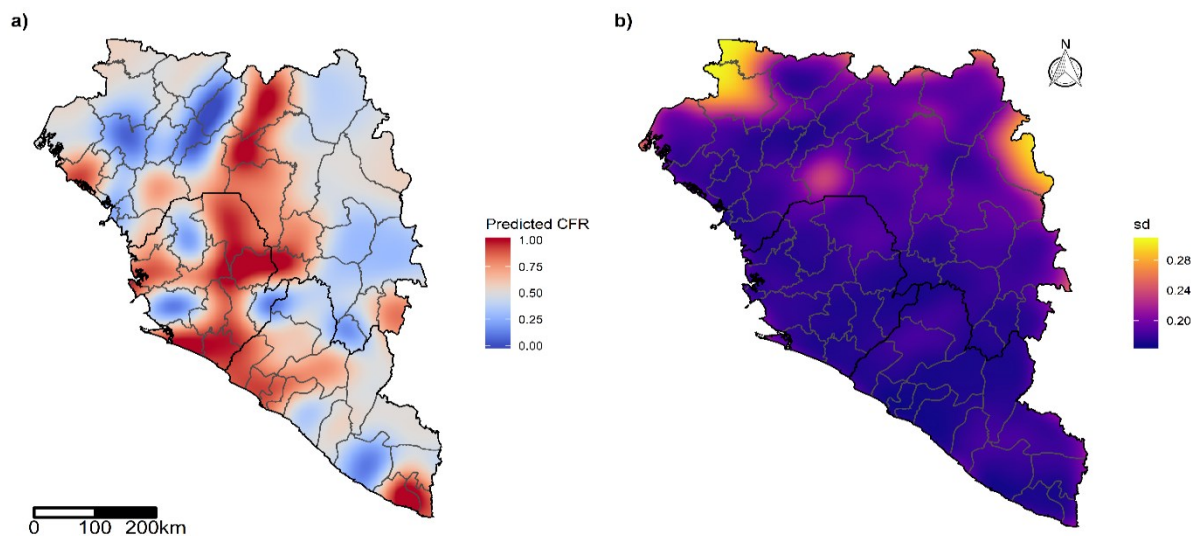


Figure 5. 9. Isopleth map for predicted CFR adjusted for imputation.

5.5.6 Directional semivariograms

Coupled with distance, the direction could also have influenced the semivariance between districts [55]. To investigate anisotropy, I estimated semivariograms for different directions, 0° , 45° , 90° and 135° (Figure 5.10). The directional semivariograms did not differ substantially from the omnidirectional variogram; thus, I assumed an isotropic spatial process and report omnidirectional semivariograms throughout this study.

To gain more insight into the isotropy properties, I perform a nonparametric hypothesis test of isotropy using the georeferenced district-level CFR residuals. I used the “GuanTestGrid” function from the R package “spTest” to carry out the nonparametric hypothesis test [142]. The resulting insignificant p-value (0.33) confirms conclusions drawn from the directional semivariograms that the dependence between district-level CFR residuals is a function of only the distance and not the direction of the spatial separation between sampling locations.

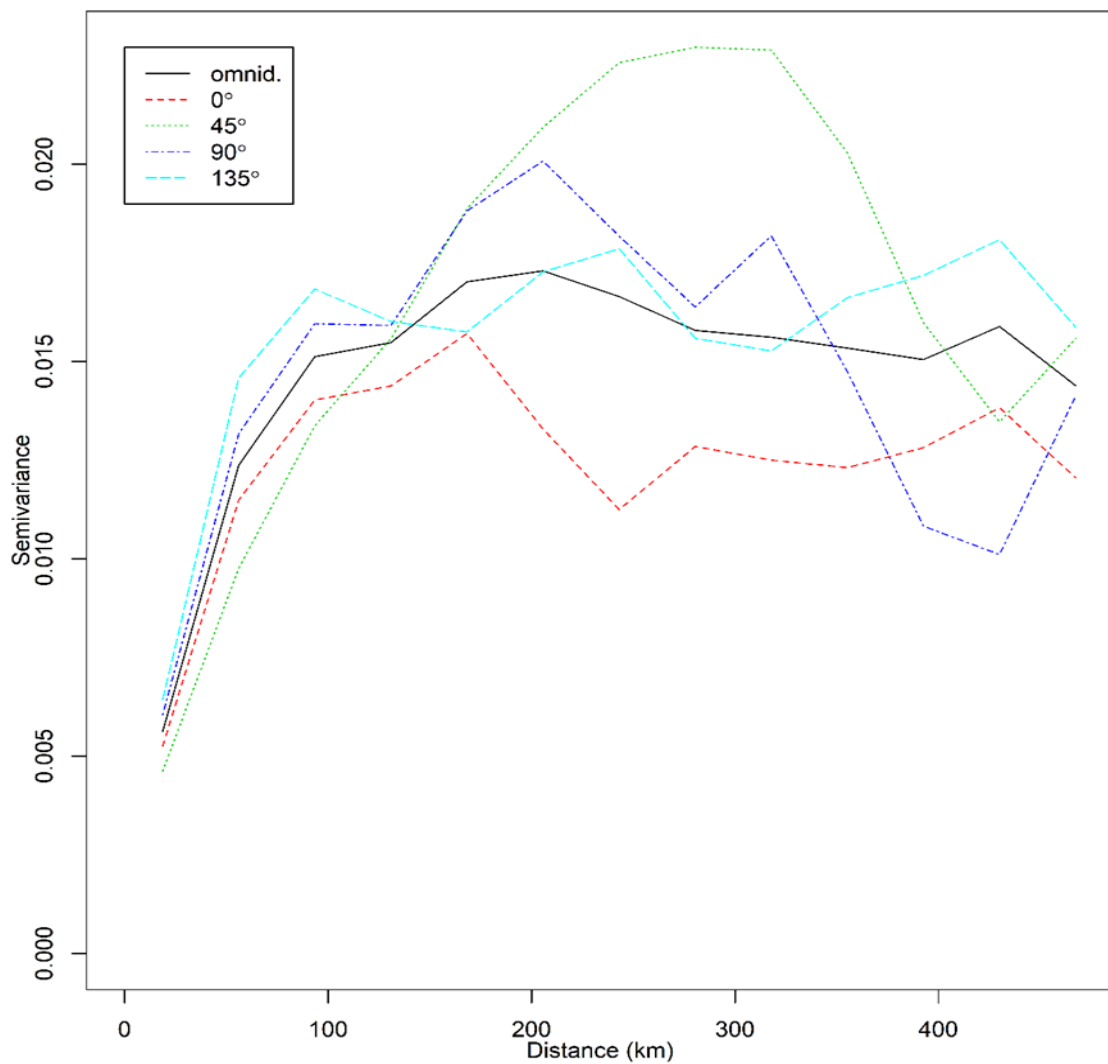


Figure 5. 10. Directional semivariograms for overall (i.e. Sierra Leone, Guinea, and Liberia).

5.5.7 Comparing Gaussian and exponential models in mapping CFR

Although I fitted a Gaussian model to the semivariograms in the main analysis, as a sensitivity analysis, I show that maps of residuals and kriged CFRs are similar for both the Gaussian and exponential models (Figure 5.11 and Figure 5.12).

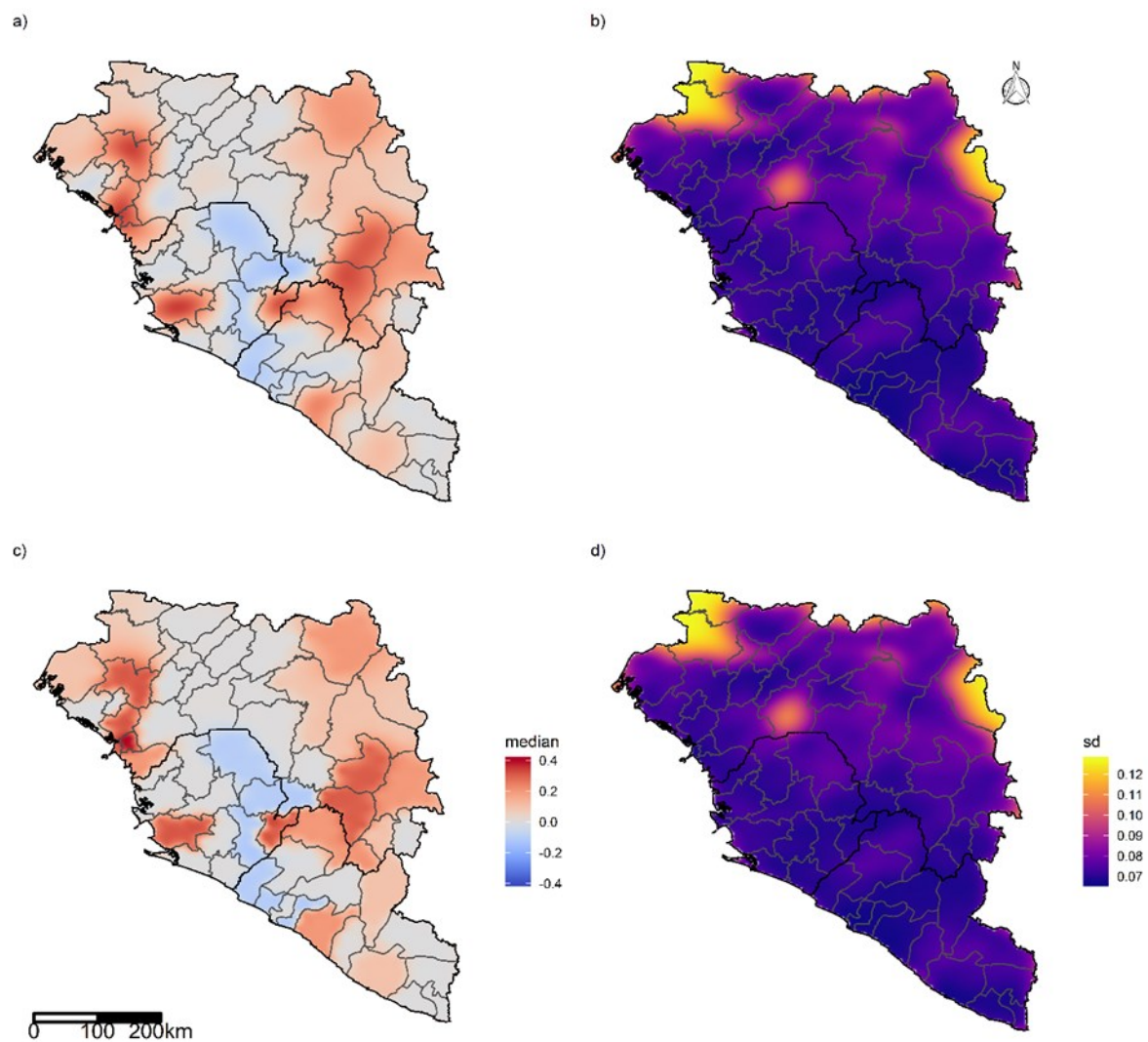


Figure 5. 11: Isopleth maps for kriged residuals. [a–b] median and standard deviation for the overall semivariogram fitted with a Gaussian model. [c–d] median and standard deviation for the overall semivariogram fitted with an exponential model.

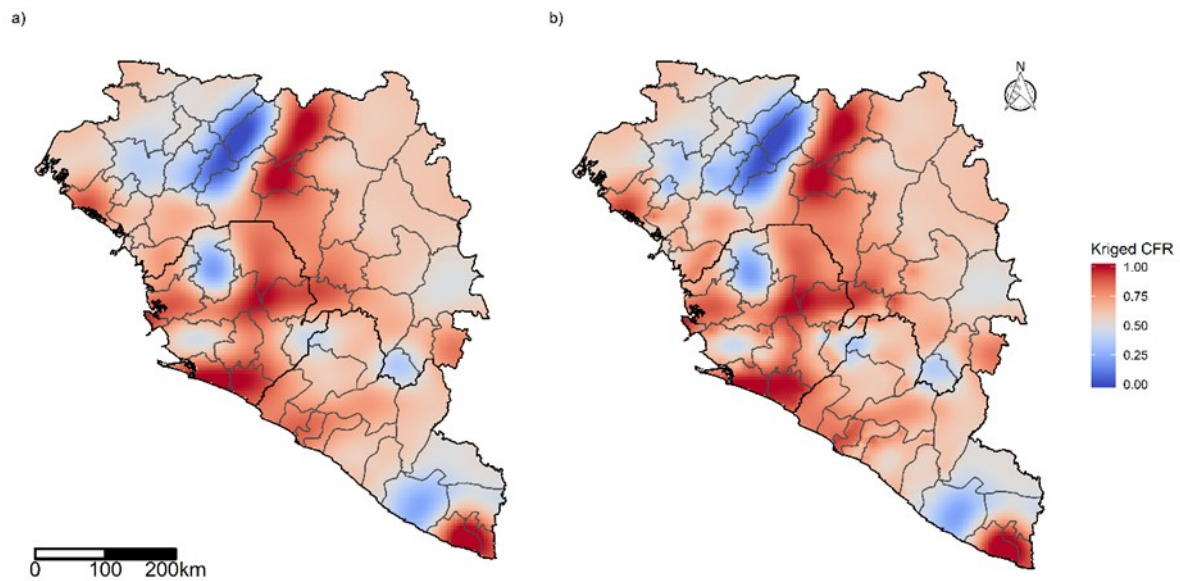


Figure 5. 12: Isopleth map for kriged CFRs (i.e. predicted CFR adjusted for imputation plus the kriged residuals). a) Kriged CFR for semivariogram fitted with a Gaussian model. b) Kriged CFR for semivariogram fitted with an exponential model.

5.5.8 Number of cases and number of known deaths per district

The total number cases per district are presented in Figure 5.13. For confirmed, probable and suspected cases, the Western district in Sierra Leone, that includes Freetown (the capital city) and sub-urban areas, reported the 8813 cases the highest for the three countries combined. Dinguiraye, in Guinea, reported just one case, the lowest for all three countries combined.

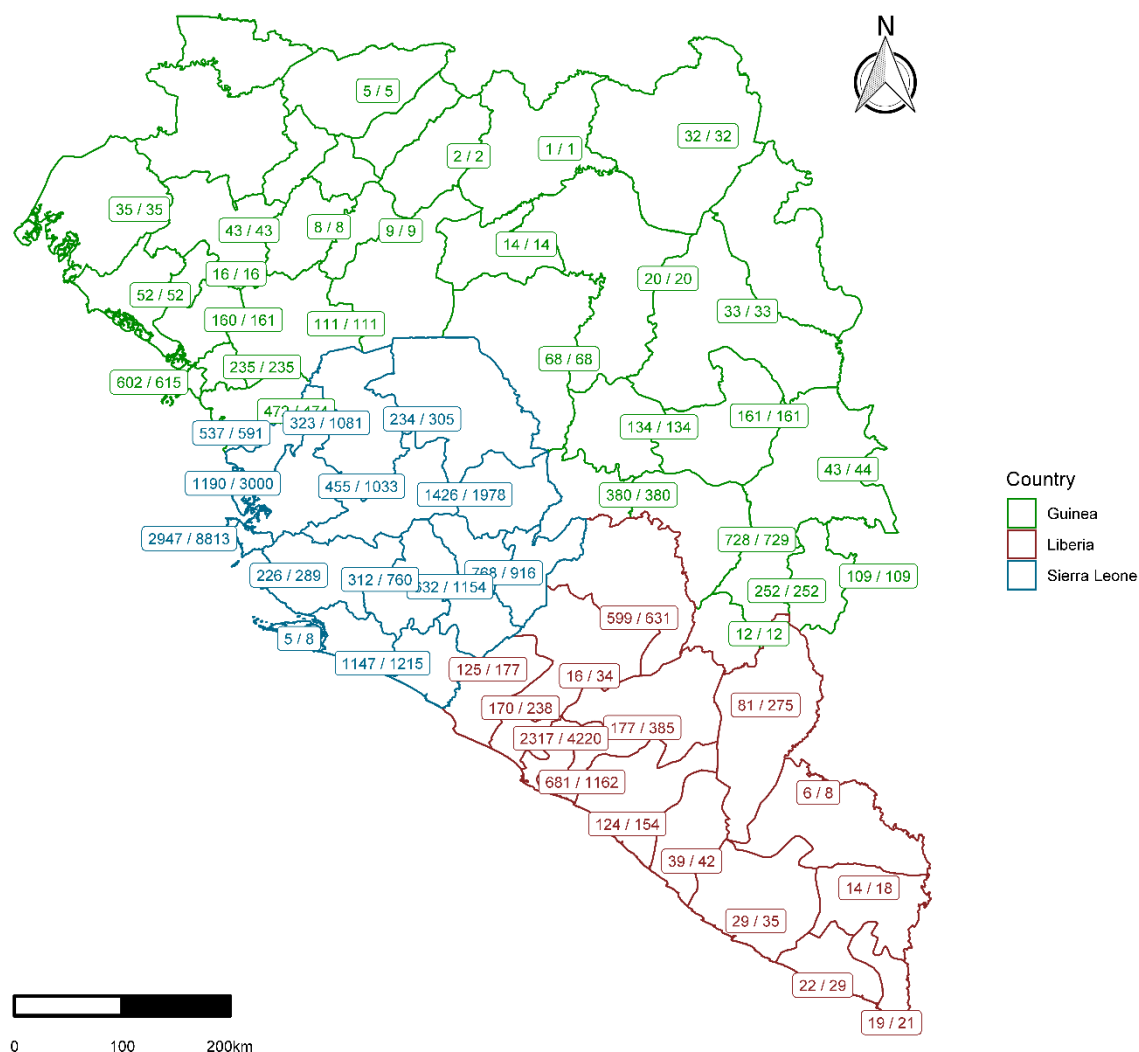


Figure 5. 13: Ebola virus disease cases (i.e. confirmed, probable and suspected) with known deaths/ total number of cases (i.e. confirmed, probable and suspected) per district for Sierra Leone, Guinea and Liberia.

The known deaths per district are also presented in Figure 5.13. For confirmed, probable and suspected cases, the Western district in Sierra Leone, that includes Freetown (the capital city) and sub-urban areas, reported 2947 deaths the highest for the three countries combined.

Chapter 6: Comparison of machine learning inference methods for estimating case fatality ratios: An Ebola outbreak simulation study

In this chapter, expanding on the algorithmic framework in Chapters 4 and 5, and using simulated data, I examined five MaLe (i.e. LR, RF, BRT, BART and ANN) inference methods to evaluate data imputation performance and the resulting case fatality ratio (CFR) estimates; varying the outbreak dataset size, the scale and type of data missingness (i.e. missing completely random - MCAR, missing at random – MAR, or missing not at random - MNAR), and the model training/validation ratio. Across the board, I investigated if adjustment for the sensitivity and specificity of MaLe inference methods reduces the bias in CFR estimates.

6.1 Introduction

Machine learning (MaLe) algorithms – computer algorithmic techniques that learn patterns in data – are increasingly used in epidemiology [143, 144] in areas including diagnostics, predictive analytics, and missing data imputation [143-145]. Forna *et al.* used Boosted Regression Trees (BRTs), a MaLe approach, to impute missing survival outcomes, adjust imputations for model imperfection and combine the imputations with observed outcomes to re-examine case fatality ratio (CFR) estimates (i.e. the probability that cases die due to the infection) for the West African Ebola epidemic [91, 146].

MaLe inference methods typically achieve encouraging out-of-sample performance levels, often with minimal data pre-processing [147]. For example, BRTs allow for missingness in predictors, capture complex nonlinear relationships between outcomes and predictors, and are insensitive to outliers within the predictors [58]. However, a major challenge is the need skilfully to optimise hyperparameters to maximise out-of-sample performance. Additionally,

it is unclear how data characteristics, such as the type of missingness, influence the performance of MaLe algorithms and the estimates based, in part, on the imputed data.

Thus, I sought MaLe inference methods, like BART that are resilient to hyperparameters choices. However, the literature suggests that even when hyperparameters are minimally dependent on data, MaLe inference methods perform best when trained on data for the specific problem for which they are to be implemented [64]. Beyond the need to optimise MaLe algorithms before they are fit for purpose [148], a practical challenge in the use of MaLe algorithms for epidemiological analysis remains how they perform for various outbreak data characteristics.

In this simulation study, I characterise the inferential and predictive performance of five algorithms (i.e. logistic regression, random forest, BRTs, BART and ANN) to estimate CFR for Ebola virus disease (EVD) in the presence of missing survival outcome data. Specifically, I vary the simulated size of the outbreak dataset, the type and scale of data missingness, and the model training/validation ratio. Our aim was not to perform a technical appraisal of MaLe inference methods as in previous studies [143, 149], but rather to demonstrate how different MaLe inference methods behave under different (simulated) infectious disease outbreak data characteristics. I aim to help epidemiologists make more informed choices as they explore the use of the available MaLe toolbox.

6.2 Methods

6.2.1 Simulated Data

I simulate an outbreak dataset of 12,049 cases (i.e. complete case count of confirmed, probable, suspected cases) to mimic that of the 2013–2016 Ebola epidemic in West Africa [17]. I included 20 predictors and the survival outcome variable. Predictors include

demographic predictors such as age group and case classification and clinical predictors such as fever occurrence and unexplained bleeding (see additional Results Table 6.2 for the complete list). I let $\mathbf{X} = \{X_1, X_2, \dots, X_{20}\}$ be a vector of the predictors sampled with replacement from the observed dataset such that the distribution of each simulated predictor is similar to that of its observed progenitor. The survival outcome variable was simulated such that 76.5% died and 23.5% survived, similar to the pattern observed in the West African Ebola epidemic dataset [91]. Using the coefficients from a generalised linear model, the probability of survival Y was modelled as a function of ten predictors ($\mathbf{X}^s \subset \mathbf{X}$): age, country, reporting delay, case classification, hospitalisation status, quarter, difficulty breathing, fever, fatigue and anorexia. The uniform distribution with bounds of 0 and 1 was used to generate the survival probabilities. The generalised linear model could be written as:

$$Y = \mathbf{X}^s \beta + \mu$$

The parameter μ is the error distribution and β are coefficients of the model. β could be written as a 2×1 matrix as follows:

$$\beta = \begin{bmatrix} \beta_o \\ \beta_k \end{bmatrix}$$

The parameter β_o is the coefficient of the intercept and β_k are the coefficients of the simulated predictors.

Thus, β^T , the transpose of the 2×1 coefficient matrix for the simulated (i.e. generalised linear model) model is as follows:

$$\beta^T = \begin{bmatrix} -0.13 & 0.11 & 0.25 & 0.31 & 0.21 & 0.2 & 0.13 & 0.16 & 0.11 & 0.13 \\ 0.11 & 0.10 & 0.05 & 0.06 & -0.02 & 0.05 & -0.04 & -0.03 & 0.01 & -0.03 \\ -0.11 & 0.21 & 0.20 & 0.20 & 0.22 & 0.19 & 0.21 & 0.13 & 0.09 & -0.05 \\ -0.01 & -0.03 & & & & & & & & \end{bmatrix}$$

In this matrix, for the categorical predictors, age has 16 levels, country and case classification have three levels each, quarter has seven levels and all the other categorical predictors have 2 levels each. Reporting delay is a continuous predictor with mean = 5.68 days and standard deviation = 8.91 days (see additional Results Table 6.2 for details).

To these predictors in the simulated model, I added ten others ($X^a \subset X$): unexplained bleeding, confusion, joint pain, jaundice, conjunctivitis, vomiting, diarrhoea, headache, muscle pain and chest pain. These added predictors independent from the outcome, were used to fit the MaLe models.

Five algorithms were evaluated in this study.

6.2.2 Logistic regression

Logistic regression (LR) is a generalised linear model and is usually applied because of its simplicity and interpretability [150]. It performs well when the relationship between the logit of the outcome probability and the predictors is linear, but it could perform poorly if complex nonlinear relationships exist between the outcome and predictors. Prediction accuracy is reduced in the presence of missing data as LR cannot inherently handle missingness in either the predictors or the outcome variable.

6.2.3 Random Forest

Random forest (RF) is a decision tree-based model where a recursive binary splitting of data is carried out until a stopping criterion is reached [66]. Decision trees are built on bootstrap samples of the training data. However, with RF, instead of using the full set of predictors

each time a tree is built from a bootstrap sample, a random sample of m predictors is used [87]. This approach reduces the risk that RF models overfit the observed data. Thus, RF models can achieve high predictive performance levels when adequately parameterised. RF, as implemented in this study, can inherently handle missingness in predictors. RF's tuneable hyperparameters and their purposes are summarised in Table 6.1.

6.2.4 Boosted Regression Trees

A BRT model is also a decision tree-based model requiring recursive binary splitting of data [58]. BRT reduces the bias and inaccuracies of decision trees by applying randomness into the stage-wise fitting, then introduces stochasticity to help avoid overfitting of the model to data. A unique feature of BRT is that subsequent trees are fitted to the residuals of previous trees and then added to the existing fitted function to update the residuals [58]. With BRT, numerous shallow trees can be fitted, and the model learns slowly with these weak learners eventually to produce a powerful predictive model [87]. BRT, as implemented in this study, can inherently handle missingness in predictors. BRT's tuneable hyperparameters and their purposes are summarised in Table 6.1.

6.2.5 Bayesian Additive Regression Trees

BARTs can be used to characterise complex relationships between the outcomes and the predictors [151]. BARTs assemble weaker learners (i.e. shallow trees with low-level predictor interactions) each explaining a small part of the relationship between an outcome and its predictors using a prior distribution, and inference is accomplished via an iterative Bayesian back-fitting Markov chain Monte Carlo (MCMC) algorithm [151]. BARTs inherently incorporate high-order interactions between predictors and the outcome. BART models can be computationally expensive due to the Bayesian nature of their implementation and, as

implemented in this study, can inherently handle missingness in predictors. BART's tuneable hyperparameters and their purposes are summarised in Table 6.1.

6.2.6 Artificial Neural Networks

An ANN is a structured set of interconnected identical processing units (e.g. perceptron) that attempts to mimic the behaviour of biological neurons [152]. These interconnections send signals from one unit to another. Signals are information from predictors which are then weighted through the network to predict the outcome. Numerous interconnected ANN neurons with updated weights “learn” complex relationships between the predictors and the outcome. While ANNs could have complex architecture, the simplest ANN (implemented in our simulation) consists of one input layer, one hidden layer, and one output layer. The power of ANN comes from the nonlinearity of the hidden neurons thus ANN could potentially achieve high predictive performance levels [152]. Fitting ANNs can be challenging and requires experience in optimising hyperparameters. Furthermore, ANN, as implemented in this study, cannot inherently handle missingness in predictors. ANN's tuneable hyperparameters and their purposes are summarised in Table 6.1.

Table 6. 1: Summary of the machine learning algorithms and hyper-parameters investigated.

Inference method and	Hyperparameters	Function of the hyperparameter in the model	Hyperparameter
MaLe algorithms	tuned		space considered (lower bound, upper bound)
Logistic Regression (LR)	–		–
Random Forest (RF)	ntree	Number of trees	(100,1000)
[153]	mtry	Number of predictors randomly selected as candidates for splitting a node	(3,10)
	nodesize	Forest average number of unique data points in a terminal node	(10,30)

Boosted Regression	ntrees	Integer specifying the total number of trees to fit	(100,2000)
Trees (BRTs) [81]	interaction.depth	Integer specifying the maximum depth of each tree (i.e., the highest level of predictor interactions allowed)	(2,10)
	bag.fraction	Fraction of training dataset observations randomly selected for each tree. This introduces random variation into the model fit	(0.5,0.75)
	shrinkage (Also known as the learning rate)	Shrinkage parameter applied to each tree	(0.001,0.05)
Bayesian Additive	num_trees	The number of trees to be grown in the sum-of-trees model.	(5,20)
Regression Trees (BART) [154]	num_burn_in	Number of MCMC samples to be discarded as ``burn-in"	(10,30)
	num_iterations_after_burn_in	Number of MCMC samples to draw from the posterior distribution of the fitted function	(100,300)
	beta	Power hyperparameter in tree prior distribution for whether or not a node is nonterminal	(1,2)

Artificial Neural	size	Number of units in the hidden layer	(5,20)
Network ANN [155]	decay	Parameter for weight decay	(10^{-8} ,0.002)
	abstol	Value below which the modelling fitting is stopped to prevent overfitting	(0.001,0.002)
	reltol	If the optimizer is unable to reduce the fit criterion by a factor of at least (1 – reltol), modelling fitting is stopped	(10^{-8} , 10^{-5})

6.2.7 Outbreak data characteristics investigated for this study

I briefly describe the outbreak data characteristics investigated in this simulation study.

6.2.7.1 Outbreak dataset size

Infectious disease outbreak sizes vary from one outbreak to another. For example, as seen in the ongoing Ebola outbreak in the Democratic Republic of the Congo [19], large Ebola outbreaks are no longer regarded as one-off and, in an era of increasing human mobility and interconnectivity, without adequate preventive and control measures, large Ebola epidemics could become more common. In this simulation study, to investigate outbreak dataset size, I down sampled the simulated dataset from 100% of cases (12,049 cases) to 75% (9,037 cases), and 50% (6,025 cases) of cases and tested in each scenario for model performance and CFR estimation.

6.2.7.2 Type and scale of missingness

Infectious disease outbreak datasets are almost never complete, and this can introduce biases in estimating epidemiological parameters (e.g. CFR) from outbreak data. Data can be missing for many reasons in an outbreak setting. For example, surveillance teams cannot always cover hard to reach communities and in providing critical care in very challenging circumstances, clinicians often fail to record some information on case reporting forms.

Data missingness can be classified into three mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [79]. MCAR means that the data are missing non-systematically and that any sample drawn from the data is representative of the underlying population [79]. MAR means that the probability of missingness in a predictor is conditional on predictors (i.e. explanatory predictors) in the

data. Finally, MNAR means that the probability of missingness in a predictor is conditional on the outcome (i.e. dependent variable). Our simulation study tests the robustness of the MaLe algorithms in estimating unbiased CFRs given the mechanism underlying the missingness.

To investigate the proportion of missingness, I generated 10%, 20% and 40% of each type of missingness in the survival outcome variable. For each proportion of outcome missingness, the same proportion of missingness in the predictors is simulated as MCAR (i.e. assuming that the missingness in all the predictors is MCAR).

6.2.7.3 Model training/validation ratio

The training/validation ratio (proportion of training data) has a direct influence on model hyperparameterisation [87]. In this simulation study, I tested three training/validation splits of the datasets: (i) 50% data for training and 50% for validation; (ii) 65% data for training and 35% for validation; and (iii) 80% data for training and 20% for validation.

6.2.8 Experimental Setup

I simulated the survival outcomes with ten predictors but fitted models with 20 potential predictors. To investigate whether the inference methods eliminate the ten irrelevant predictors from the simulation, I estimated how important individual predictors were by contrasting inferential performances. Inferential performance, measured as the mean squared error (mse), characterises the change in performance from permuting the values of each predictor and comparing those to predictions made on the unpermuted simulated data [102].

In fitting each inference method to the data, I performed 5-fold cross validation. In cross validation, the dataset is divided into k sub-samples (in our case $k = 5$). A single sub-sample

is chosen as testing data, and the remaining $k - 1$ sub-samples are used as training data. The procedure is repeated k times, in which each of the k sub-samples is used exactly once as the testing data [156]. The k results are averaged, and the resulting single estimate is used to evaluate each inference method during hyperparameter optimisation.

For each inference method considered (LR, RF, BRT, BART and ANN), I used the literature to guide the choice of upper and lower bounds of the hyperparameter values used to optimise performance (Table 1) [58, 66, 150-152]. Conditional upon these bounds, 50 random grid searches were carried out to identify the optimised hyperparameters for each algorithm. These optimised hyperparameters were then used for model validation (using data held out for model validation) and CFR estimation based on data with simulated missingness (both predictors and outcome). For model validation, outcome imputation performance was characterised using the sensitivity, specificity, percentage correctly classified (PCC) and the area under the receiver operating characteristic curve (AUC). Each model returned a probability for each missing outcome. Converting these probabilities into imputed binary values required the identification of a threshold probability. I selected our threshold to obtain equal sensitivity and specificity [91].

Within this algorithmic framework, I varied the outbreak dataset size, the type and scale of data missingness, and the model training/validation ratio for 100 simulations, each time estimating the model performance and CFR. The true CFR of the simulated dataset was also calculated each time. A step-by-step detailed description of the algorithmic framework is provided in Chapter 3.

In earlier work, assuming survival outcomes were MAR, Forna *et al.* adjusted the imputed CFR to account for imperfect sensitivity and specificity of the BRT inference method [91]. As

a sensitivity analysis, I adjusted the unadjusted CFR estimates to investigate whether adjusting the imputed survival outcomes reduces the bias in the true CFR for all inference methods. The reduction in bias was calculated as follows:

$$\text{Unadjusted CFR bias} = |\text{Unadjusted CFR} - \text{True CFR}|$$

$$\text{Adjusted CFR bias} = |\text{Adjusted CFR} - \text{True CFR}|$$

$$\text{Reduction in bias (percent)} = \frac{\text{Adjusted CFR bias}}{\text{Unadjusted CFR bias}} \times 100, (4)$$

Our main analyses present adjusted CFR for MAR, while adjusted CFR for MCAR and MNAR are presented in the sensitivity analysis.

The *mlr* package; which provides a unified interface to MaLe in R was utilised [157]. For the algorithms which do not allow for missingness in predictors, I imputed the missing predictor values using the mean observed predictor values to allow all algorithms to be applied to the same sized datasets.

6.3 Results

Figure 6.1 shows the inferential performances for all 20 predictors and each inference method. Across inference methods, the inferential performance (mse) for the ten predictors included in the simulation process was higher 1% (median, range: 0%–6.7%) than that for the other predictors added for model fitting 0.3% (median, range: 0%–1.5%).

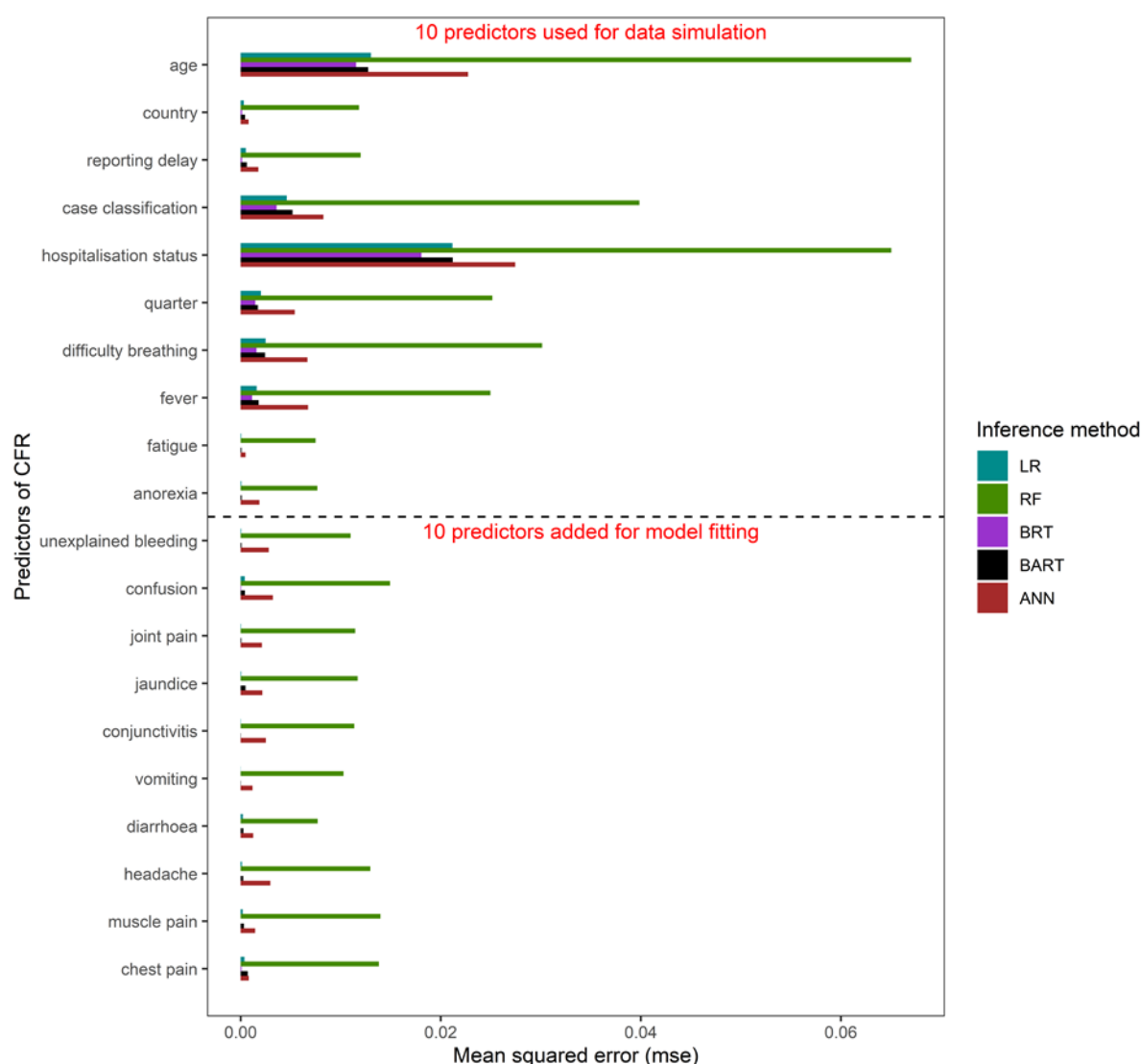


Figure 6. 1: Inferential performances (mean squared error [mse]) for all 20 predictors of case fatality ratio (CFR) estimated for each inference method (i.e. LR, RF, BRT, BART, ANN). The dotted horizontal line separates the ten predictors used in the simulation process from the ten predictors added for model fitting.

Figure 6.2 shows performance results for simulated MCAR missingness in both the survival outcome and all the predictors. Inference methods performed consistently better than chance when selecting a probability threshold such that sensitivity equals specificity. The performance of the models decreased as the proportion of missingness increased. For

instance, across inference methods, dataset size and proportion of training data used, the AUC decreased by 7% (median, range: 1%–16%) when missingness was increased from 10% to 40%. At 40% missingness, the AUC of BRTs was slightly greater, 2% (median, range: -14%–16%), compared to the other inference methods combined. Across inference methods, the proportion of missingness and proportion of training data used, the AUC increased by 1% (median, range: -4%–6%) when outbreak size was increased from 50% to 100%. The AUC increased by 1% (median, range: -5%–4%) when the proportion of data used for model training increased from 50% to 80%.

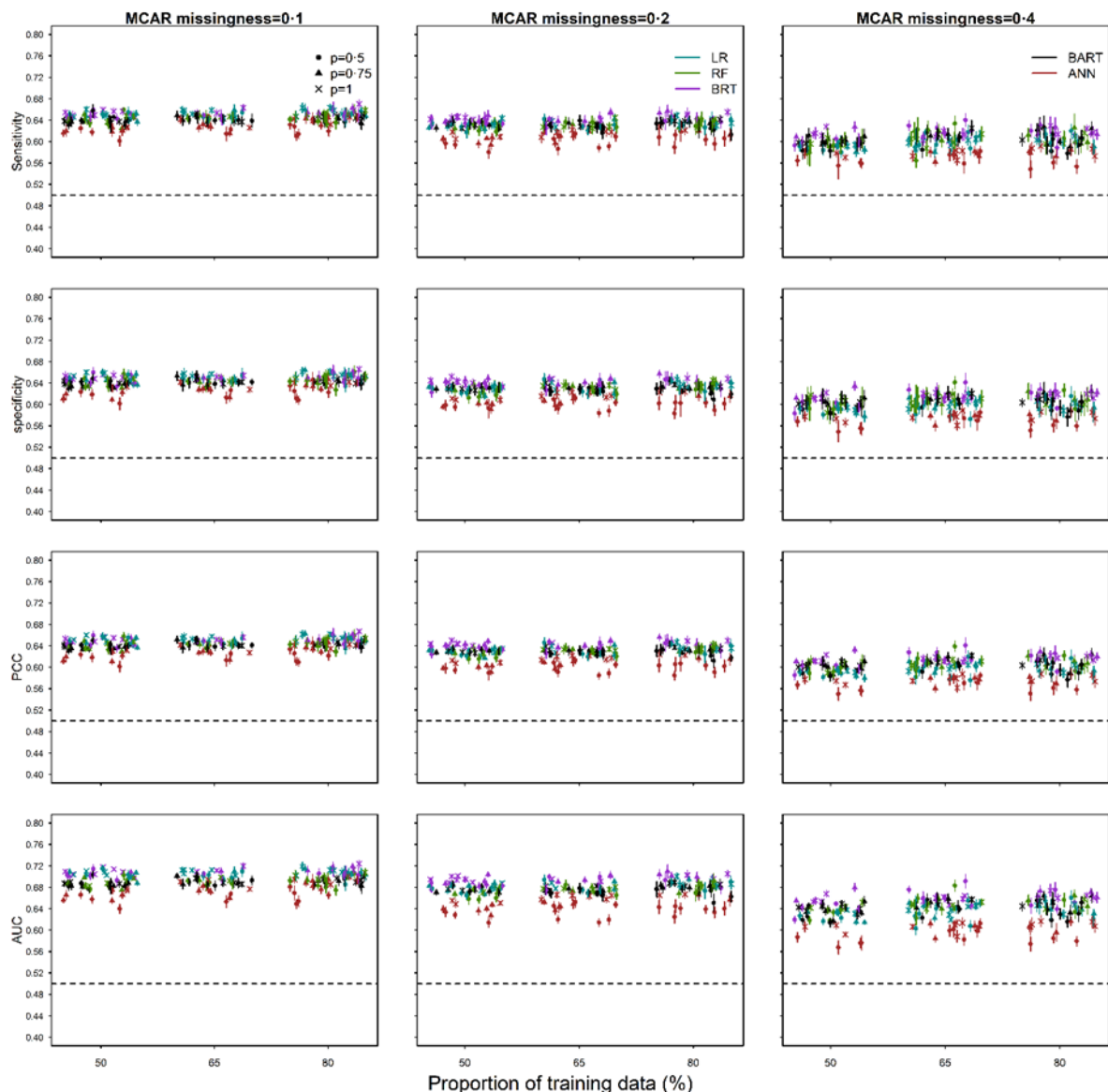


Figure 6. 2: Survival outcomes and predictors missing completely at random (MCAR):

Imputation performance (sensitivity, specificity, percentage correctly classified [PCC], area under the receiver operating characteristic curve [AUC]) as a function of the proportion of data used for model training and a) the proportion of outbreak data (p); where $p=1$ corresponds to the full dataset of 12,049 cases (100%); b) the proportion of simulated missingness of 0.1 (10%), 0.2 (20%) and 0.4 (40%). The imputation performance at which imputations are as good as chance is indicated by the dotted horizontal line at 0.5.

Figure 6.3 shows performance results for MAR missingness in the survival outcome and MCAR missingness in the predictors. The performance profile for this scenario is similar to that in the MCAR simulations.

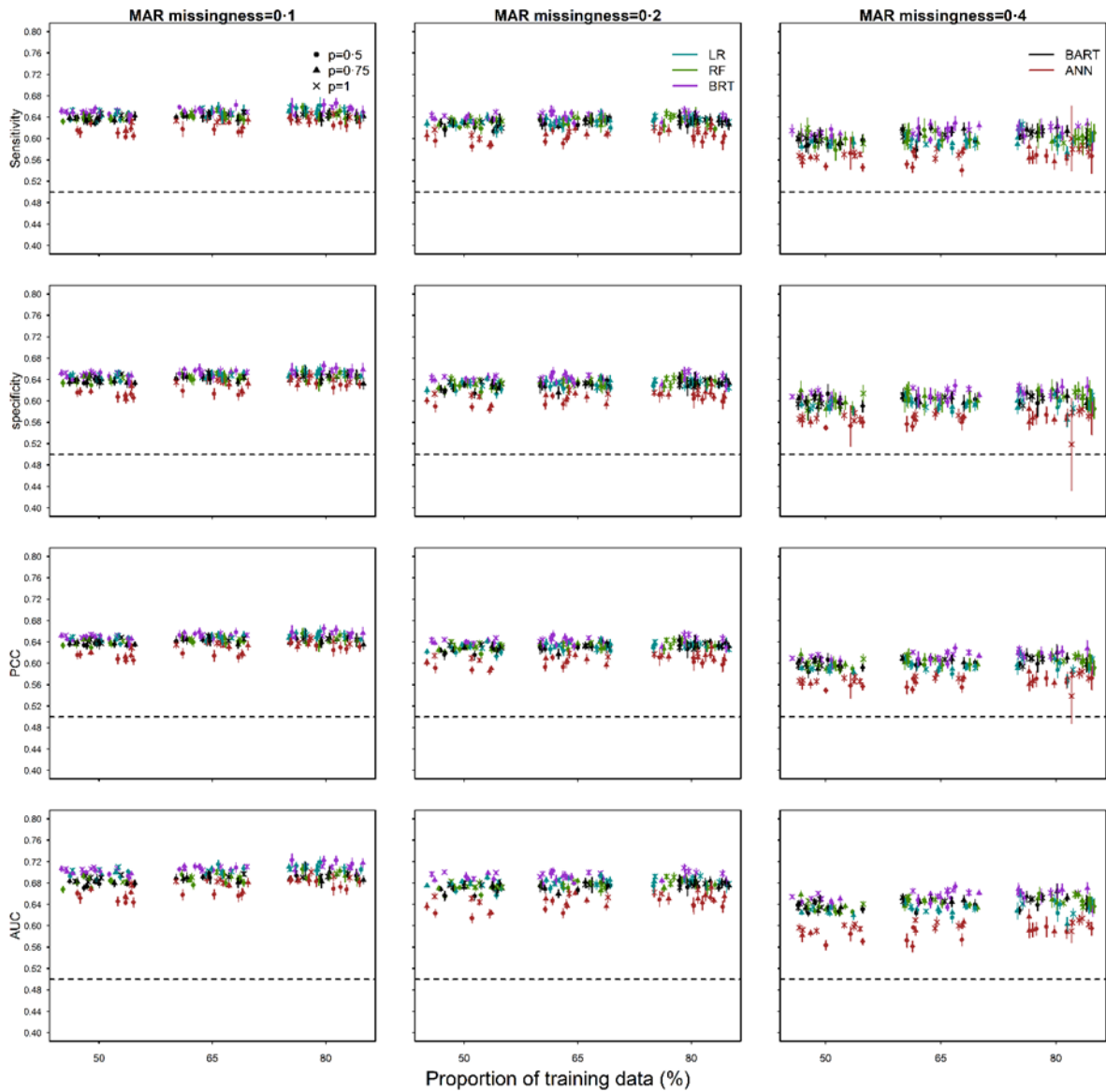


Figure 6. 3: Survival outcome missing at random (MAR) and predictors missing completely at random (MCAR): Imputation performance (sensitivity, specificity, percentage correctly classified [PCC], area under the receiver operating characteristic curve [AUC]) as a function of the proportion of data used for model training and a) the proportion of outbreak data (p); where $p=1$ corresponds to the full dataset of 12,049 cases (100%); b) the proportion of simulated missingness of 0.1 (10%), 0.2 (20%) and 0.4 (40%). The imputation performance at which imputations are as good as chance is indicated by the dotted horizontal line at 0.5.

Figure 6.4 shows performance results for MNAR missingness in the survival outcome and MCAR missingness in the predictors. Below 20% missingness, inference methods performed consistently better than chance (AUC 62% (median, range: 53%–68%)), when selecting a probability threshold such that sensitivity equals specificity. AUC decreased by 12% (median, range: 6%–16%) when the missingness was increased from 10% to 40%. At 40% missingness, sensitivity, specificity and PCC and estimates overlapped for most inference methods, and AUC estimates were only slightly above 50%. Similarly, to MCAR, performance improved with increased outbreak dataset size and on average, performance slightly increased with increases in the percentage of data used for model training.

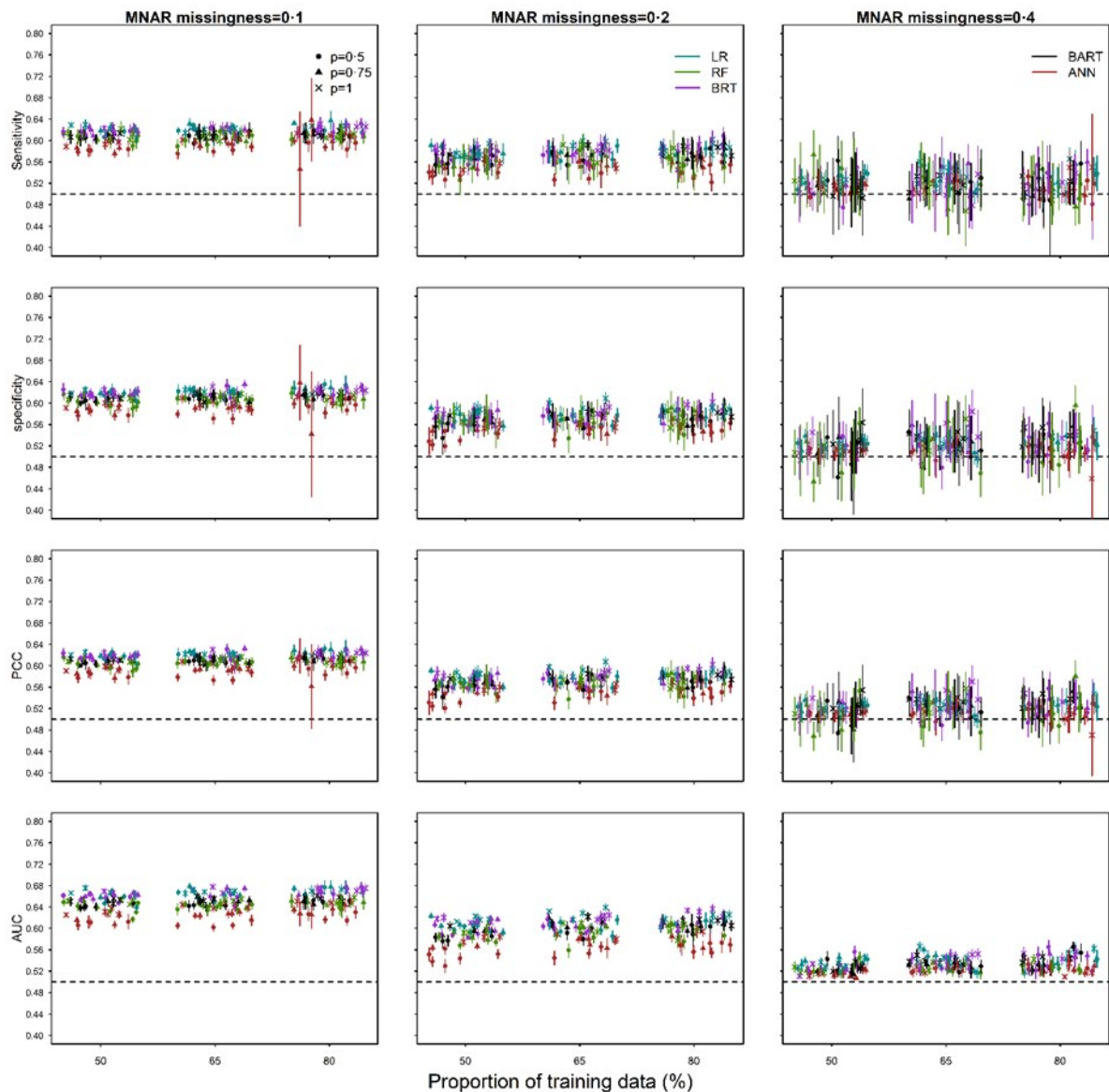


Figure 6. 4: Survival outcome missing not at random (MNAR) and predictors missing completely at random (MCAR): Imputation performance (sensitivity, specificity, percentage correctly classified [PCC], area under the receiver operating characteristic curve [AUC]) as a function of the proportion of data used for model training and a) the proportion of outbreak data (p); where $p=1$ corresponds to the full dataset of 12,049 cases (100%); b) the proportion of simulated missingness of 0.1 (10%), 0.2 (20%) and 0.4 (40%). The imputation performance at which imputations are as good as chance is indicated by the dotted horizontal line at 0.5.

Figure 6.5 shows the unadjusted CFR estimates for MCAR, MAR and MNAR missingness. For all of the inference methods considered in these experiments, the bias in CFR increased with increasing proportions of missingness. For MCAR missingness, LR and ANN tended to overestimate the CFR while RF, BRT and BART tended to underestimate the CFR. For MAR missingness, all models tended to underestimate the CFR, with increasing uncertainty compared to the CFR estimates obtained with MCAR. For MNAR missingness, the bias and uncertainty in the CFR estimates were comparatively higher. There were no distinguishable patterns in CFR estimates due to differences in outbreak dataset size. The uncertainty in CFR estimates usually increases with outbreak size. However, in my simulation this was not observed. The proportion and type of missingness in survival outcomes are therefore driving the uncertainty in CFR. Finally, the CFR estimates appeared to be independent of the training/validation ratio.

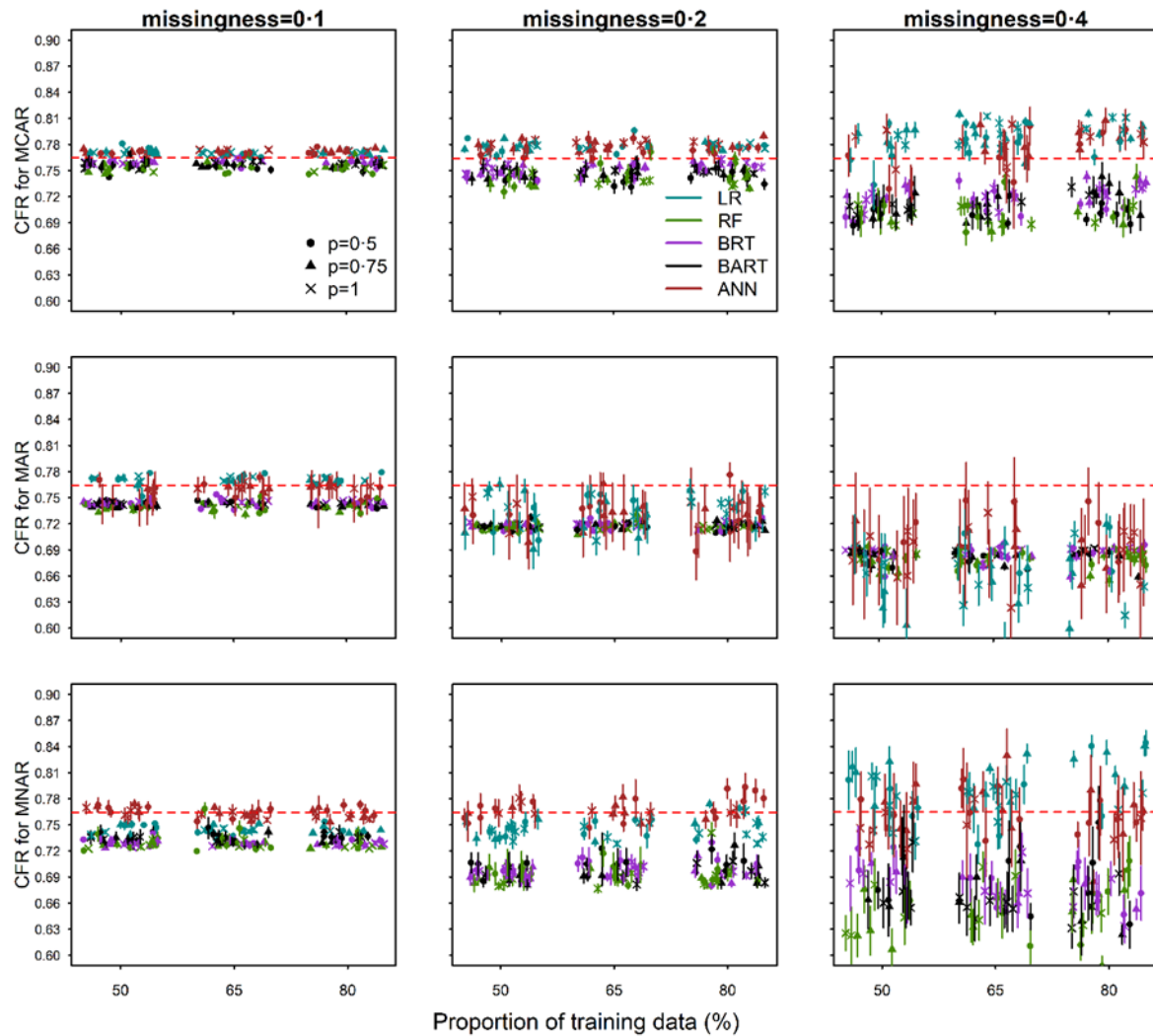


Figure 6. 5: The relationship between the unadjusted CFR estimate for each missingness type (MCAR, MAR, MNAR) and the proportion of data used for model training and a) proportion of outbreak data (p); where $p=1$ corresponds to the full dataset of 12,049 cases (100%); b) proportion of simulated missingness of 0.1 (10%), 0.2 (20%) and 0.4 (40%). The true CFR of the complete simulated data (without simulated missingness) is indicated by the red dotted horizontal line.

Figure 6.6 shows results where the CFRs for MAR and MCAR survival outcomes are adjusted for imperfect sensitivity and specificity of the MaLe models. For MAR, the overall reduction in bias across inference methods, proportion of missingness, outbreak size and proportion

of training data was 0.5% (median, range: 0%–11%). While the overall reduction in bias for MCAR was 2.25% (median, range: 0%–10%). MCAR-adjusted CFRs are upwardly biased when missingness is 20% or 40% but the MAR-adjusted CFRs are relatively less biased for the same proportions of missingness.

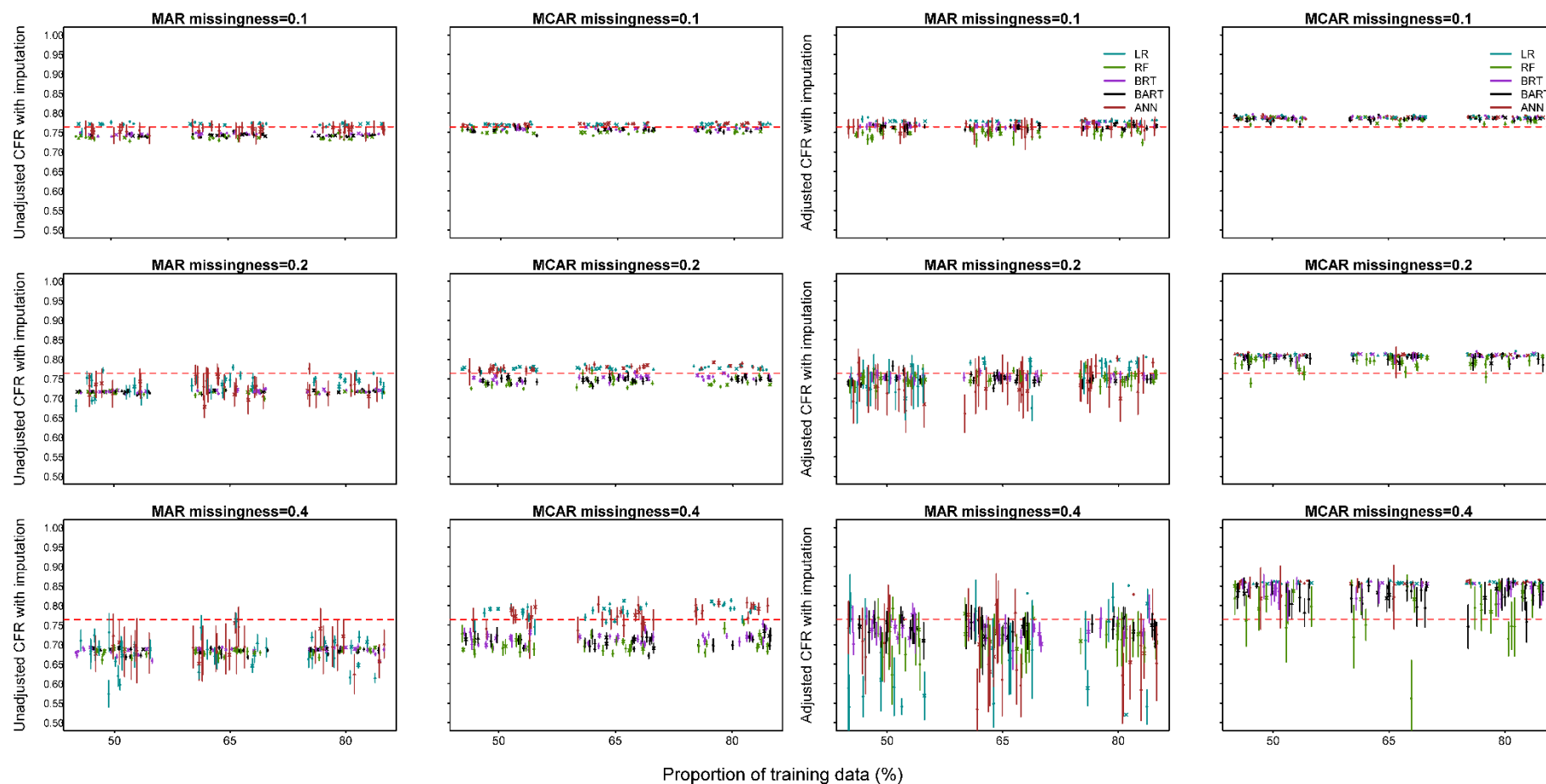


Figure 6. 6: The unadjusted and adjusted CFR estimates for MAR and MCAR missingness in survival outcomes and predictors missing completely at random (MCAR). The true CFR of the complete simulated data (without simulated missingness) is indicated by the red dotted horizontal line.

6.4 Discussion

MaLe algorithms are used in various epidemiological applications [158-160]. An increasingly common application is to obtain and improve estimates of epidemiological parameters of infectious diseases [97]. Our simulation study examined the inferential and predictive performance of five MaLe algorithms (i.e. LR, RF, BRT, BART and ANN) to estimate the CFR for Ebola varying outbreak data characteristics including the outbreak dataset size, the type and magnitude of data missingness, and the model training/validation ratio.

The inferential performance profile (Figure 6.1) shows that the inference methods correctly identified the ten predictors that largely explained the simulated variance in CFR. RF picks up signals from unimportant predictors because it is inherently programmed to explore predictors independent of data during model fitting as a way of increasing model generalisability [87].

Corroborating the existing literature, our simulations demonstrate that MaLe algorithms perform better with large (outbreak) database sizes [161]. Algorithms like ANN usually require even more data to achieve optimal performance and produce unbiased CFR estimates. The suboptimal performance for outbreaks of small sizes could be attributed to less information in small datasets resulting in MaLe classifiers that are less generalisable. In practice, the amount of data is an important factor to consider when deciding whether to use MaLe to impute survival outcomes.

The type and scale of missingness in the survival outcomes substantially influence the performance of MaLe algorithms. The performance profile of the inference methods for MCAR and MAR missingness is consistent with the published literature [162-164]. For example, missing data handling methods like listwise deletion and multiple imputations

have been previously shown to be unbiased in the presence of MCAR and MAR missingness [162]. In addition, consistent with published work [162], I found that as the proportion of missingness increased, none of the MaLe algorithms performed ideally for survival outcome MNAR missingness. In fact, no inferential method performed particularly well when 40% of outcomes are MNAR. Thus, MNAR CFR estimates showed more bias and more uncertainty. In instances of >40% missingness in survival outcomes, at least 50% random follow up of cases could reduce the bias and uncertainty in CFR estimates. For missingness MAR, CFR estimates were below the true CFR. Using a similar approach implemented by Forna *et al.* [91], I demonstrated that, for BRT and BART, adjusting for imperfect sensitivity and specificity reduced bias in the CFR estimates. Our simulations show that for missingness > 10% imputation using MaLe may bias CFR estimates, in line with existing literature suggesting that datasets with limited missingness are not influenced to the same extent by data imputation methods as are datasets with substantial levels of missingness [80]. Based on these results, I encourage epidemiologists to initially investigate the scale and type of missingness (using statistic tests such as Little's MCAR test [92]) before using MaLe algorithms for CFR estimation. Key domain knowledge and outbreak context should drive data analysis and interpretation of results.

I found that model performance increases with increases in model training/validation ratio. These results suggest that provided the outbreak dataset is sufficiently large, training MaLe models on at least half (50%) of the data could optimise performance as techniques like *k*-fold cross validation ensure that models are internally validated to prevent overfitting to training data.

Our results do not investigate the type of missingness in the predictors of survival outcome; I assumed that the missingness in all predictors was MCAR. Thus, these experiments are a simplification of the possible complexities in observed outbreak data. I investigated the proportion of missingness in the predictors and imputed the missing values using the mean before implementing LR and ANN. Imputing for missingness in predictors, at least in part, could explain the relatively reduced performance of these two algorithms. ANN parameters usually require expert optimisation which may explain part of the reduction in performance. RF, BRT and BART algorithms as implemented in our experiments inherently handled the missingness in the predictors [66, 91, 151]. By resampling with replacement from the 2013–2016 West African Ebola outbreak dataset, these simulation results are only contextually relevant to an infinite population with the exact characteristics of that dataset.

This work is not a technical appraisal of MaLe inference methods, but rather it is meant to guide epidemiologists in making more informed choices as they consider the use of the available MaLe tools for CFR estimation and, more broadly, for infectious disease outbreak analysis.

To conclude, these scenarios and results illustrate the potential of MaLe algorithms to describe outbreak patterns, impute survival outcomes and thus improve CFR estimation. I confirm that algorithms like BRT that inherently handle missingness in predictors generally perform better and adjusting for imperfect sensitivity and specificity reduces the bias in the CFR estimates based, in part, on imputed data. Broadly, I demonstrate that the nature and context of the epidemic under investigation should inform the use of MaLe algorithms for infectious disease epidemiological analyses.

6.5 Additional results

6.5.1 Survival outcome, demographic and clinical predictors considered for this simulation study.

Table 6.2 describes the survival outcome and predictors simulated in this study. Except for reporting delay (the difference between the date of reporting and the of symptom onset in days) which is continuous, all other simulated predictors are categorical. Missingness in the survival outcome was simulated as MCAR/MAR/MNAR while missingness in all predictors was simulated as MCAR.

Table 6. 2: Survival outcome, demographic and clinical predictors considered for this simulation study.

Predictor	Description	Mean (SD)	Type of missingness investigated
Outcome			
Survival outcome	Survival outcome in two classes: Alive, Dead	NA	MCAR/MAR/MNAR
Predictors			
Age (categorised)	Age in 16 classes: 0-4, 5-9,10-14, 15-19,20-24, 35-39,40-44,	NA	MCAR

	45-49,50-54, 55-59, 60-64,65-69, 70-74,75+		
Country of origin	Country in three classes: Sierra Leone, Guinea, Liberia	NA	MCAR
Reporting delay	The difference between the date of reporting and the of symptom onset (days)	5.68 (8.91)	MCAR
Case classification	Case classification in three classes: confirmed, probable and suspected	NA	MCAR
Hospitalisation status	Hospitalisation status in two classes: Yes, No	NA	MCAR

Quarter (Date of reporting aggregated at 3-month intervals)	Quarter in seven classes: January-March 2014, April-June 2014, July-September 2014, October-December 2014, January-March 2015, April-June 2015, July-September 2015	NA	MCAR
Difficulty breathing	Difficult breathing in two classes: Yes, No	NA	MCAR
Fever	Fever in two classes: Yes, No	NA	MCAR
Fatigue	Fatigue in two classes: Yes, No	NA	MCAR
Anorexia	Anorexia in two classes: Yes, No	NA	MCAR
Unexplained bleeding	Unexplained bleeding in two classes: Yes, No	NA	MCAR

Confusion	Confusion in two classes: Yes, No	NA	MCAR
Joint pain	Joint pain in two classes: Yes, No	NA	MCAR
Jaundice	Jaundice in two classes: Yes, No	NA	MCAR
Conjunctivitis	Conjunctivitis in two classes: Yes, No	NA	MCAR
Vomiting	Vomiting in two classes: Yes, No	NA	MCAR
Diarrhoea	Diarrhoea in two classes: Yes, No	NA	MCAR
Headache	headache in two classes: Yes, No	NA	MCAR
Muscle pain	Muscle pain in two classes: Yes, No	NA	MCAR
Chest Pain	Chest pain in two classes: Yes, No	NA	MCAR

*SD=standard deviation, * NA = Not applicable

6.6 Sensitivity analysis

As a sensitivity analysis, I present results for the comparison of unadjusted and adjusted CFR for MCAR and MNAR missingness in the survival outcomes.

6.6.1 Comparing the unadjusted and adjusted CFR estimates for MCAR missingness in survival outcomes

Figure 6.7 shows that for MCAR missingness in survival outcome and after sensitivity and specificity adjustment, overall reduction in CFR bias across inference methods, proportion of missingness, outbreak size and proportion of training data was 2.25% (median, range: 0%–10%).

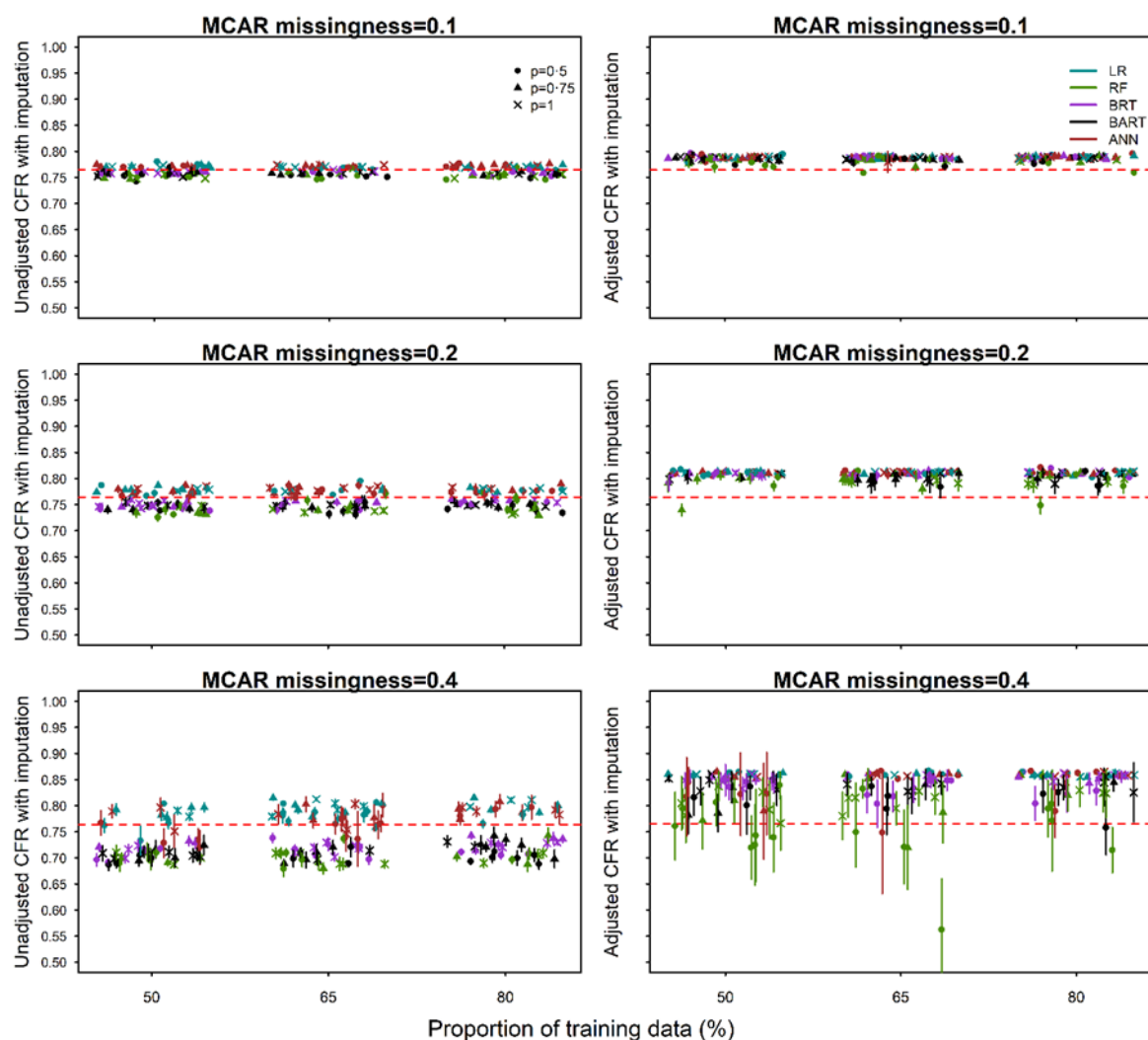


Figure 6. 7: Comparing the unadjusted and adjusted CFR estimates for MCAR missingness in survival outcomes and predictors missing completely at random (MCAR). The true CFR of the complete simulated data (without simulated missingness) is indicated by the red dotted horizontal line.

6.6.2 Comparing the unadjusted and adjusted CFR estimates for MNAR missingness in survival outcomes

Figure 6.8 shows that for MNAR missingness in survival outcome and after sensitivity and specificity adjustment, overall reduction in CFR bias across inference methods, proportion of missingness, outbreak size and proportion of training data was 1.5% (median, range: 0%–16%).

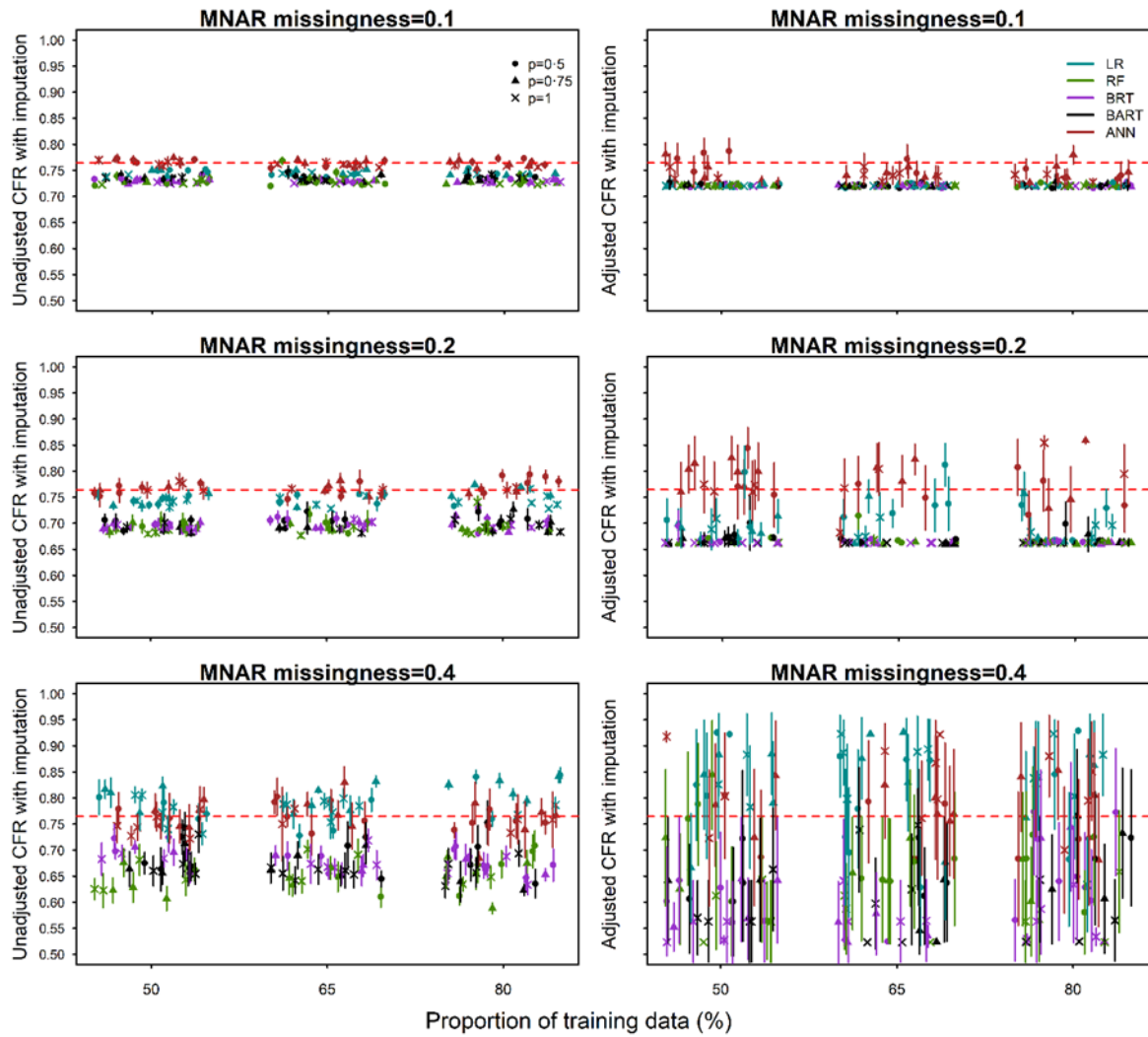


Figure 6. 8: Comparing the unadjusted and adjusted CFR estimates for MNAR missingness in survival outcomes and predictors missing completely at random (MCAR). The true CFR of the complete simulated data (without simulated missingness) is indicated by the red dotted horizontal line.

Chapter 7: Discussion of findings: Subnational Analysis and Modelling of the Ebola Epidemic in West Africa, 2013-2016

The main motivation behind this thesis has been to use machine learning algorithms (i.e. BRT) to improve estimates and understand the determinants of EVD CFR for the 2013–2016 Ebola epidemic in West Africa, accounting for the missingness in survival outcomes in a scientifically robust manner. Additionally, using geospatial techniques, I assessed the patient mix in detail to characterise key sources of the sub-national spatiotemporal heterogeneity in the CFR. Finally, using simulated outbreak data, I investigated the use of BRT and other machine learning inference methods (LR, RF, BART and ANN) to estimate CFR under varying outbreak scenarios. Outbreak scenarios investigated differed in outbreak size, proportions of missingness in outbreak data, the type of missingness in outbreak data and the model training/validation ratio.

7.1 Summary of Findings

In chapter 2, I described the epidemiological characteristics of the 2013-2016 Ebola epidemic dataset used in this thesis. The entire dataset for Sierra Leone, Guinea and Liberia combined consisted of 33,338 confirmed, probable and suspected cases of Ebola. Survival outcomes were recorded for 18,644 (56%) cases. Informed by the existing literature [17], I identified potential predictors of Ebola CFR from the data and quantified the proportion of missingness in each predictor. I then investigated the type of missingness in survival outcomes and established that survival outcomes are not MCAR and that the missingness depended on some of the potential predictors. This initial data characterisation and descriptive background form the basis of the analyses in subsequent chapters of this thesis.

In Chapter 3, relying on previously published work [91, 146], I described in detail, the statistical approaches and MaLe algorithms used throughout this thesis. The model performance measures defined include the sensitivity, specificity, percentage correctly classified and the AUC. The models returned probabilities, and these were converted to binary values using a threshold probability. I describe different approaches to assign threshold probability and throughout this thesis, I selected a threshold of equal sensitivity and specificity which allows for making predictions without trading off death and survival outcomes. To prevent model overfitting, I describe *k*-fold cross validation a computationally efficient technique that internally validates the models to prevent overfitting. I also describe bootstrapping a technique that allowed me to quantify the uncertainty in the performance measures and the CFRs without making any assumptions about the functional distribution of the outbreak data. With the statistical approaches underpinning my algorithmic framework well defined, I ended this chapter describing details of the strategy used to evaluate model performance and estimate CFR in subsequent chapters.

In chapter 4, from which original content has been published [91], using the methodology outlined in chapter 3, I used the large-scale dataset of the 2013-2016 West African Ebola epidemic to predict survival outcomes for Ebola cases, adjust the predictions for model imperfection and re-estimated more representative estimates of CFR. I also identified and explored the relevant clinical and demographic predictors of Ebola CFR. I found that district, hospitalisation status, age, case classification, and quarter explained 93.6% of the variance in the naïve CFR. I concluded that, used in conjunction with other resources, adjusted CFR estimates will inform public health contingency planning for future Ebola epidemics, and help better allocate resources and evaluate the effectiveness of future Ebola outbreak/epidemic interventions.

In chapter 5, for which original content is also published [146], I implemented BRT and geospatial (i.e. semivariogram and kriging) algorithmic framework described in chapter 3 and, using the case fatality predictions from the BRT models and the observed CFR, I estimated spatial and temporal-level residuals and these residuals were investigated for spatial and temporal autocorrelation. BRT modelling accounted for most of the spatiotemporal variation and interactions in CFR, but moderate spatial autocorrelation remained. Combining district-level CFR estimates and kriged district-level residuals provided the best linear unbiased predicted map of CFR accounting for the both explained and unexplained spatial variation. Temporal autocorrelation was not observed in the district-level residuals from the BRT estimates. Chapter 5 provides new insight into the epidemiology of the 2013–2016 West African Ebola epidemic with a view of informing future public health contingency planning, resource allocation and impact assessment. The analytical framework developed, coupled with key domain knowledge, could be deployed in real time to support the response to ongoing and future epidemics/outbreaks.

Finally, in chapter 6, using simulated data, and I investigated the performance of five MaLe algorithms (i.e. LR, RF, BRT, BART and ANN). I estimated CFR for varying outbreak sample sizes, types and proportions of missingness, and model training/validation ratios. The scenarios and results presented in this chapter illustrate the potential of MaLe algorithms for CFR estimation. Adjusting for imperfect sensitivity and specificity of MaLe inference methods does reduce bias in CFR estimates. Algorithms like RF, BRT, and BART that inherently handle missingness in data generally performed better. However, there is no one size fit all approach, rather key domain knowledge (i.e. the knowledge of the field the data belong to – infectious disease epidemiology) and the nature and context of the epidemic

under investigation, should inform the use of MaLe algorithms for CFR estimation and infectious disease epidemiological analyses based on datasets with missing data.

7.2 Future Work and Limitations

Although the overarching aim of this thesis was to impute for unavailable (unreported) data to estimate more representative CFR for the 2013–2016 West African Ebola epidemic, the major limitation of this work has been the quality of the available data itself. In previous work and building on West African Ebola experience, Cori *et al.* identified key outbreak data quality issues and gave practical suggestions for improvement in data quality for future outbreaks [165]. For instance, if the data used in this thesis had included other known clinical predictors like viral load [117, 118], CFR estimates could be further refined, and MaLe models could then be used in subsequent case triaging and prognosis. Using data for the West African epidemic that included viral load, age, temperature, bleeding, jaundice, dyspnoea, dysphagia, and time-to presentation as predictors, Colubri *et al.* used LR to build a prognostic tool for Ebola [166]. If similar predictors are available, the algorithmic framework implemented throughout this thesis could also be adapted as an efficient prognostic tool that can inherently handle the missingness in the predictors. The output from these analyses are themselves useful to inform resource allocation, care planning and provide a baseline CFR against which the CFR observed with interventions (such as a new treatment) could be assessed.

Although all three countries (i.e. Sierra Leone, Guinea and Liberia) collected data throughout the epidemic, surveillance systems and case reporting procedures were not standardised within and across countries making it very difficult reliably to compare CFR estimates. For instance, Liberia switched from an aggregate system of reporting to an individual-based

system of reporting that could have affected data curation at country level [77]. For the recent EVD outbreak in the Democratic Republic of the Congo, case reporting used by six neighbouring countries became increasingly complex and discordant during a 6-month period [167]. South Sudan removed its 'probable' case reporting category by November 2018, Uganda did not include the category in the 6-month period. In most countries, nonbleeding symptoms were included for reporting 'suspected' cases but, in Uganda, symptom criteria were limited to signs of unusual bleeding [167]. While it is epidemiologically important to classify the reported Ebola cases as per WHO guidelines, the different case classification systems used by different countries and consequently in the published literature make the comparison of CFR estimates across different studies difficult. For instance, it would have been interesting to compare the complete district-level CFR maps with district CFR outliers estimated by Garske *et al.* [78]. However, the CFR estimates in Garske *et al.* are for confirmed cases only whilst maps produced in this thesis are for confirmed, probable and suspected cases. The algorithmic framework could be adapted to produce maps for confirmed cases only and these maps could then be compared with those in Garske *et al.* While this type of cross-analysis is possible for studies using similar datasets from the same outbreak, comparing CFR estimates across different countries and/or different outbreaks could be more challenging. Future work could include developing robust statistical approaches to compare CFR across different outbreaks. However, at policy and implementation levels, significant effort is still needed globally to strengthen and harmonise outbreak surveillance systems.

In chapter 5, future work could compare CFR estimates for every point on the linear unbiased map with additional localised epidemiological and demographic parameters.

Because district-level (higher administrative unit) estimates were used for these analyses,

the findings do not explicitly reflect the variation in smaller administrative units such as chiefdoms, districts (the sub-administrative level in Liberia) and sub-prefectures in Sierra Leone, Liberia and Guinea, respectively. With more detailed data on patient residence, it would be interesting to compare these results with those obtained based on the analysis of smaller administrative units. Using kriging based on estimated semivariograms for interpolation of CFR estimates for districts without survival data is not assumption-free. After including known predictors and adjusting the BRT model for imperfect sensitivity and specificity, in interpolating CFR, I assume that there are no known structural reasons for the districts without survival outcomes to be importantly different from districts with observed CFRs.

Finally, whilst the simulation study allows for the validation of the algorithmic framework and the investigation of outbreak data characteristics; simulations simplify the complexities of observed outbreak data. For instance, I did not go further to investigate the type of missingness in the predictors of survival outcome; I assumed that the missingness in all predictors is MCAR. However, I did investigate the proportion of missingness in the predictors and did simple mean imputation before implementing LR and ANN. Imputing for missingness in predictors, at least in part, could explain the relatively lower performance of these two algorithms compared to RF, BRT and BART, which inherently handled the missingness in predictors. In future work, these simplified assumptions could be further relaxed to investigate how the data complexities would affect the performance of MaLe models and the epidemiological parameters (including CFR) they estimate.

The MaLe inference methods investigated in this thesis are not exhaustive. Methods like XGBoost, a variant of tree ensembled models have shown high performances in previous

experiments [168]. Automated machine learning methods like Tree-based Pipeline Optimization Tool (TPOT) that do data cleaning, predictor engineering, model selection, and hyperparameter optimisation in one operation are rapidly being developed and deployed [169]. While these advanced approaches would ultimately make MaLe more accessible to epidemiologists, incorrect inferences could also become ubiquitous without key domain knowledge underpinning the interpretation of results.

7.3 Implications of Research

The research presented here has focused on the use of BRT to impute for unreported survival outcomes and produced improved (i.e. representative of all cases and adjusted for BRT imputation performance) CFR estimates for the 2013–2016 West African Ebola epidemic. To my knowledge, the dataset used in this study (33 338 cases) is the largest case-level data used to characterise the CFR of EVD. Semivariograms and kriging analysis have also been used to investigate spatiotemporal autocorrelation in subnational CFR. Reliable national CFR estimates and subnational CFR maps will be important in strategic planning and resource allocation in different future outbreak/epidemic contexts not just limited to West Africa.

This thesis has also identified and explored relevant clinical and demographic predictors of the CFR. I found that district, hospitalisation status, age, case classification, and quarter (date of reporting aggregated at 3-month intervals) explained 93.6% of the variance in the naïve CFR. Most bleeding predictors (except for unexplained bleeding) were removed from the simplest model that maximised CFR predictions, which corroborates the existing literature suggesting that external bleeding was not a prominent characteristic of the West African Ebola epidemic. Hiccup occurrence was also excluded from the simplest model.

However, from these findings, I cannot tell whether the importance of these clinical predictors in explaining the observed variance in CFR is related to changing clinical presentation of the EVD over time or was caused by changes in reporting practices throughout the West African epidemic.

Although adjusted CFRs with imputation are more representative of the patient mix, the wide left-skewed confidence intervals for confirmed, probable and suspected cases bring into focus the case classification system used in this study. CFR estimates for confirmed cases are more robust suggesting that, the suspected cases which were less validated epidemiologically could be driving the imprecision in CFR estimates. Analysis in Chapter 6 show that adjusting CFR, irrespective of the nature of survival outcome missingness, does not improve precision when the proportion of missingness is greater than 20%. Thus, data collection in outbreak settings should reduce the number of predictors investigated and random-sample follow-ups to ascertain missing outcomes (e.g. at least 50% random follow-up if missingness were 40%).

Since subnational estimates of CFR are spatiotemporally heterogeneous, methods (such as semivariograms and kriging) to resolve heterogeneity in Ebola CFR at a fine spatial resolution will be beneficial for informing future public health contingency planning, resource allocation and impact assessment. These tools can also be used to map and compare subnational CFR estimates for every point on the map with additional localised epidemiological and demographic parameters. Since data collected during outbreaks/epidemics will sometimes present with substantial missing records, MaLe and geospatial analysis (coupled with key domain knowledge) are powerful tools to impute

unrecorded outcomes and re-estimate reliable baseline CFR estimates and indeed other relevant epidemiological parameters from outbreak data.

The simulation study presented in this thesis is not a technical appraisal of MaLe models, but rather meant to help epidemiologists make more informed choices as they explore the use of available MaLe toolbox for CFR estimation and, more broadly, for infectious disease outbreak analysis. For outbreaks of smaller size, epidemiologists should prioritise following up all cases to ascertain survival outcomes before estimating CFRs. While survival outcomes ascertainment may not always be possible for large outbreaks, epidemiologists must be careful not to overfit models with large data and should implement MaLe algorithms such as RF, BRT and BART that are simple to calibrate given the available data. Data collected during outbreaks are complex and usually context specific. For instance, the case classification guidelines used in West Africa are different from those been used for the recent Ebola outbreak in the Democratic Republic of the Congo [17, 167]. Epidemiologists are encouraged initially to investigate context specific data characteristics like the proportion and type of missingness before any attempt to use MaLe algorithms for CFR estimation. In addition, broader statistical techniques like adjusting for model sensitivity and specificity, K -fold cross validation and bootstrapping, described and implemented in this thesis, are essential for internal validation of MaLe models, quantifying the uncertainty in MaLe model performance and ultimately estimating unbiased epidemiological parameters.

Epidemiologists should be aware of these techniques and should be ready to implement them to improve the performance and report the uncertainty of MaLe algorithms used for CFR estimation. Generally, rather than just the availability of MaLe tools, key domain knowledge, outbreak context and broader statistical considerations should drive analysis and interpretation of results.

The MaLe algorithmic framework presented in this thesis to re-estimate more representative Ebola CFR and generate unbiased subnational CFR maps from 2013–2016 West Africa Ebola epidemic data will also apply to other Ebola outbreaks. Furthermore, the algorithmic framework can be adapted to other infectious diseases (e.g. coronaviruses, dengue and zika viruses) epidemiological analyses based on datasets with missing data.

7.4 Conclusions

Adjusted CFR estimates improved the naïve CFR estimates obtained without imputation and were more representative of the patient mix. Used in conjunction with other resources, adjusted CFR estimates and unbiased CFR maps would inform future public health contingency planning, resource allocation and impact assessment. The algorithmic framework developed in this thesis, coupled with key domain knowledge could be deployed in real time to support the response to ongoing and future EVD outbreaks/epidemics. While BRT was a robust approach for CFR imputation, I demonstrate that there are no lock-and-key approaches when selecting algorithms for CFR imputation, rather the nature and context of the epidemic under investigation should inform the MaLe algorithm used for infectious disease epidemiological analysis. Robust national and subnational estimates of baseline Ebola CFR are essential for the assessment and evaluation of any control measures that may be implemented during an outbreak/epidemic and, thus, the work presented here is a significant contribution to such endeavours.

8 References

1. Kuhn JH, Becker S, Ebihara H, et al. Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. *Archives of virology* **2010**; 155:2083-103.
2. Languon S, Quaye O. Filovirus disease outbreaks: a chronological overview. *Virology: research and treatment* **2019**; 10:1178122X19849927.
3. Van Kerkhove MD, Bento AI, Mills HL, Ferguson NM, Donnelly CA. A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. *Scientific data* **2015**; 2:150019.
4. Breman JG, Heymann DL, Lloyd G, et al. Discovery and description of Ebola Zaire virus in 1976 and relevance to the West African epidemic during 2013–2016. *The Journal of infectious diseases* **2016**; 214:S93-S101.
5. Nicastrì E, Kobinger G, Vairo F, et al. Ebola Virus Disease: Epidemiology, Clinical Features, Management, and Prevention. *Infectious Disease Clinics* **2019**; 33:953-76.
6. Lefebvre A, Fiet C, Belpois-Duchamp C, Tiv M, Astruc K, Glélé LA. Case fatality rates of Ebola virus diseases: a meta-analysis of World Health Organisation data. *Médecine et maladies infectieuses* **2014**; 44:412-6.
7. Leroy EM, Kumulungui B, Pourrut X, et al. Fruit bats as reservoirs of Ebola virus. *Nature* **2005**; 438:575-6.
8. BAIZE S, LEROY EM, GEORGES AJ, et al. Inflammatory responses in Ebola virus-infected patients. *Clin Exp Immunol* **2002**; 128:163-8.
9. Rodriguez LL, De Roo A, Guimard Y, et al. Persistence and Genetic Stability of Ebola Virus during the Outbreak in Kikwit, Democratic Republic of the Congo, 1995. *The Journal of Infectious Diseases* **1999**; 179:S170-S6.

10. Deen GF, Broutet N, Xu W, et al. Ebola RNA Persistence in Semen of Ebola Virus Disease Survivors - Final Report. *N Engl J Med* **2017**; 377:1428-37.
11. Maruyama T, Rodriguez LL, Jahrling PB, et al. Ebola Virus Can Be Effectively Neutralized by Antibody Produced in Natural Human Infection. *Journal of Virology* **1999**; 73:6024-30.
12. MacIntyre CR, Chughtai AA. Recurrence and reinfection—a new paradigm for the management of Ebola virus disease. *International Journal of Infectious Diseases* **2016**; 43:58-61.
13. Mina MJ, Kula T, Leng Y, et al. Measles virus infection diminishes preexisting antibodies that offer protection from other pathogens. *Science* **2019**; 366:599-606.
14. Jacobs M, Rodger A, Bell DJ, et al. Late Ebola virus relapse causing meningoencephalitis: a case report. *The Lancet* **2016**; 388:498-503.
15. Osterholm MT, Moore KA, Kelley NS, et al. Transmission of Ebola viruses: What we know and what we do not know. *mBio* **2015**.
16. Feldmann H. Ebola — A Growing Threat? *New England Journal of Medicine* **2014**; 371:1375-8.
17. WHO Ebola Response Team. Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *N Engl J Med* **2014**; 2014:1481-95.
18. Chertow DS, Kleine C, Edwards JK, Scaini R, Giuliani R, Sprecher A. Ebola virus disease in West Africa—clinical manifestations and management. *N Engl J Med* **2014**; 371:2054-7.
19. Barry A, Ahuka-Mundeke S, Ahmed YA, et al. Outbreak of Ebola virus disease in the Democratic Republic of the Congo, April–May, 2018: an epidemiological study. *The Lancet* **2018**; 392:213-21.
20. Chowell G, Nishiura H. Transmission dynamics and control of Ebola virus disease (EVD): a review. *BMC Medicine* **2014**; 12:196.

21. WHO Ebola Response Team. West African Ebola epidemic after one year—slowing but not yet under control. *N Engl J Med* **2015**; 372:584-7.
22. Nouvellet P, Garske T, Mills HL, et al. The role of rapid diagnostics in managing Ebola epidemics. *Nature* **2015**; 528:S109.
23. Stamm LV. Ebola Virus Disease: Rapid Diagnosis and Timely Case Reporting are Critical to the Early Response for Outbreak Control. *The American Journal of Tropical Medicine and Hygiene* **2015**; 93:438-40.
24. Dhillon RS, Srikrishna D, Kelly JD. Deploying RDTs in the DRC Ebola outbreak. *The Lancet* **2018**; 391:2499-500.
25. Celentano DD, Mhs S. *Gordis epidemiology*. Elsevier, **2018**.
26. Lipsitch M, Donnelly CA, Fraser C, et al. Potential biases in estimating absolute and relative case-fatality risks during outbreaks. *PLoS neglected tropical diseases* **2015**; 9:e0003846.
27. Grennan D. What Is a Pandemic? *JAMA* **2019**; 321:910-.
28. Vaz RG, Mkanda P, Banda R, et al. The Role of the Polio Program Infrastructure in Response to Ebola Virus Disease Outbreak in Nigeria 2014. *The Journal of Infectious Diseases* **2016**; 213:S140-S6.
29. Johnson KM, Breman JG, Burke J, et al. Ebola haemorrhagic fever in Zaire, 1976. *Bulletin of the World Health Organisation* **1978**; 56:271-93.
30. World Health Organisation. Ebola Situation Report. 15 April **2015**:1-16.
31. World Bank. The Economic Impact of the 2014 Ebola Epidemic: Short and Medium Term Estimates for Guinea, Liberia, and Sierra Leone, **2014**:1-29.

32. Gostin LO, Friedman EA. A retrospective and prospective analysis of the west African Ebola virus disease epidemic: robust national health systems at the foundation and an empowered WHO at the apex. *The Lancet* **2015**; 385:1902-9.
33. Moon S, Sridhar D, Pate MA, et al. Will Ebola change the game? Ten essential reforms before the next pandemic. The report of the Harvard-LSHTM Independent Panel on the Global Response to Ebola. *The Lancet* **2015**; 386:2204-21.
34. Nielsen CF, Kidd S, Sillah AR, Davis E, Mermin J, Kilmarx PH. Improving burial practices and cemetery management during an Ebola virus disease epidemic—Sierra Leone, 2014. *MMWR Morbidity and mortality weekly report* **2015**; 64:20.
35. Green A. WHO and partners launch Ebola response plan. *The Lancet* **2014**; 384:481-.
36. Gostin LO, Lucey D, Phelan A. The Ebola Epidemic: A Global Health Emergency. *JAMA* **2014**; 312:1095-6.
37. Gulland A. UK's plan to build community care centres for Ebola patients is questioned. *BMJ (Clinical research ed)* **2014**; 349:g6788-g.
38. Camacho A, Kucharski A, Aki-Sawyer Y, et al. Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements : a Real - time Modelling Study. *PLOS Currents Outbreaks* **2015**; 7:1-28.
39. Farrar JJ, Piot P. The Ebola Emergency - Immediate Action, Ongoing Strategy. *The New England journal of medicine* **2014**:1545-6.
40. Langreth R, Chen C, Nash J, Lauerman J. Ebola drug made from tobacco plant saves US aid workers. *Business Week* **2014**; 8:2014.
41. Dunning J, Sahr F, Rojek A, et al. Experimental Treatment of Ebola Virus Disease with TKM-130803: A Single-Arm Phase 2 Clinical Trial. *PLOS Medicine* **2016**; 13:e1001997.

42. The PREVAIL II Writing Group, for the Multi-National PREVAIL II Study Team. A randomized, controlled trial of ZMapp for Ebola virus infection. *The New England journal of medicine* **2016**; 375:1448.
43. Enserink M. Debate erupts on 'repurposed' drugs for Ebola. *Science* **2014**; 345:718-9.
44. Kupferschmidt K. Successful Ebola treatments promise to tame outbreak. *Science* **2019**; 365:628-9.
45. de Wit E, Feldmann H, Munster VJ. Tackling Ebola: new insights into prophylactic and therapeutic intervention strategies. *Genome Medicine* **2011**; 3:5.
46. Henao-Restrepo AM, Longini IM, Egger M, et al. Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: interim results from the Guinea ring vaccination cluster-randomised trial. *The Lancet* **2015**; 386:857-66.
47. Henao-Restrepo AM, Camacho A, Longini IM, et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). *The Lancet* **2017**; 389:505-18.
48. Shukarev G, Callendret B, Luhn K, Douoguih M. A two-dose heterologous prime-boost vaccine regimen eliciting sustained immune responses to Ebola Zaire could support a preventive strategy for future outbreaks. *Human Vaccines & Immunotherapeutics* **2017**; 13:266-70.
49. Mutua G, Anzala O, Luhn K, et al. Safety and immunogenicity of a 2-dose heterologous vaccine regimen with Ad26. ZEBOV and MVA-BN-Filo Ebola vaccines: 12-month data from a phase 1 randomized clinical trial in Nairobi, Kenya. *The Journal of infectious diseases* **2019**; 220:57-67.

50. Wells CR, Pandey A, Parpia AS, et al. Ebola vaccination in the Democratic Republic of the Congo. *Proceedings of the National Academy of Sciences* **2019**; 116:10178-83.
51. Bell BP. Overview, control strategies, and lessons learned in the CDC response to the 2014–2016 Ebola epidemic. *MMWR supplements* **2016**; 65.
52. Shultz JM, Espinel Z, Espinola M, Rechkemmer A. Distinguishing epidemiological features of the 2013–2016 West Africa Ebola virus disease outbreak. *Disaster Health* **2016**; 3:78-88.
53. CAMERON D, JONES IG. John Snow, the Broad Street Pump and Modern Epidemiology. *International Journal of Epidemiology* **1983**; 12:393-6.
54. Bengtsson L, Gaudart J, Lu X, et al. Using Mobile Phone Data to Predict the Spatial Spread of Cholera. *Scientific Reports* **2015**; 5:8923.
55. Brooker S, Kabatereine N, Tukahebwa E, Kazibwe F. Spatial analysis of the distribution of intestinal nematode infections in Uganda. *Epidemiol Infect* **2004**; 132:1065-71.
56. Wong ZSY, Zhou J, Zhang Q. Artificial Intelligence for infectious disease Big Data Analytics. *Infection, Disease & Health* **2019**; 24:44-8.
57. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol* **2019**.
58. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol* **2008**; 77:802-13.
59. Wüthrich MV. Machine learning in individual claims reserving. *Scandinavian Actuarial Journal* **2018**; 2018:465-80.
60. Mullainathan S, Spiess J. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* **2017**; 31:87-106.

61. Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P. Variation in demersal fish species richness in the oceans surrounding New Zealand: An analysis using boosted regression trees. *Marine Ecology Progress Series* **2006**; 321:267-81.
62. Bibault J-E, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: State of the art and future prospects. *Cancer Letters* **2016**; 382:110-7.
63. Szymczak S, Biernacka JM, Cordell HJ, et al. Machine learning in genome-wide association studies. *Genetic epidemiology* **2009**; 33:S51-S7.
64. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases* **2018**; 66:149-53.
65. Bhatt S, et al. The global distribution and burden of dengue. *Nature* **2013**; 496:504-7.
66. Gu W, Vieira A, Hoekstra R, Griffin P, Cole D. Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. *Epidemiol Infect* **2015**; 143:2786-94.
67. Koch T. Mapping medical disasters: Ebola makes old lessons, new. *Disaster Med Public Health Prep* **2015**; 9:66-73.
68. Raso G, Matthys B, N'Goran EK, Tanner M, Vounatsou P, Utzinger J. Spatial risk prediction and mapping of Schistosoma mansoni infections among schoolchildren living in western Côte d'Ivoire. *Parasitology* **2005**; 131:97-108.
69. Srividya A, Michael E, Palaniyandi M, Pani S, Das P. A geostatistical analysis of the geographic distribution of lymphatic filariasis prevalence in southern India. *The American journal of tropical medicine and hygiene* **2002**; 67:480-9.
70. Nunes C. Tuberculosis incidence in Portugal: spatiotemporal clustering. *International Journal of Health Geographics* **2007**; 6:30.

71. Gleason BL, Foster S, Wilt GE, et al. Geospatial analysis of household spread of Ebola virus in a quarantined village – Sierra Leone, 2014. *Epidemiology and Infection* **2017**; 145:2921-9.
72. Wong R, Harris JK. Geospatial Distribution of Local Health Department Tweets and Online Searches About Ebola During the 2014 Ebola Outbreak. *Disaster Medicine and Public Health Preparedness* **2018**; 12:287-90.
73. Kramer AM, Pulliam JT, Alexander LW, Park AW, Rohani P, Drake JM. Spatial spread of the West Africa Ebola epidemic. *Royal Society open science* **2016**; 3:160294.
74. Fang L-Q, Yang Y, Jiang J-F, et al. Transmission dynamics of Ebola virus disease and intervention effectiveness in Sierra Leone. *Proceedings of the National Academy of Sciences* **2016**; 113:4488-93.
75. Timothy JWS, Hall Y, Akoi-Boré J, et al. Early transmission and case fatality of Ebola virus at the index site of the 2013–16 west African Ebola outbreak: a cross-sectional seroprevalence survey. *The Lancet Infectious Diseases* **2019**; 19:429-38.
76. Wilkinson A, Fairhead J. Comparison of social resistance to Ebola response in Sierra Leone and Guinea suggests explanations lie in political configurations not culture. *Critical Public Health* **2017**; 27:14-27.
77. McNamara LA, Schafer IJ, Nolen LD, et al. Ebola Surveillance — Guinea, Liberia, and Sierra Leone. *MMWR Supplements* **2016**; 65:35-43.
78. Garske T, Cori A, Ariyaratnam A, et al. Heterogeneities in the case fatality ratio in the West African Ebola outbreak 2013–2016. *Phil Trans R Soc B* **2017**; 372:20160308.
79. He Y. Missing data analysis using multiple imputation: getting to the heart of the matter. *Circ Cardiovasc Qual Outcomes* **2010**; 3:98-105.

80. Saunders JA, Morrow-Howell N, Spitznagel E, Doré P, Proctor EK, Pescarino R. Imputing missing data: A comparison of methods for social work researchers. *Social work research* **2006**; 30:19-31.
81. Ridgeway G, Southworth MH, RUnit S. Package 'gbm'. Viitattu **2013**; 10:40.
82. Freeman ME. Package ' PresenceAbsence '. **2015**.
83. Manel S, Williams HC, Ormerod SJ. Evaluating presence absence models in ecology; the need to count for prevalence. *Journal of Applied ecology* **2001**; 38:921-31.
84. Lobo JM, Jiménez-valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* **2008**; 17:145-51.
85. Altman DG, Bland JM. Diagnostic Tests 1: Sensitivity and Specificity. *BMJ: British Medical Journal* **1994**; 308:1552-3.
86. McPherson JM, Jetz W, Rogers DJ. Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions - Possibilities and limitations. *Ecological Modelling* **2006**; 192:499-522.
87. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol. 112. Springer, **2013**.
88. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence* **1995**; 14:1137-43.
89. Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, Giacobini M. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData mining* **2011**; 4:12.
90. Sergeant E. Package ' RSurveillance '. **2016**.

91. Forna A, Nouvellet P, Dorigatti I, Donnelly CA. Case fatality ratio estimates for the 2013 – 2016 West African Ebola epidemic: application of Boosted Regression Trees for imputation. *Clin Infect Dis* **2019**; ciz678, doi:<http://dx.doi.org/10.1093/cid/ciz678>.
92. Li C. Little's test of missing completely at random. *The Stata Journal* **2013**; 13:795-809.
93. Haaskjold YL, Bolkan HA, Krogh KØ, et al. Clinical features of and risk factors for fatal Ebola virus disease, Moyamba District, Sierra Leone, December 2014–February 2015. *Emerging infectious diseases* **2016**; 22:1537.
94. Focosi D, Maggi F. Estimates of Ebola virus case-fatality ratio in the 2014 West African outbreak. *Clinical Infectious Diseases* **2015**; 60:829-.
95. Faye O, Andronico A, Faye O, et al. Use of viremia to evaluate the baseline case fatality ratio of Ebola virus disease and inform treatment studies: a retrospective cohort study. *PLoS Med* **2015**; 12:e1001908.
96. WHO Ebola Response Team. Ebola virus disease among children in West Africa. *New England Journal of Medicine* **2015**; 372:1274-7.
97. Dorigatti I, Donnelly CA, Laydon DJ, et al. Refined efficacy estimates of the Sanofi Pasteur dengue vaccine CYD-TDV using machine learning. *Nature Communications* **2018**; 9:3644.
98. World Health Organisation. Case definition recommendations for Ebola or Marburg virus diseases. Interim Guideline **2014**:1-3.
99. Barry M, Touré A, Traoré FA, et al. Clinical predictors of mortality in patients with Ebola virus disease. *Clinical infectious diseases* **2015**; 60:1821-4.
100. Pigott DM, Bhatt S, Golding N, et al. Global distribution maps of the leishmaniasis. *Elife* **2014**; 3:e02851.
101. Liu C, Berry PM, Dawson TP, Pearson RG. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* **2005**; 28:385-93.

102. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**;1189-232.
103. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *American journal of epidemiology* **1978**; 107:71-6.
104. Reiczigel J, Földi J, Ózsvári L. Exact confidence limits for prevalence of a disease with an imperfect diagnostic test. *Epidemiology & Infection* **2010**; 138:1674-8.
105. Xu Z, Jin B, Teng G, et al. Epidemiologic characteristics, clinical manifestations, and risk factors of 139 patients with Ebola virus disease in western Sierra Leone. *Am J Infect Control* **2016**; 44:1285-90.
106. Dallatomasina S, Crestani R, Sylvester Squire J, et al. Ebola outbreak in rural West Africa: epidemiology, clinical features and outcomes. *Tropical Medicine & International Health* **2015**; 20:448-54.
107. Bah EI, Lamah M-C, Fletcher T, et al. Clinical presentation of patients with Ebola virus disease in Conakry, Guinea. *New England Journal of Medicine* **2015**; 372:40-7.
108. Qin E, Bi J, Zhao M, et al. Clinical features of patients with Ebola virus disease in Sierra Leone. *Clinical infectious diseases* **2015**; 61:491-5.
109. Francica JR, Varela-Rohena A, Medvec A, Plesa G, Riley JL, Bates P. Steric shielding of surface epitopes and impaired immune recognition induced by the ebola virus glycoprotein. *PLoS pathogens* **2010**; 6:e1001098.
110. Bwaka MA, Bonnet M-J, Calain P, et al. Ebola hemorrhagic fever in Kikwit, Democratic Republic of the Congo: clinical observations in 103 patients. *Journal of Infectious Diseases* **1999**; 179:S1-S7.

111. Whitty CJ. The contribution of biological, mathematical, clinical, engineering and social sciences to combatting the West African Ebola epidemic. *Phil Trans R Soc B* **2017**; 372:20160293.
112. Funk S, Ciglenecki I, Tiffany A, et al. The impact of control strategies and behavioural changes on the elimination of Ebola from Lofa County, Liberia. *Phil Trans R Soc B* **2017**; 372:20160302.
113. Qureshi AI, Chughtai M, Bah EI, et al. High survival rates and associated factors among Ebola virus disease patients hospitalized at Donka National Hospital, Conakry, Guinea. *Journal of vascular and interventional neurology* **2015**; 8:S4.
114. Ghani A, Donnelly C, Cox D, et al. Methods for estimating the case fatality ratio for a novel, emerging infectious disease. *American journal of epidemiology* **2005**; 162:479-86.
115. Pelat C, Ferguson NM, White PJ, et al. Optimizing the precision of case fatality ratio estimates under the surveillance pyramid approach. *American journal of epidemiology* **2014**; 180:1036-46.
116. Langer M, Portella G, Finazzi S, et al. Intensive care support and clinical outcomes of patients with Ebola virus disease (EVD) in West Africa. *Intensive care medicine* **2018**; 44:1266-75.
117. de La Vega M-A, Caleo G, Audet J, et al. Ebola viral load at diagnosis associates with patient outcome and outbreak evolution. *The Journal of clinical investigation* **2015**; 125:4421.
118. Fitzpatrick G, Vogt F, Gbabei M, et al. The contribution of Ebola viral load at admission and other patient characteristics to mortality in a Medecins Sans Frontieres Ebola case management centre, Kailahun, Sierra Leone, June–October 2014. *The Journal of infectious diseases* **2015**; 212:1752-8.

119. Michel G, Bisegger C, Fuhr DC, Abel T. Age and gender differences in health-related quality of life of children and adolescents in Europe: a multilevel analysis. *Quality of life research* **2009**; 18:1147.
120. WHO Ebola Response Team. Ebola virus disease among male and female persons in West Africa. *New England Journal of Medicine* **2016**; 374:96-8.
121. Huber C, Finelli L, Stevens W. The economic and social burden of the 2014 Ebola outbreak in West Africa. *The Journal of infectious diseases* **2018**; 218:S698-S704.
122. Mbala-Kingebeni P, Villabona-Arenas C-J, Vidal N, et al. Rapid confirmation of the Zaire Ebola virus in the outbreak of the Equateur province in the Democratic Republic of Congo: implications for public health interventions. *Clinical Infectious Diseases* **2018**; 68:330-3.
123. Lefebvre A, Fiet C, Belpois-Duchamp C, Tiv M, Astruc K, Aho Glélé LS. Case fatality rates of Ebola virus diseases: A meta-analysis of World Health Organisation data. *Medecine et Maladies Infectieuses* **2014**; 44:412-6.
124. Carroll MW, Matthews DA, Hiscox JA, et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature* **2015**; 524.
125. GADM. GADM database of global administrative areas. Global Administrative Areas. **2020**. <https://gadm.org/>.
126. Lai P-C, So F-M, Chan K-W. Spatial epidemiological approaches in disease mapping and analysis. CRC press, **2008**.
127. Anselin L. Local Indicators of Spatial Association—LISA. *Geographical Analysis* **1995**; 27:93-115.
128. Bivand RS, Wong DW. Comparing implementations of global and local indicators of spatial association. *Test* **2018**; 27:716-48.

129. Deng M, Liu Q, Cheng T, Shi Y. An adaptive spatial clustering algorithm based on Delaunay triangulation. *Computers, Environment and Urban Systems* **2011**; 35:320-32.
130. Srividya A, Michael E, Palaniyandi M, Pani SP, Das PK. A geostatistical analysis of the geographic distribution of lymphatic filariasis prevalence in southern India. *The American Journal of Tropical Medicine and Hygiene* **2002**; 67:480-9.
131. Lau CL, Won KY, Becker L, et al. Seroprevalence and spatial epidemiology of lymphatic filariasis in American Samoa after successful mass drug administration. *PLoS neglected tropical diseases* **2014**; 8:e3297.
132. Yue Y, Sun J, Liu X, et al. Spatial analysis of dengue fever and exploration of its environmental and socio-economic risk factors using ordinary least squares: A case study in five districts of Guangzhou City, China, 2014. *International Journal of Infectious Diseases* **2018**; 75:39-48.
133. Jian X, Olea RA, Yu Y-S. Semivariogram modeling by weighted least squares. *Computers & Geosciences* **1996**; 22:387-97.
134. Cressie NA. Spatial prediction and kriging. *Statistics for Spatial Data* (Cressie NAC, ed) New York: John Wiley & Sons **1993**:105-209.
135. Rossoni FD, Felix VB. geotoolsR: Tools to Improve the Use of Geostatistic. **2017**.
136. Carrat F, Valleron A-J. Epidemiologic mapping using the “kriging” method: application to an influenza-like epidemic in France. *Am J Epidemiol* **1992**; 135:1293-300.
137. Davis MW, Grivet C. Kriging in a global neighborhood. *Journal of the International Association for Mathematical Geology* **1984**; 16:249-65.
138. Durbin J, Watson GS. Testing for serial correlation in least squares regression. III. *Biometrika* **1971**; 58:1-19.

139. Miaou S-P, Song JJ, Mallick BK. Roadway traffic crash mapping: a space-time modeling approach. *Journal of Transportation and Statistics* **2003**; 6:33-58.
140. Abdulhafedh A. How to detect and remove temporal autocorrelation in vehicular crash data. *Journal of transportation technologies* **2017**; 7:133.
141. Kleinschmidt I, Sharp B, Clarke G, Curtis B, Fraser C. Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in KwaZulu Natal, South Africa. *American journal of epidemiology* **2001**; 153:1213-21.
142. Weller ZD. spTest: An R Package Implementing Nonparametric Tests of Isotropy. *Journal of Statistical Software* **2018**; 83:1-24.
143. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology* **2019**; 188:2222-39.
144. Gilbert PB, Luedtke AR. Statistical Learning Methods to Determine Immune Correlates of Herpes Zoster in Vaccine Efficacy Trials. *The Journal of Infectious Diseases* **2018**; 218:S99-S101.
145. Midani FS, Weil AA, Chowdhury F, et al. Human Gut Microbiota Predicts Susceptibility to *Vibrio cholerae* Infection. *The Journal of Infectious Diseases* **2018**; 218:645-53.
146. Forna A, Dorigatti I, Nouvellet P, Donnelly CA. Spatiotemporal variability in case fatality ratios for the 2013–2016 Ebola epidemic in West Africa. *International Journal of Infectious Diseases* **2020**; 93:48-55.
147. Feurer M, Klein A, Eggenberger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. In: *Advances in neural information processing systems*. 2962-70.

148. Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *The Journal of Machine Learning Research* **2017**; 18:826-30.
149. Aristodemou L, Tietze F. The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. *World Patent Information* **2018**; 55:37-51.
150. Pregibon D. Logistic Regression Diagnostics. *Ann Statist* **1981**; 9:705-24.
151. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **2010**; 4:266-98.
152. Nagy H, Watanabe K, Hirano M. Prediction of sediment load concentration in rivers using artificial neural network model. *Journal of Hydraulic Engineering* **2002**; 128:588-95.
153. Ishwaran H, Kogalur UB, Kogalur MUB. Package 'randomForestSRC'. **2019**.
154. Kapelner A, Bleich J. bartMachine: Bayesian Additive Regression Trees: R package version, **2016**.
155. Ripley B, Venables W, Ripley MB. Package 'nnet'. R package version **2016**; 7:3-12.
156. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Vol. 1. Springer series in statistics New York, **2001**.
157. Bischl B, Lang M, Kotthoff L, et al. mlr: Machine Learning in R. *The Journal of Machine Learning Research* **2016**; 17:5938-42.
158. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology* **2010**; 63:826-33.
159. Artzi M, Liberman G, Nadav G, et al. Differentiation between treatment-related changes and progressive disease in patients with high grade brain tumors using support

vector machine classification based on DCE MRI. *Journal of neuro-oncology* **2016**; 127:515-24.

160. Wei W, Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association* **2011**; 18:370-5.

161. Naushad SM, Ramaiah MJ, Pavithrakumari M, et al. Artificial neural network-based exploration of gene-nutrient interactions in folate and xenobiotic metabolic pathways that modulate susceptibility to breast cancer. *Gene* **2016**; 580:159-68.

162. Meeyai S. Logistic regression with missing data: a comparisson of handling methods and effects of percent missing values. *Journal of Traffic and Logistics Engineering* **2016**; 4:128-34.

163. Zhang Q, Rahman A, D'este C. Impute vs. ignore: Missing values for prediction. In: *The 2013 International Joint Conference on Neural Networks (IJCNN)*. IEEE:1-8.

164. Baraldi AN, Enders CK. An introduction to modern missing data analyses. *Journal of school psychology* **2010**; 48:5-37.

165. Cori A, Donnelly CA, Dorigatti I, et al. Key data for outbreak evaluation: building on the Ebola experience. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2017**; 372:20160371.

166. Colubri A, Hartley M-A, Siakor M, et al. Machine-learning Prognostic Models from the 2014–16 Ebola Outbreak: Data-harmonization Challenges, Validation Strategies, and mHealth Applications. *EClinicalMedicine* **2019**; 11:54-64.

167. Medley AM, Mavila O, Makumbi I, et al. Case Definitions Used During the First 6 Months of the 10th Ebola Virus Disease Outbreak in the Democratic Republic of the Congo—

Four Neighboring Countries, August 2018–February 2019. Morbidity and Mortality Weekly Report **2020**; 69:14.

168. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery, **2016**:785–94.

169. Olson RS, Moore JH. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In: Hutter F, Kotthoff L, Vanschoren J, eds. Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer International Publishing, **2019**:151-60.