List of Tables

List of Figures

Chapter 4

Modelling to identify determinants of sepsis mortality

4.1 Chapter overview

In this chapter I present an extension of the mortality models from Chapter 3 to address difficulties arising from the problems of separation, overparametrisation, collinearity and missing data. I use Bayesian logistic regression following multiple imputation of missing data and dimensionality reduction using factor analysis of mixed data (FAMD), to show that, broadly, inferences from the original models are sound. The association of receipt of TB therapy with survival persisted across all models. There was no clear association between more rapid administration of antibacterials or volume of intravenous fluid administered and survival, though in both cases the 95% credible intervals of effect size incorporated a clinically significant effect. A subgroup analysis using a propensity-score approach suggested that the association of TB therapy with survival was strongest in immunosuppressed and/or anaemia participants, though numbers were small and confidence intervals wide. The role of early administration of TB therapy in septic adults in Malawi is unknown, though the analyses here suggest a potentially significant impact, and, I suggest, contribute to the equipoise necessary for clinical trials.

4.2 Introduction and chapter aims

In Chapter 3, I presented an initial attempt at developing models to identify the determinants of mortality sepsis. However there were some problems with the approaches used, both

technical, and conceptual. The technical issues are: that the model may include too many parameters and be overfit; that a complete-case analysis was undertaken which can introduce bias; that included variables may be collinear which can contribute to increased variance in parameter estimates; and, perhaps most importantly, some parameters perfectly predicted outcome and so were excluded from the model. Conceptually, I argue that it is difficult to interpret parameters without first making explicit the causal model for data generation. I will cover each of these points below, before describing the techniques I used to address them.

First, the model may be overfit. Adding covariates to a model may reduce bias in parameter estimates but will increase variance (the so called bias-variance trade off[1]) which may make interpretation difficult. More parameters in a model results in a better fit to the data, but an overfit model may end up fitting to noise, rather than identifying the true data-generating process, and so would have biased inferences and poor out of sample prediction. Rules of thumb based on ~ 10 outcomes per included predictor have been suggested[2], but the process of selecting variables to include is difficult with no consensus as to how it should be achieved. Common approaches include stepwise inclusion strategies, where variables are sequentially added or removed based on some criteria of model fit or statistical significance, but these can introduce significant bias This is because the statistics used to test the parameters (and generate confidence intervals around effect sizes etc.) are based on an assumption that a single hypothesis is being tested, an assumption which is violated by the stepwise model building process. It can be shown that standard errors are too small, that p-values are biased towards zero and parameter estimates biased away from zero [2].

How, then, to select variables to include in the models presented here? *A priori* selection of variables for theoretical reasons is likely ideal, but this becomes difficult when there are a large number of potentially important predictors. Dimensionality reduction techniques (such as principal components analysis) or shrinkage methods (lasso or ridge regression) have been suggested as alternative predictor variable selection techniques[1–3]. I use a dimensionality reduction technique called factor analysis of mixed data[4] (FAMD) and compare out-of sample prediction (using cross validation) of models built using this technique to the original model - see method below.

A further problem in modelling mortality in studies of sick inpatients is collinearity, where some predictor variables can be predicted with high accuracy by other predictor variables. For example, shocked patients are likely to have elevated lactate, low blood pressure, low bicarbonate, and high heart rate and so parameter estimates become very large or unidentifiable) when these are all entered a regression model together[5]. An advantage of principal-components type dimensionality reduction (including FAMD) is that they can solve this problem by generation new coordinate systems that are constrained to be orthogonal. Missing data bias

too can be significant; I address this by using multiple imputation by chained equations where new values for each missing data point are generated by models based on all other data[6].

The finally technical modelling challenge to overcome is the phenomenon of separation, where some covariates perfectly predict outcome[5]. In the maximum likelihood framework (usually used to fit logistic regression models) the parameter estimates for such covariates are non identifiable (essentially infinite), and hence the covariates are often excluded from the model, even though they are often very strong predictors of outcome. I use Bayesian logistic regression with weakly informative priors to overcome this problem. In the Bayesian modelling framework, probability encodes our belief about parameter values, and parameter estimates are generated from a combination of the data (using the *likelihood*, see Chapter 8) and a *prior*, quantifying our belief about the parameter values prior to the modelling process. To overcome the problem of separation, Gelman et al [7] suggest using our belief that very large or very small parameter estimates (e.g. an odds ratio of 100) are unusual, which is encoded as a student's t distribution with 3 degrees of freedom, centre zero and scale 2.5. In the broadest sense, a distribution not dissimilar to a normal distribution, but with longer tails: we think that the parameters will be close to 0, but allow a chance they may be larger. In effect, this pulls the infinite parameter estimates from the data closer to 0.

Even if the parameters of the regression can be correctly specified, however, correct interpretation of predictor effects is often difficult or impossible without a clear hypothesised causal structure. For example, consider a hypothesised causal structure of death in sepsis in Figure 4.1, which I express as a directed acyclic graph (DAG); nodes represent collections of variables which theoretically specify host status (age, sex, immune status including HIV status and CD4 cell count), infection type (e.g. causative pathogen, site), disease severity (e.g physiological variables quantifying shock, hypoxia etc.), therapies administered, and outcome. Arrows (called edges in the DAG framework) show causality: host status influences infection (e.g. TB is more common in HIV) and severity (patients with advanced HIV may have more severe infection), for example, and therapies administered is likely to be influenced by disease severity (perhaps sicker patients receive antimicrobials more quickly), host status (clinicians are likely to administer different therapies to HIV-infected patients), and infection type. A standard analysis of sepsis would construct a predictive multivariable model for death by including factors which the analyst felt likely to be associated with mortality, which would usually include HIV status, CD4 cell count, physiologic variables (such as presence of shock) and infection variables (e.g. presence of bloodstream infection [BSI]). The effects of the predictor variables are often then interpreted as the independent effect of the included predictors, after controlling for all others; however, this may not be the case.

For example, severity is at least in part a mediator of the effect of HIV on outcome, so the

interpretation of the coefficient of HIV in such a model is the residual effect of HIV once disease severity is accounted for. It is likely that there are direct effects of host and infection factors on outcome (dotted edges in Figure 4.1, not least because measured variables in a study are unlikely to wholly quantify disease severity, but if not then controlling for disease severity will completely remove the effect of HIV status on mortality, which may not be the analysts intention, or interpretation of parameters. This has been called the "Table 2 fallacy."[8] It is important therefore to clearly define the effect that is being sought from an analysis (e.g. the effect of HIV status on mortality) and to ascertain which factors need to be controlled for based on this. It may be that a number of different models are necessary to estimate parameters of interest, if more than one parameter is of interest. The causal inference framework provides tools to do this using DAGs[9], and the *dagitty* package in R[10] automates this framework so, when provided with a DAG, it can output the variables that must be conditioned upon to estimate the causal effect of an exposure on an outcome. In this chapter, therefore, I am clear that the aim of the analysis is to provide an estimate of the effect of treatments administered on mortality; the class of antimicrobial administered (antibacterial, antifungal, antimycobacterial or antimalarial) as well as the time-to-antimicrobial for different classes, and the volumes of intravenous fluid administered. This will inform the overarching aim of the thesis - to develop novel antimicrobial strategies for sepsis in sSA to improve outcomes.

4.3   Methods

Assuming the causal model in Figure 4.1, an estimation of the effect of administered treatment will require correcting for (or conditioning on) host, infection, and severity variables (assuming a direct effect of infection and host on outcome, as seems likely) i.e. all the variables that were included in the logistic regression model in Chapter 3. To solve the problem of nonidentifiability of the models including malaria and meningitis status, I refit the models in a Bayesian framework with weakly informative priors. A student's t distribution centred on 0 with three degrees of freedom and a scale of 2.5 was used as the prior for all parameters, following Gelman et al[7]. The model was fit using the *brms* package in R[11], which acts as a front end to the Stan probabilistic programming language[12]. Four Markov-chain Monte-Carlo (MCMC) chains each with 1000 iterations and a burn-in of 500 iterations were used with default *brms* settings. Convergence was assessed using traceplots and assessing for autocorrelation using the Gelman-Rubin diagnostic ($\hat{R}$) with a target of $\hat{R} < 1.1$). Parameter estimates were expressed as medians and 95% credible intervals.

To correct for missing-data bias, missing data were imputed using multiple imputation of
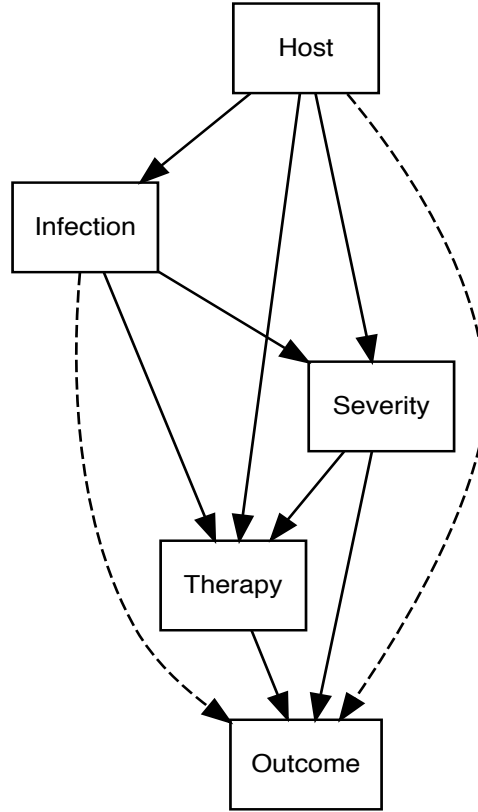
Figure 4.1: Hypothesised causal structre of death in sepsis. Host factors (e.g. age, sex, immune status) influence the type of infection; disseminated TB is more common in HIV, for example. Severity (variables quantifying e.g. shock or respiratory failure) is influenced by infection type and host factors. Therapy encodes which antimicrobials were administered and rapidity of administration of antimicrobials, and is influenced by disease severity (sicker patients may be given different therapies), host factors (HIV status may influence treatment) and the infection type (for example, malaria rapid diagnostic tests influenceing rapidity of malaria treatment). Dotted edges from host and infection to outcome are because it is not clear *a priori* whether the effect of infection and host factors are entirely mediated by disease severity: in fact, even if this were the case in a theoretical sense, the available severity variables are unlikley to completely account for the causative effect of infection type on mortality and so conditioning on all available severity variables is likely to leave some residual causative effect of infection type. See text for further discussion

chained equations using default settings in the *mice* package in R[13], with each missing variable predicted by all other missing variables n the model, to produce 5 imputed datasets. Models were fit using *brms* and then pooled parameter values calculated by taking medians and 95% confidence intervals of pooled posterior parameter estimates from all imputed datasets. The parameter estimates from complete-case analysis and multiple imputation are both presented.

One of the concerns of this model is that it is overfit - that is, there are so many parameters that it will fit to noise in the data rather than to the true data-generating process. To assess whether this was the case, I performed dimensionality reduction to collapse the predictor variables into a smaller number of variables, refit models using these variables, and then compared all models predictive ability using leave-one out cross validation. The dimensionality reduction technique that I used was factor analysis of mixed data (FAMD) from the *FactoMineR* package in R[4]. This technique uses principal component analysis (PCA) for continuous variables and multiple correspondence analysis (MCA) to generate a new orthogonal coordinate system which maximises explained variance in each FAMD axis. FAMD axis one therefore explains the most variance in the dataset, followed by FAMD axes 2 and 3, and so on. As well as reducing the dimensionality of the dataset, this technique has the advantage of ensuring an orthogonal coordinate system to tackle the problem of collinearity. The raw covariate values were used to generate these new coordinate system.

Because the exposures of interest are the therapies administered to the participants in the study, treatment variables (receipt of antibacterials, antifungals, antimalarials, antimycobacterials, and IV fluid) were left untransformed. These variables and a number of transformed FAMD variables (ranging form 1 to 5) were as predictors in new models to predict death by 28 days. The out of sample predictive ability of the models was assessed by performing leave-one-out cross validation using the *loo* package in R[14]. This estimates the out-of sample predictive ability of the model by estimating a quantity called the expected log pointwise predictive density ($ELPD$) essentially the log of the likelihood for a new, unseen dataset conditional on the current data. This quantity is estimated using leave-one-out cross validation to produce and estimate of the $ELPD$, hereafter referred to as $ELPD_{loo}$. The standard error of $ELPD_{loo}$ for a model is also calculated and so two models can be compared by comparing the $ELPD_{loo}$ difference and standard error; if the difference is greater than twice the standard error (i.e. a 95% confidence interval, assuming normality) we can be confident that one model would be expected to have greater out-of-sample predictive ability than the other[14].

The relationship between time-to-antimicrobials and mortality was assessed, initially in bivariate associations using nonparametric locally estimates scatterplot smoothing (LOESS) regression which performs a rolling linear regression[15] and estimates the probability of

death by 28 days as a function of the predictor variables. Only for antibacterials were there sufficient data to construct regression models which used time to antibacterial therapy as a predictor for death by 28 days, alongside the other treatment variables and the first three FAMD dimensions. In view of possible nonlinear relationship between time to therapy and death apparent in the bivariate plots both linear and second-order polynomial models were fit. Coefficient estimates are presented, but because interpretation of polynomial coefficients is challenging, predicted probability plots with 95% credible intervals with the levels of the other covariates set to their mean values were plotted, using all the posterior draws to generate the median prediction and 95% credible intervals.

Finally, to attempt to correct for confounding using a different method, a propensity-score matching approach was used to produce an unbiased estimate of the effect of receipt of TB therapy on 28-day mortality. Variables that had been identified as being associated with mortality from the models described above, along with variables that were associated with receipt of TB therapy apparent on bivariable analysis were included in a (maximum-likelihood fit) logistic regression model to generate a propensity score. Because HIV-uninfected participants did not have a CD4 count measured, a new dichotomous variable was used which was coded as 1 for HIV-infected participants with a CD4 count below 100 cells μL^{-1} and 0 for everyone else. Participants were then matched 1:1 on this propensity score without calliper restriction using the *MatchIt* package in R[16] and the distribution of covariates in this new cohort examined using kernal density plots and histograms. Effect of TB therapy on mortality was then expressed as risk ratios, and subgroup analysis carried out to explore whether there was any effect modification of the apparent effect of TB therapy in advanced immunosuppression (defined as CD4 cell count below μL^{-1}), anaemia (defined as haemoglobin below 8g dL^{-1}) or confirmed TB.

4.4 Results

Bayesian logistic regression with weakly informative priors succeeding in fitting the models from Chapter 3; the inferences - particularly concerning the apparent association between TB therapy and survival - were largely unchanged, including after multiple imputation of missing data (Table 4.1). Because of concerns about overfitting, dimensionality reduction with FAD was carried out; the first 3 FAMD dimensions explained 34% of the variance in the dataset, a not inconsiderable amount for 21 predictor variables. The composition of FAMD dimensions one,two and three are shown in Figure 4.2A-B expressed as a plot of the squared correlation ratio (for categorical variables) and the squared correlation coefficient (for continuous variables) for each of the original variables included in the analysis with the

FAMD dimensions. Graphically, FAMD dimension one appeared to show an association with mortality (Figure 4.2C-D.)

Table 4.1: Unadjusted and adjusted odds ratios of death by 28 days in sepsis from Bayesian logistic regression, for complete case analysis (CCA) and following multiple imputation of missing data.

Variable	aOR (95% CrI)	
	CCA	Imputed
Host Variables		
Age (per 5 years increase)	0.92 (0.66-1.25)	0.87 (0.65-1.14)
Male sex (vs female)	0.91 (0.22-3.51)	0.67 (0.18-2.36)
HIV Infected (vs uninfected)	0.21 (0.03-1.22)	0.32 (0.06-1.59)
Haemoglobin (per g dL^{-1})	**0.71 (0.54-0.91)**	**0.69 (0.52-0.90)**
Severity Variables		
Temperature (per °C)	0.72 (0.30-1.73)	0.56 (0.26-1.18)
Heart rate (per 10 min^{-1})	1.20 (0.87-1.72)	1.14 (0.85-1.57)
Mean arterial BP (per 10 mmHg)	1.20 (0.73-1.93)	1.14 (0.70-1.80)
Respiratory rate (per 10 min^{-1})	**0.25 (0.08-0.66)**	**0.38 (0.16-0.88)**
Oxygen saturation (per 5%)	0.73 (0.48-1.11)	**0.67 (0.45-0.99)**
GCS (per 1 unit)	0.76 (0.50-1.12)	0.75 (0.51-1.10)
Unable to stand	**13.79 (2.88-74.50)**	**13.64 (3.35-64.82)**
Lactate (per 1 mmol L^{-1})	1.12 (0.91-1.39)	1.13 (0.92-1.41)
White cell count (per 1x10^9 L^{-1})	0.96 (0.84-1.07)	0.94 (0.83-1.05)
Platelet count (per 100x10^9 L^{-1})	1.13 (0.70-1.78)	0.94 (0.60-1.43)
Bicarbonate (per 1 mmol L^{-1})	0.97 (0.81-1.17)	0.95 (0.81-1.11)
Urea (per 1 mmol L^{-1})	1.20 (1.00-1.45)	1.17 (1.00-1.37)
Creatinine (per 10 mmol L^{-1})	0.99 (0.92-1.08)	0.99 (0.93-1.08)
Diagnosis		
BSI (vs no BSI)	**0.04 (0.00-0.48)**	**0.04 (0.00-0.40)**
TB (vs no TB)	1.12 (0.25-5.00)	0.72 (0.18-2.69)
Malaria (vs no malaria)	0.01 (0.00-2.27)	**0.00 (0.00-0.41)**
Meningitis (vs no meningitis)	**68.53 (1.29-27384.82)**	**37.00 (1.03-6237.92)**
Treatment Received		
Received antibacterial (vs none)	8.38 (0.20-6631.38)	1.46 (0.10-30.00)
Received antifungal (vs none)	1.39 (0.23-8.60)	1.19 (0.24-5.65)
Received antimalarial (vs none)	0.03 (0.00-8.68)	0.08 (0.00-13.22)
Received antimycobacterial (vs none)	**0.11 (0.02-0.58)**	**0.12 (0.02-0.56)**
IV fluid (per L)	0.82 (0.29-2.21)	0.79 (0.31-1.98)

Note:
CCA = Complete case analysis, BP = Blood pressure, GCS = Glasgow coma scale, BSI = Bloodstream infection, TB = tuberculosis. All odds ratios are for as increase in the variables shown.

The first 5 FAMD dimensions were then used to fit models predictive of death by 28 days, along with untransformed treatment variables; the primary interest here was to see if the apparent effect of treatment administered would change under these models. Five models were fit using one, two, three, four or five FAMD dimensions; parameter estimates from these models are shown in Figure 4.3A, along with the parameter estimates form the original model using all, untransformed, parameters. Parameter estimates from treatment variables were largely unchanged across, though uncertainty was markedly increased in the original models.
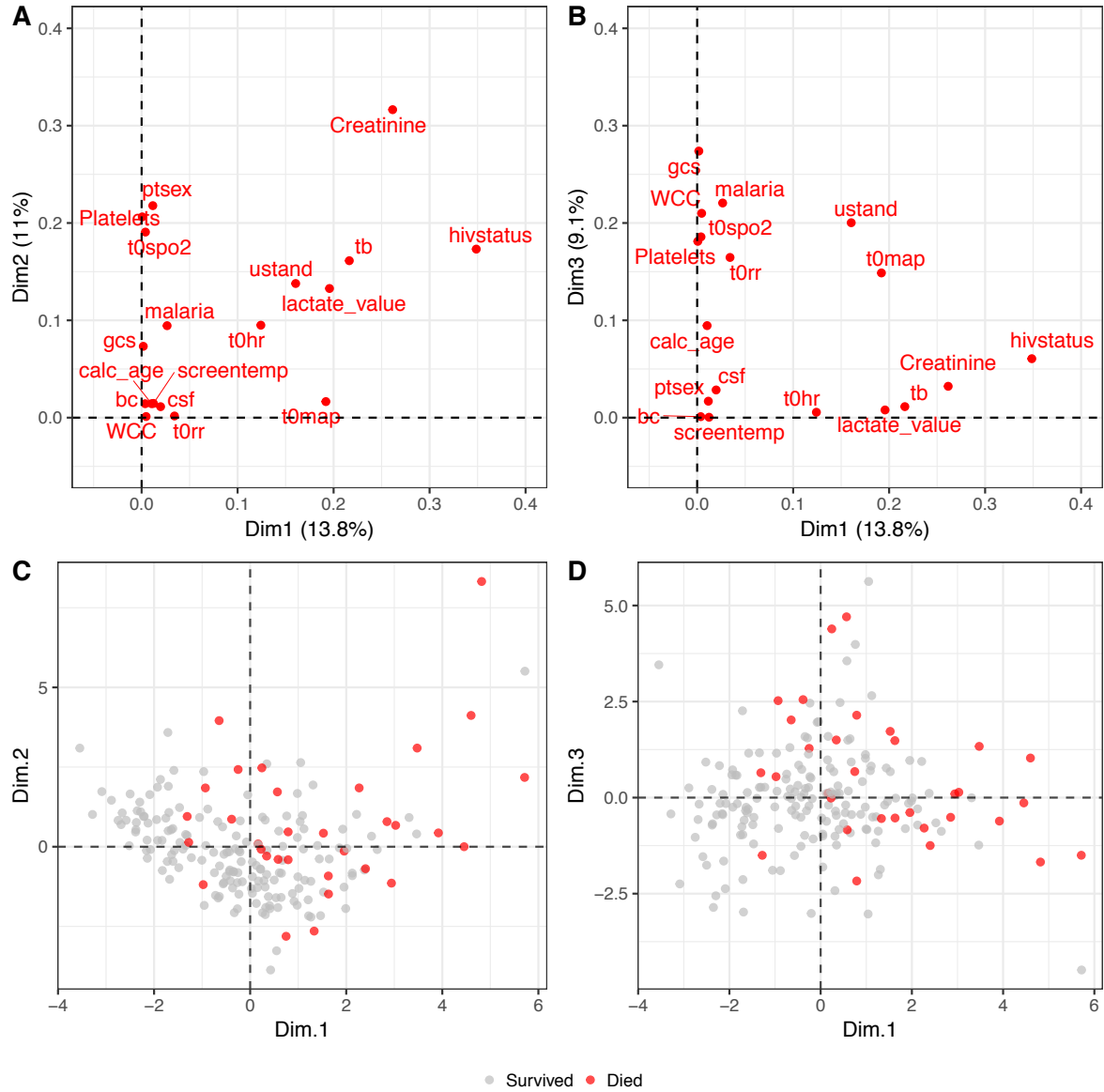
Figure 4.2: Dimensionality reduction of dataset using factor analysis of mixed data (FAMD); this is a combination of principal components analysis (PCA) for continuous variables and multiple correspondence analysis (MCA) for categorical variables, resulting in a new orthogonal coordinate system which maximises explained variance in each FAMD axis. A and B show the squared correlation ratio (for categoriacal variables) and the squared correlation coefficient (for continuous variables) with dimensions 1 and 2 (A) or 1 and 3 (B), along wih the proportion of variance explained by each axis. C shows the location of all individuals in the FAMD space, with patients who died by 28 days coloured red to show that Dim.1 seems to be associated with mortality.
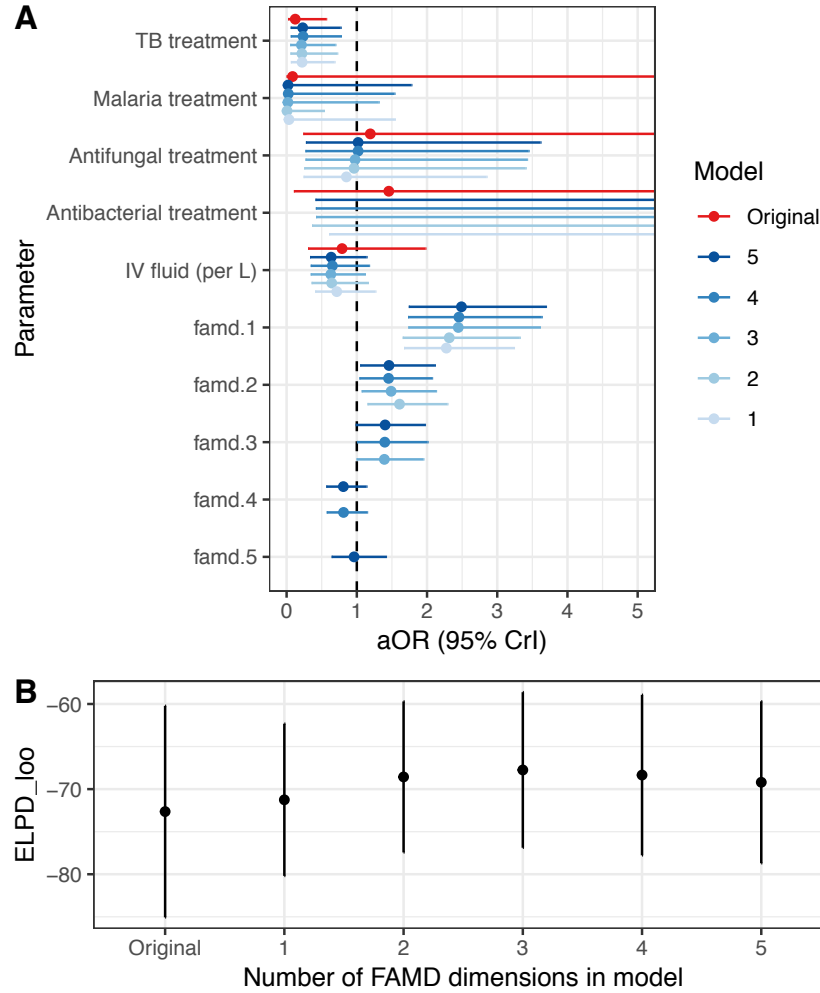
Figure 4.3: Modelling the effect of reciept of different treatments following dimensionality reduction with factor analysis of mixed data (FAMD). A (Top) shows parameter estimates for treatment variables only from the original imputed model using all raw covariate values. Models 1-5 use the first 1,2,3,4 or 5 transformed dimensions from FAMD. Parameter estimates and inferences are essentially unchanged, though there is less uncertainty generally in the estimates from the FAMD models. This would be expected as fewer parameters with less collinearity, are used. B (bottom) shows the estimated ELPD (expected log predictive denisty) from leave-one out cross validation from all the models, along with the standard error of the estimate. This is a measure of out of sample predictive accuracy: bigger (less negative) is better. One of the concerns of the original model is that it is overfit and so would have poor ELPD. In absolute terms this is true but the magnitude of the difference is much less than the standard error, meaning that out of sample prediction for all the models is broadly similar, giving confidence in the original model inferences.

Nevertheless, inferences were largely unchanged: we can be confident only that the odds ratio of the effect of TB treatment is different to zero.

The out-of-sample predictive ability of the models was assessed using the expected log predictive density (ELPD) estimate form leave one out cross validation. In absolute terms, all FAMD models greater ELPD than the original model but any differences were small compared to the standard error of the ELPD estimate. We can not be confident that any model has different out of sample predictive accuracy and therefore can be as confident in the parameter estimates form the original (untransformed) model as any other.

4.4.1 Exploring time-to antibacterials and IV fluid as determinants of mortality

Exploration of bivariate associations of mortality with time to antimicrobials and volume of intravenous fluid received are shown in Figure 4.4, where LOESS moving linear regression provides a nonparametric estimate of probability of death by 28 days as a function of treatment variables. Time to antimalarial therapy is not shown in this plot as no patient who received antimalarial therapy died. Volume of intravenous fluid administered does has no apparent effect on 28 day mortality (Figure 4.4A). It might be expected that any effect would be most apparent in participants with shock: stratifying the analysis by shock (defined as mean arterial blood pressure below 75mmHg, Figure (Figures 4.4B) once again showed no apparent relationship. Neither time to antimycobacterial or antifungal therapy showed any apparent association though confidence intervals are wide (Figures 4.4C and D).

There was no apparent relationship between time to antibacterials and2 8-day mortality up to around 40 hours, when there was a suggestion of an increased probability of death (Figure 4.4E). To explore this further, I used a logistic regression analysis, including only patients who received antibacterials (n = 207) using both linear models, fitted in a Bayesian framework as before (and following imputation of missing data), and, in view of a possible nonlinear effect, second order polynomial models. Model three from the analysis above was used (incorporating the first three FAMD dimensions), as the best fitting model. The estimates of the coefficients of the linear model is shown in Table 4.2 and the predicted probability of death by 28 days shown in Figure 4.4. In both cases it is not possible to fully rule in or out an effect of antibacterial delay. The 95% credible interval of the adjusted odds ratio for death per hour of antibacterial delay from the linear model crossed one (aOR 1.01 95% [CrI 0.98-1.04]) though incorporated a clinically relevant effect size, and the uncertainty in predictions from the polynomial of a late nonlinear effect of antibacterial delay are so wide that it is not possible to draw any conclusions.
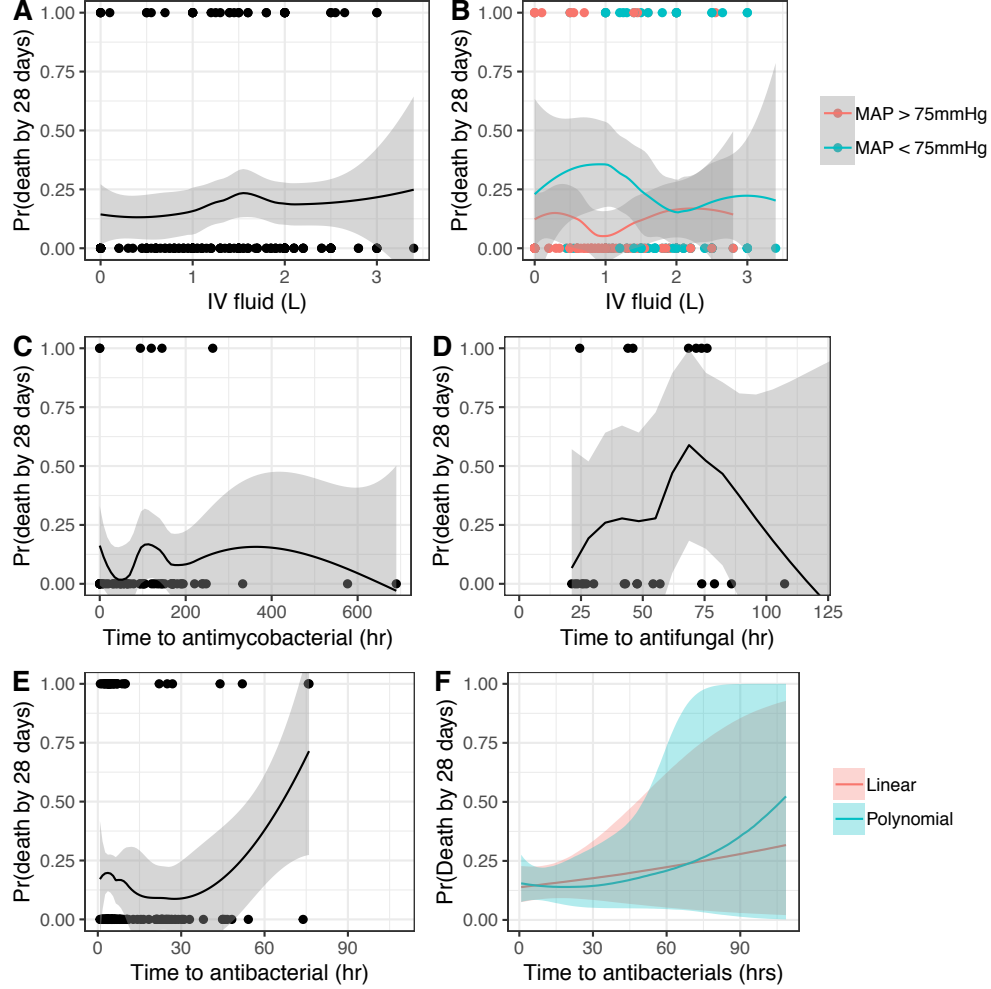
Figure 4.4: Associations of IV fliud volume and time-to-antimicrobials with death by 28 days. A-E show nonparametric regression (LOESS) of outcome (with death coded as 1 for died and 0 for survived) against various covariates; the regression line can be interpreted as the probability of death by 28 days and can be used to assess for a bivariate relationship and also the nature of any relationship (i.e. linear versus nonlinear). A: IV fluid (L), B: IV fluid stratified by presence or absence of shock (defined as MAP < 75mmHg), C: Time to antimycobacterials, D: Time to antifungals E: Time to antibacterials, with a possible late, nonlinear relationship. F: Models of time-to-antimicrobials as a predictor of mortality considering time-to-antibacterials to have a linear or second order polynomial effect. In both cases the uncertainty in the effect is such that there is no convincing relationship. Overall, there is no convincing relationship between any of these variables and death by 28 days.

Table 4.2: Adjusted odds ratio of death by 28 days per hour delay in antibacterials

Variable	aOR (95% CrI)
Time to antibacterials (per hour)	1.01 (0.98-1.04)
IV fluid (per L)	0.65 (0.36-1.16)
Received antimalarial (vs none)	0.02 (0.00-1.37)
Received antifungal (vs none)	1.03 (0.28-3.49)
Received antimycobacterial (vs none)	0.18 (0.05-0.61)
famd.1	2.56 (1.81-3.73)
famd.2	1.43 (1.04-2.01)
famd.3	1.38 (1.00-1.90)

Note:
The variables famd1,2 and 3 are the three transformed dimensions following dimensionality reduction using factor analysis of mixed data that account for the most variability in the dataset.

4.4.2 Propensity score matching and subgroup analysis

Finally, I used propensity score matching, a different method to attempt to generate unbiased estimates of the effect of receipt of TB therapy on mortality. First I examined bivariate associations of receipt of TB therapy (Table 4.3 in the chapter appendix). Patients who received TB therapy were almost all HIV-infected (88% [46/52] vs 60% [95/161] in the no-TB therapy group, p < 0.001) with lower CD4 count (median 60 vs 123 cells μL^{-1}, p = 0.006) and Haemoglobin (median 9.7 vs 11.1 g dL^{-1}), and received more antimalarials (11% [6/53] vs 3% [6/172], p = 0.037) and IV fluids (median 1.5L vs 1.2L, p= 0.02), though most of these associations would be expected to pull an estimate of the mortality effect of TB therapy towards the null, rather than inflate an effect size. More patients with a positive diagnostic test for TB received TB therapy, as might be expected (53% [28/53] of those receiving TB therapy had a positive diagnostic test for TB, versus 28% [48/172] not receiving therapy, p = 0.001), though almost all the TB treatment was empiric, as the treating clinicians did not have access to urinary LAM results (which were batch processed on frozen urines) or mycobacterial blood culture results (which take up to 6 weeks to become positive).

Factors associated with receipt of TB therapy (HIV status, CD4 count, diagnosis of TB, receipt of antimalarial therapy and volume of IV fluid received) and factors associated with mortality from the models presented above (haemoglobin, respiratory rate, oxygen saturation, inability to stand, bloodstream infection and diagnosis of malaria) were used as predictors in a logistic regression to predict receipt of TB therapy. Predictions from this model were used to generate a propensity score for each participant, and then each participant who received TB therapy was matched with one participant who did not to generate a new cohort, with better matching of covariates 4.6. The propensity-score adjusted risk ratio of survival to 28
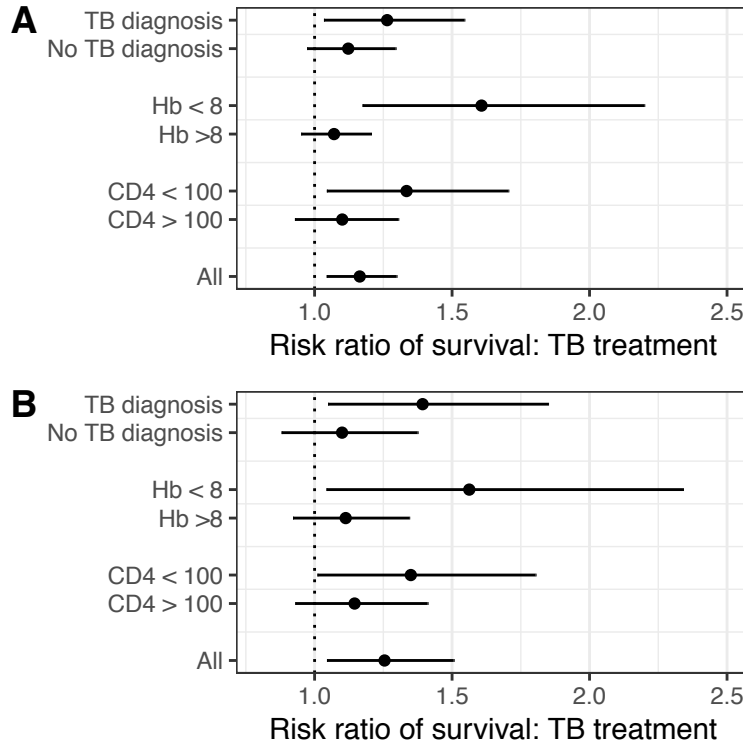
Figure 4.5: Subgroup analysis of effect of TB therapy on mortality. A (Top) shows crude (unadjusted) risk ratio for survival to 28 days; RR > 1 favours TB therapy, RR < 1 favours no TB therapy. A significant effect is seen in the immunosupressed, anaemic, and to a lesser extent, those with a confirmed diagnosis of TB. B (Bottom) shows the same analysis for the propensity-score matched cohort, showing that the overall and subgroup effects are essentially unchanged.

days in this cohort upon receipt of TB therapy was 1.25 (95% CI 1.04-1.51), similar to the unadjusted estimate (Figure 4.5). Mortality benefit seemed higher in the immunosuppressed and anaemic in absolute terms, though with significant uncertainty in the estimates (4.5): RR 1.56 (95% CI 1.04-2.24) in those with haemoglobin below 8g dL^{-1} compared to 1.11 (95% CI 0.92-1.34) above 8g dL^{-1}.

4.5 Discussion

I have presented, in this chapter, an analysis using a number of statistical techniques to further explore the associations of mortality in this cohort, and to address the difficulties encountered with the modelling strategies presented in Chapter 3. In addition I have presented an explicit hypothesised causal structure of death to facilitate a causal interpretation of the model parameters. In Chapter 3, maximum likelihood fitting of logistic regression models was likely biased because some predictors were perfectly associated with outcome, and were

excluded - the phenomenon of separation. This was overcome using Bayesian logistic regression with weakly informative priors. As expected, malaria and meningitis were strongly associated with survival and death, respectively, but more importantly inferences from the model were largely unchanged once these strong predictors were included in the model. The strongly protective effect of TB therapy persisted. Unexpectedly, bloodstream infection showed a protective effect; given *S.* Typhi was the commonest bloodstream infection isolate, which is associated with a low mortality, this is certainly plausible.

To address a concern that the model with so many parameters might be overfit, I performed dimensionality reduction with FAMD and refit the models using transformed predictor variables. As might be expected in absolute terms the out of sample predictive accuracy (as measured by ELPD) improved, but by an amount much less than the error in the estimation of ELPD. This supports the inferences from the original model.

As described in Chapter 3, time-to-antibacterials is thought to be one of the major modifiable determinants of death in sepsis in high-income settings, but in a crude (unadjusted) analysis there was no signal that this was the case in this cohort. I have expanded that analysis here, and adjusted for putative confounders, but the findings are unchanged. Fitting a logistic regression model and modelling time-to-antibacterials with a linear effect (and adjusting for confounders) found an adjusted odds ratio of 1.01 for death per hour delay with a 95% credible interval (95% CrI 0.98 - 1.04) that includes a clinically relevant effect size (i.e. the aOR of 1.04 [1.02-1.05] from the large observational study of sepsis care in New York[17]). There was a some suggestion from the raw data that a lengthy delay in antibacterials could be associated with an increase in mortality, perhaps with a late nonlinear effect; however the estimates from a nonlinear (second order polynomial) model had such high uncertainty that it is not possible to draw any firm conclusions. The analysis I present here, therefore, is consistent both with time-to-antibacterials having a similar effect on outcome as in high-income settings, and with no effect. Larger studies would be needed to distinguish these two scenarios. Similarly, uncertainties in estimates of effect of IV fluid administration are such that the estimates could contain a clinically relevant effect, and the analysis presented here does not advance the conclusions drawn in Chapter 3.

Again, however, the protective effect of TB therapy is significant and robust to correction for the putative confounders included in the analyses presented here. Using two techniques - logistic regression and propensity score matching - results in the same conclusions: that receipt of TB therapy was associated with survival in this cohort. The effect size of receipt of TB therapy (aOR 0.17 (0.03-0.74) is much greater than the effect of antibacterial delay, even in high income settings, highlighting the fact that determinants of mortality in sepsis are likely to be different in sSA. Subgroup analysis found that the effect is perhaps driven by a mortality

benefit in the immunosuppressed and/or anaemic, and the effect size in these subgroups was greater than in those with a diagnosis of TB, albeit with wide confidence intervals. If true, this would suggest that there is a benefit to empiric TB therapy outside those in whom a diagnosis of TB has been made in this study, and could contribute to the equipoise needed to consider clinical trials of empiric TB therapy in sepsis in sSA. This very tentative suggestion (numbers are small and confidence intervals wide) is dependent on the diagnostic accuracy of any tests used; the recently developed fujiLAM urinary LAM test shows significantly higher sensitivity than the Alere LAM test used here, which could negate the benefit of empiric TB therapy.

4.5.1   Limitations

I have used a number of techniques to account for confounding, but this will only address the included variables; it is very likely that there are unmeasured confounders, and these could seriously bias the conclusions drawn. This is certainly possible, but any confounder that acted to produce a spurious association between TB therapy and survival would have to be associated with both TB therapy and survival. It seems likely that the clinicians looking after the participants in this study would be more likely to treat sicker patients with TB therapy, and produce bias in the opposite direction. Nevertheless, unmeasured confounders are just that - unmeasured - and it is not possible to address this question with the data here. Given that almost all the participants who received TB therapy were HIV-infected, all conclusions regarding this should be largely be applied to people living with HIV; it is not clear (and perhaps unlikely) that the associations described here would also be present in the HIV-uninfected.

I have made an attempt to put the modelling here in the framework of causal inference. If the hypothesised causal framework presented above was right and I had truly adjusted for all confounders, then the estimates of effect I present would be true causal effects. In fact this is unlikely: both because of unmeasured confounders, and that the causal pathways I have hypothesised are almost certainly a vast oversimplification. I have made choices about inclusion of variables both in the logistic regression and propensity score analysis which could also introduce bias.

4.6   Conclusions and further work

In conclusion, the findings from the modelling work in Chapter three are largely unchanged when the models are expanded to address possible bias from excluding parameters due to

separation, missing data bias, and overfitting. Malaria and meningitis are strongly associated with survival to and death by 28 days respectively; BSI seems to be associated with survival to 28 days. The association of receipt of TB therapy with survival persists, and is greater in those who are anaemic and immunosuppressed. This is a finding that deserves further exploration; the place of early TB therapy in the treatment of HIV-infected participants with sepsis is unknown, but the data presented here may contribute to the equipoise needed for clinical trials.

4.7 Appendix

Below I show bivariable associations of receipt of TB therapy and variable distributions of the propensity-score matched cohort.

Table 4.3: Bivariable associations of receipt of TB treatment in sepsis

Variable	TB treatment	No TB treatment	p
Host Variables			
Age (years)	37.7 (32.5-42.9)	35.6 (26.8-43.6)	0.487
Male sex	30/53 (57%)	84/172 (49%)	0.349
HIV Infected*[*]	**46/52 (88%)**	**97/161 (60%)**	**<0.001**
Taking ART[†]	35/46 (76%)	82/97 (85%)	0.250
CD4 count[†] (μL^{-1})	**60.0 (26.2-114.8)**	**123.0 (39.0-274.0)**	**0.006**
Haemoglobin (x10^9 g dL^{-1})	**9.7 (7.4-11.3)**	**11.1 (8.6-13.9)**	**0.001**
Severity Variables			
Temperature (°C)	38.5 (38.0-39.2)	38.4 (37.9-39.0)	0.487
Heart rate (min^{-1}))	125.0 (110.0-134.0)	119.5 (99.8-132.0)	0.051
Systolic BP (mmHg)	92.0 (81.0-107.0)	99.0 (86.0-120.0)	0.133
Diastolic BP (mmHg)	67.0 (56.0-71.0)	65.0 (57.0-78.8)	0.486
Mean arterial BP (mmHg)	76.0 (65.3-83.7)	77.2 (65.1-91.2)	0.272
Respiratory rate (min^{-1}))	34.0 (30.0-38.0)	34.0 (32.0-37.0)	0.503
Oxygen saturation (%)	96.0 (94.0-98.0)	96.0 (95.0-98.0)	0.871
GCS	15.0 (15.0-15.0)	15.0 (15.0-15.0)	0.566
Unable to stand	13/53 (25%)	50/172 (29%)	0.601
Lactate (mmol L^{-1})	3.2 (2.4-4.9)	3.4 (2.2-5.3)	0.796
White cell count (x10^9 L^{-1})	6.4 (4.6-9.1)	6.6 (4.3-11.7)	0.595
Platelet count (x10^9 L^{-1})	225.5 (146.8-303.2)	215.0 (145.0-296.0)	0.498
Bicarbonate (mmol L^{-1})	18.0 (16.0-21.0)	20.0 (17.0-22.5)	0.065
Urea (mmol L^{-1})	5.0 (3.8-8.7)	4.6 (3.3-7.7)	0.174
Creatinine (mmol L^{-1})	76.0 (59.0-105.0)	75.5 (59.0-102.2)	0.824
Diagnosis			
BSI	4/53 (8%)	20/172 (12%)	0.611
TB	**28/53 (53%)**	**48/172 (28%)**	**0.001**
Malaria	6/53 (11%)	15/172 (9%)	0.592
Meningitis	1/53 (2%)	3/172 (2%)	1.000
No diagnosis	**15/53 (28%)**	**96/172 (56%)**	**0.001**
Treatment Received			
Antibacterials	47/53 (89%)	160/172 (93%)	0.383
Time to Antibacterials (hr)	5.1 (3.8-9.7)	5.4 (3.6-13.4)	0.844
Antifungals	8/53 (15%)	18/172 (10%)	0.337
Time to Antifungals (hr)	45.4 (25.2-60.6)	50.9 (33.2-78.3)	0.243
Antimalarials	**6/53 (11%)**	**6/172 (3%)**	**0.037**
Time to Antimalarials (hr)	4.5 (3.0-11.7)	12.5 (3.3-21.7)	0.631
IV fluid (ml)	**1.5 (1.0-2.0)**	**1.2 (0.5-2.0)**	**0.020**

Note:

BP = Blood pressure, GCS = Glasgow coma scale. Numeric variables are presented as median (IQR) and categorical variables as proportions. P-values are from Kruskal-Wallace test for continuous variables and Fisher's exact test for categorical variables.

[*] Participants with HIV status unknown not included in this row
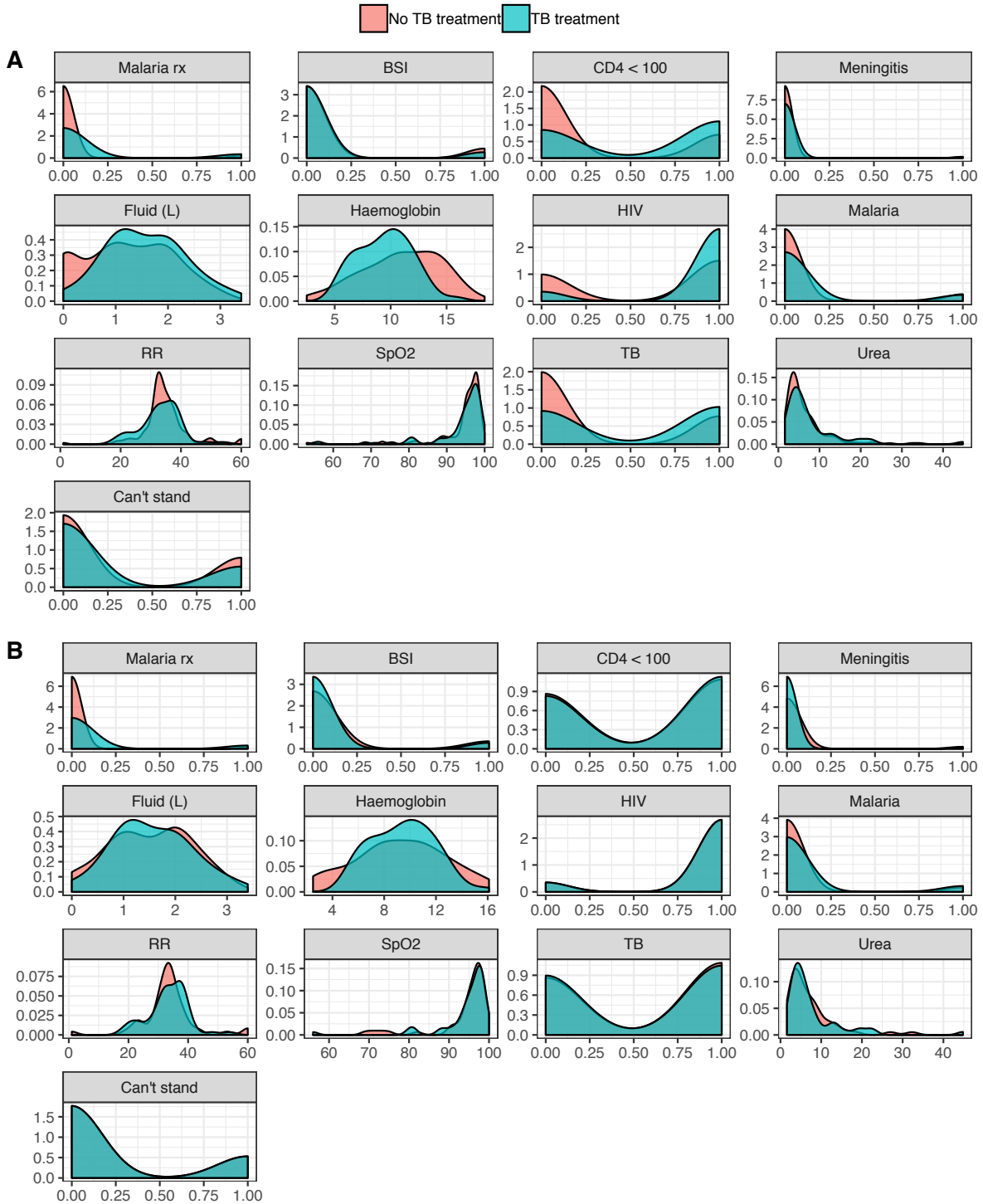
[†] Includes only HIV-infected participants

Figure 4.6: Variable distributins following propensity score matching. A: original cohort. B: Propesnity score matched cohort. BSI = bloodstream infection, RR = respiratory rate, Sp02 = Capilliary oxygen saturation, TB = tuberculosis. Categrorical variables (Malaria rx, BSI, CD4 < 100, Meningitis, HIV, Malaria, TB, Can't stand) are coded as 1 for present and 0 for absent.

References

1 Hastie T, Tibshirani R, Friedman JH(H. *The elements of statistical learning : data mining, inference, and prediction.* Springer 2009.

2 Harrell FE. *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis.* Springer 2001.

3 Tibshirani R. Regression Shrinkage and Selection via the Lasso. 1996;**58**:267–88. doi:10.2307/2346178

4 Lê S, Josse J, Husson F. FactoMineR : An R Package for Multivariate Analysis. *Journal of Statistical Software* 2008;**25**:1–18. doi:10.18637/jss.v025.i01

5 Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press 2007.

6 Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* Hoboken, NJ, USA:: John Wiley & Sons, Inc. 1987. doi:10.1002/9780470316696

7 Gelman A, Carlin JB, Stern HS *et al. Bayesian data analysis.* 3rd ed. Chapman; Hall/CRC 2004.

8 Westreich D, Greenland S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology* 2013;**177**:292–8. doi:10.1093/aje/kws412

9 Pearl J, Glymour M, Jewell NP. *Causal inference in statistics : a primer.* John Wiley & Sons Ltd 2016.

10 Textor J, Zander B van der, Gilthorpe MS *et al.* Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology* 2017;**45**:dyw341. doi:10.1093/ije/dyw341

11 Bürkner P-C. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 2017;**80**:1–28. doi:10.18637/jss.v080.i01

12 Carpenter B, Gelman A, Hoffman MD *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 2017;**76**:1–32. doi:10.18637/jss.v076.i01

13 Buuren S van, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011;**45**:1–67. doi:10.18637/jss.v045.i03

14 Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 2017;**27**:1413–32. doi:10.1007/s11222-016-9696-4

15 Cleveland WS, Devlin SJ. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* 1988;**83**:596–610. doi:10.1080/01621459.1988.10478639

16 Ho DE, Imai K, King G *et al.* MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* 2011;**42**:1–28. doi:10.18637/jss.v042.i08

17 Seymour CW, Gesten F, Prescott HC *et al.* Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. *New England Journal of Medicine* 2017;**376**:2235–44. doi:10.1056/NEJMoa1703058