Prediction of disease severity in Jordanian respiratory syncytial virus (RSV) patients in the presence of missingness

By

John Jackson Resser

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

Master of Science

In

Biostatistics

December 17, 2022

Nashville, Tennessee

Approved:

Andrew Spieker, Ph.D.

Amir Asiaee Taheri, Ph.D.

ACKNOWLEDGEMENTS

I would like to use this space to recognize those whose support during my time as a graduate student has proven invaluable.

I would first like to thank Dr. Andrew Spieker. I have benefited greatly from his guidance throughout this research process, and his mentorship has been key in my development as an effective researcher. I would like to extend a special thanks to Dr. Amir Asiaee Taheri for agreeing to serve as a second reader and providing actionable feedback to enhance the quality of this work. I would like to thank all of the professors who I have learned under these past two years and the teaching assistants and classmates who have supported me in those courses.

I am thankful for the support I have received from my friends and family. They have provided me with the affirmation and encouragement I needed to navigate graduate school. I am grateful to my wife, Shannon, for enduring my ramblings about the day's statistics problem and being my greatest cheerleader. Lastly, I am thankful for my Lord and Savior, Jesus Christ, and the eternal hope that I have in Him. Colossians 3:23.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Scientific Background

Respiratory syncytial virus (RSV) is a ubiquitous, single-stranded RNA virus from the Pneumoviridae family frequently detected in children under the age of two [1, 2]. It is a leading cause of acute respiratory infection (ARI) and deaths in infants, especially in low-income and middle-income countries [3]. The virus is very contagious and can be spread through direct contact with viral droplets (for example, kissing the face of a child with RSV), touching a contaminated surface, or when an infected person coughs or sneezes [4]. A previous study approximated a global incidence of 33.8 million RSV-associated cases of ARI each year; 3.4 million (roughly 10.1%) of these cases require hospitalization [5]. It has been demonstrated that the majority of children hospitalized for an RSV-associated ARI are previously healthy, indicating that the decision-making process for who should be hospitalized is predominantly directed by either exhibition of symptoms or the judgment of the physician [6, 7, 8]. It is currently unclear as to whether a data-driven risk index could identify patients at reduced need of intervention at hospital admission, and how we can quantify patient-level risk when encountering missingness in patient covariate profiles.

A variety of quantitative methods have been proposed in an effort to reduce subjectivity in the assessment of respiratory disease severity. Some studies have devised clinical scoring systems based on patient examinations conducted upon hospital admission. One study by Bierman & Pierson established a treatment protocol for adolescents with status asthmaticus [9]. A later study utilized a clinical scoring system adapted from the one developed by Bierman & Pierson to assess patient improvement over time in infants affected by acute wheezing [10]. Another developed a global respiratory severity score (GRSS) to quantify

disease severity in infants with primary RSV infection [11]. The GRSS was independently validated by a separate research group, Kubota et al., who determined that the GRSS was clinically useful in guiding the decision-making process as to which infants younger than 10 months infected with RSV need respiratory support [12].

The aforementioned studies were primarily concerned with developing procedures to summarize patient health states at the time of presentation. In contrast, we aim to establish a quantitative framework for predicting patient risk of disease severity at the time of hospital admission. We propose a risk index that makes use of several forms of patient covariate information such as medical history, clinical, and medical examination data to generate by way of a predictive model a unique severity score for each patient. These scores represent the probability at time of hospital admission that the corresponding patient will experience a major medical intervention and/or event (MMIE). We will leverage these scores to determine whether there exist obvious subgroups in less need of intervention. The resultant model may subsequently be applied to new subjects whose outcomes are not yet known to predict their probability of experiencing an MMIE. Additionally, we seek to determine how we can best characterize MMIE risk when dealing with incomplete patient covariate information. In a world where hospital resources are finite, the ability to quantify patient risk from time of admission, even when a portion of a patient's covariate profile is unknown, is indispensable.

1.2 Data

Our data are from a three-year viral surveillance prospective cohort at Al-Bashir Government Hospital in Amman, Jordan between the months of March 2010 and March 2013. This hospital, established in 1954, is the major government-run referral center in the city of Amman. It is estimated by the Ministry of Health that this hospital provided care to at least 50-60% of children residing in Amman at the time [8]. Of the 17,557 children admitted to the pediatric ward during this time period, 11,230 (64%) were under the age of two [13]. During the three-year period, children under the age of two exhibiting a fever and/or

respiratory symptoms were enrolled within 48 hours of admission to the hospital. Throat and nasal swabs for all patients in the cohort were collected and subsequently tested for 11 respiratory viruses by reverse-transcription polymerase chain reaction (RT-PCR). The data for this research specifically relate to the subset of 1,397 subjects who tested positive for RSV, independent of tests for other viruses.

After the children were enrolled, their parents or legal guardians were given a standardized questionnaire to procure demographic and clinical data. This information included: age; sex; gestational age; birth weight; type of birth delivery (caesarean section [CS] or normal spontaneous vaginal delivery [NSVD]); child and maternal history of asthma, allergic rhinitis, and eczema; family history of asthma; history of breast feeding; number of siblings; number of household members; day care attendance; smoke exposure (cigarette and/or hookah pipe); and underlying medical conditions (UMCs). UMCs in this study population included: diabetes, heart disease, Down syndrome, kidney disease, sickle cell disease, cystic fibrosis, cancer, genetic/metabolic disorders, cerebral palsy, neurological disorders, mental retardation/developmental delay, seizure disorder, chronic diarrhea, gastroesophageal reflux disease, immunodeficiency, asthma/reactive airway disease, liver disease, hypothyroidism and/or other medical conditions.

The treating physician recorded the following information relating to physical examinations once the children were admitted: duration of symptoms, decreased activity, irritability, loss of appetite, shortness of breath, apnea, heart rate, body temperature, respiratory rate, oxygen saturation level (further categorized as $\geq$95, 90-94, $\leq$89), flaring/retractions (further categorized as flaring only, retractions, and accessory muscle use), wheezing (further categorized as audible by stethoscope only, audible without stethoscope, or not specified), and cyanosis (further categorized as circumoral, generalized cyanosis, or not specified). When the children were discharged, the following data were obtained: antibiotic use, blood culture results, chest radiography results, oxygen use, admission to the intensive care unit, mechanical ventilation, length of stay in the hospital, and status at discharge [12].

A table detailing all relevant model predictors is available in the Appendix (Table 5.1). Table 1.1 below details the severity markers considered as components for the model outcome, discussed in further detail in Section 2.1. All study data were entered into a secure Research Electronic Data Capture (REDCap) database (Vanderbilt University, Nashville, Tennessee, USA). All analyses were performed using R version 4.1.2.

Table 1.1 Severity Markers

Variable name	**Description**	**Details**
`los`	Length of hospital stay	Days (continuous)
`any_icu`	Admittance to intensive care unit	0 = No, 1 = Yes
`vent`	Receipt of mechanical ventilation	0 = No, 1 = Yes
`discharge_status`	Alive or dead prior to discharge	0 = Alive, 1 = Dead
`oxygen`	Receipt of supplemental oxygen	0 = No, 1 = Yes

Chapter 2

METHODOLOGY & RESULTS

2.1 Defining Model Outcome

Before we specify a model we must first define the model outcome, which will in this case be experiencing an MMIE. However, there exists no universally established clinical definition for what constitutes an MMIE. We must therefore carefully consider how to delineate the model outcome for this research. We wish for the outcome to serve as an indicator of disease severity and a proxy for hospitalization need. Since these can be reflected in a multitude of ways, it is appropriate to opt for a composite endpoint: a combination of at least two separate severity markers. This will allow for MMIE classification to be a function of several markers rather than just one. For this study, we consider a patient who experiences an MMIE to have met the minimum threshold for hospitalization. We will denote an MMIE by the occurrence of at least one of the following four events during a patient's hospital stay: admittance to the intensive care unit (ICU); receipt of mechanical ventilation (MV); a hospital stay (LOS) exceeding a pre-specified number of days; or death prior to discharge.

Figure 2.1 shows the distribution of LOS for the 1,383 complete-case subjects. The distribution is noticeably right-skewed, indicating that extended stays in the hospital were rare. Roughly 75% of subjects in our study population were discharged from the hospital within one week. Figures 2.2 and 2.3 display the distribution of LOS stratified by MV and ICU, respectively. Together the histograms in Figure 2.2 demonstrate that, congruent with intuition, subjects admitted to the ICU at some point during their stay remained in the hospital for a longer period of time on average than subjects who were not. Comparing LOS for the two groups by a Welch's two sample t-test, it is apparent that the means of the two groups are significantly different ($p < 0.001$). Similarly, Figure 2.3 shows that patients who

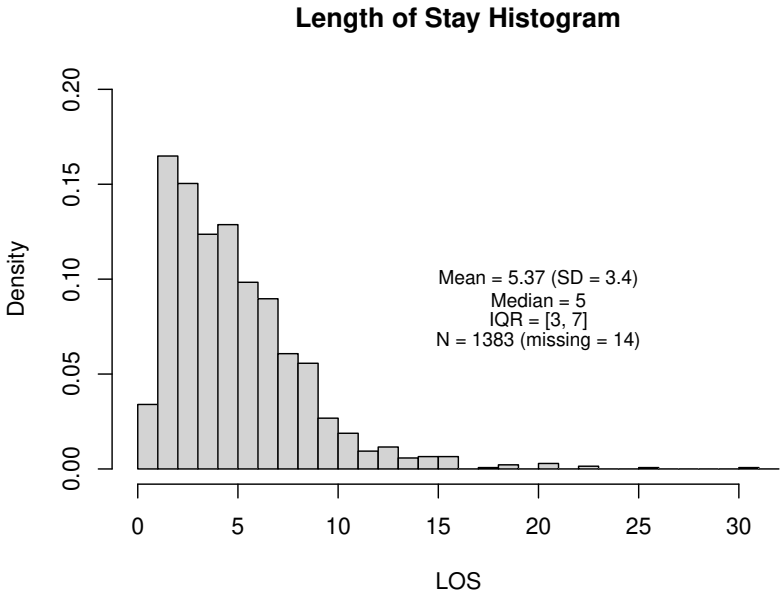

Figure 2.1 | **Histogram for hospital length of stay.** The distribution of length of stay is right-skewed, indicating that extended stays in the hospital were rare. Roughly 75% of subjects were discharged from the hospital within one week.

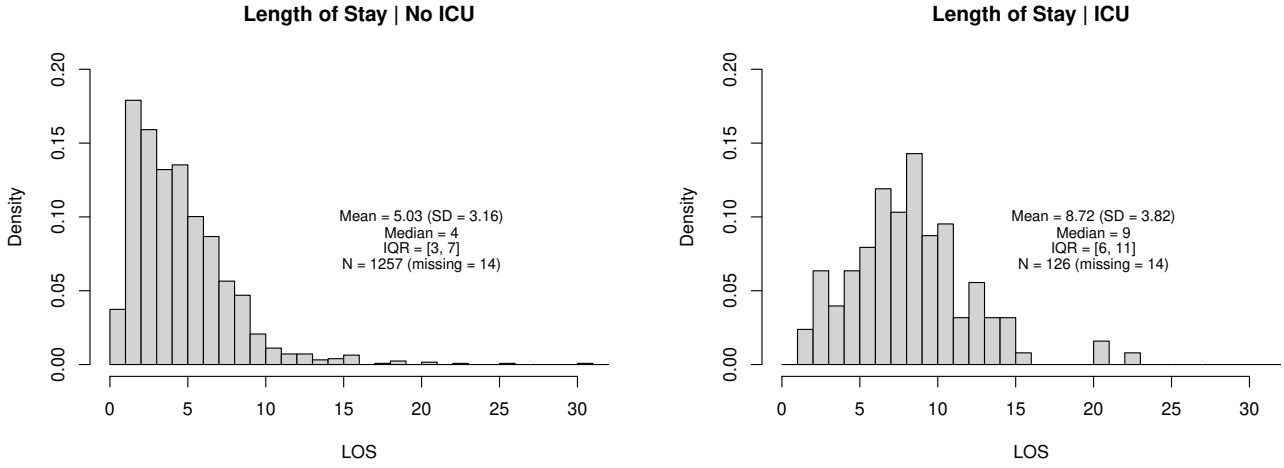

Figure 2.2 | **Histograms for hospital length of stay stratified by ICU admittance.** The left histogram shows length of stay for subjects who did not have an ICU visit. The right histogram depicts length of stay for subjects who did visit the ICU. Individuals who were admitted to intensive care had longer hospital stays on average than individuals not admitted to intensive care ($p < 0.001$).

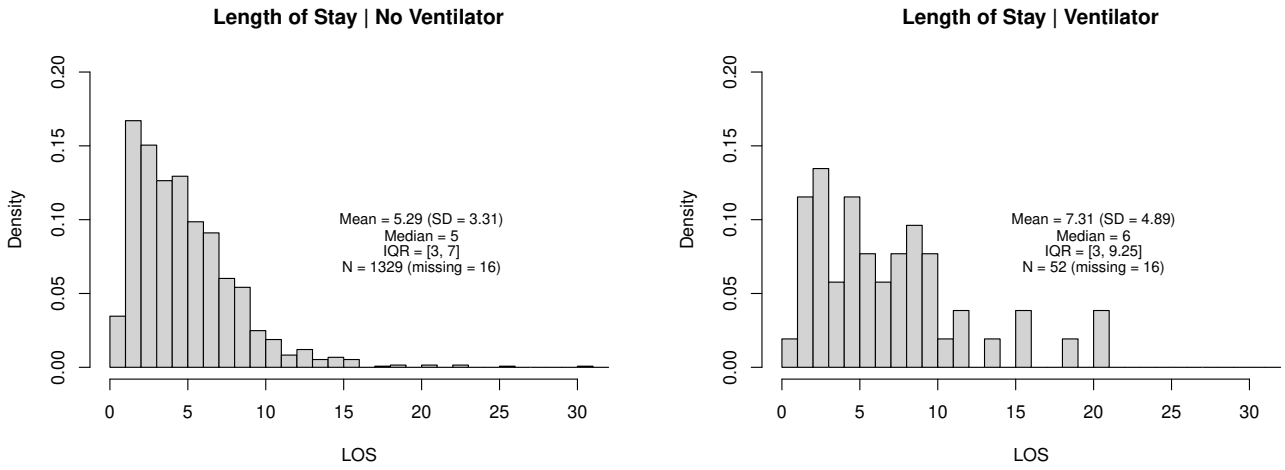

Figure 2.3 | **Histograms for hospital length of stay stratified by ventilation receipt.** The left histogram shows length of stay for subjects who did not receive mechanical ventilation. The right histogram depicts length of stay for subjects who did receive mechanical ventilation. Individuals who were on a ventilator remained at the hospital for a longer period of time on average than those who were not on a ventilator ($p < 0.001$).

received mechanical ventilation at some point during their hospitalization stayed longer on average than those who did not. The results of a Welch's t-test indicate that the means of the two groups are significantly different ($p < 0.001$). The question remains, however, as to how exactly LOS should be treated in the outcome variable configuration. At what point does LOS rise to a clinically meaningful level of severity? In making this decision, we wish to avoid the extreme situations in which nearly all patients experienced or did not experience an MMIE such that the predictive ability of the model is preserved. We determined that designating a hospital length of stay exceeding three days as severe unambiguously captures disease severity to an appropriate extent.

It was considered for a time whether a fifth element should be included in the outcome configuration: whether the patient received supplemental oxygen at some point during hospitalization. Ultimately the decision was made to exclude such a component.One basis for this is that in theory a patient can be dehospitalized and still be sent home with oxygen. We do not wish to capture within the outcome the situation where a subject is well enough

that he/she can be discharged with oxygen in tow. An even greater reason for the exclusion is that oxygen is often prescribed without regard for whether the patient actually needs it. A patient receiving oxygen does not directly imply that he/she required hospitalization or even the oxygen itself. Therefore, receipt of supplemental oxygen does not rise to the level of severity we aspire to capture with this endpoint, so the decision was made to ignore it. In summary, an MMIE is defined as experiencing at least one of the following events: ICU visit, mechanical ventilation receipt, hospital stay period exceeding three days, or dead prior to discharge. Those who met none of these criteria did not experience an MMIE. Based on this definition, out of the 1,397 observations in the data set, 923 individuals experienced a severe outcome and 458 did not. Note that for 16 individuals, it could not be determined whether the individual experienced a severe outcome on account of missing data. Among this subgroup we have complete information concerning ICU admittance, but we are missing LOS for 14 subjects, MV for 15, and discharge status for 15. For subject 341 the only missing variable was discharge status, and for subject 1162 we were only missing MV. The rest of the subjects are missing all outcome variables with the exception of ICU admittance.

2.2 Defining Notation

After defining the outcome, the next step is to select an appropriate regression model to fit to the data. Regression analysis is a statistical process used to estimate the effects of P parameters ($\beta_1, \cdots, \beta_P$, where $p = 1, \cdots, P$) on an outcome for $i = 1, \cdots, N$ subjects. The i^{th} subject's covariate information is represented by $x_i = x_{i1}, \cdots, x_{iP}$. Traditional methods for regression analysis estimate the model by a least squares approach; that is, selecting values for $\beta_1, \cdots, \beta_P$ that minimize the sum of the squared residuals for the model:

$$\text{RSS} = \sum_{i=1}^{N} \left(y_i - \mathbf{X}_i^T \boldsymbol{\beta}\right)^2 = ||y - \mathbf{X}\boldsymbol{\beta}||^2$$

Modeling a binary outcome using linear regression is not appropriate, however, as not all of the resulting probabilities will be bounded between 0 and 1. We will instead consider a

logistic regression, a particular class of regression analysis used to model the probability of a binary outcome. Logistic regression will keep the probabilities properly bounded. Let Y_i represent the MMIE outcome for the i^{th} subject, and let $\pi_i = P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i)$ represent the probability that the i^{th} subject experiences the outcome given the patient's covariate information. Thus, we fit the model

$$\pi(\mathbf{x}_i) = \text{expit}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_P x_{iP}) = \text{expit}(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})$$

, where $\text{expit}(x) = \frac{e^x}{1+e^x}$. The likelihood for logistic regression takes the following form:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{Y}) &= \prod_{i=1}^{N} \pi(\mathbf{x}_i)^{y_i}(1 - \pi(\mathbf{x}_i))^{1-y_i} \\
\log \mathcal{L}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{Y}) &= \sum_{i=1}^{N} \log[\pi(\mathbf{x}_i)^{y_i}(1 - \pi(\mathbf{x}_i))^{1-y_i}] \\
&= \sum_{i=1}^{N} [y_i \log \pi(\mathbf{x}_i) + (1 - y_i) \log(1 - \pi(\mathbf{x}_i))] \\
&= \sum_{i=1}^{N} [\log(1 - \pi(\mathbf{x}_i)) + y_i \log(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)})] \\
&= \sum_{i=1}^{N} [y_i(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))]
\end{aligned}$$

To determine the values of $\beta_0, \beta_1, \cdots, \beta_P$ that maximize $\log \mathcal{L}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{Y})$, we derive the score function by taking the partial derivative with respect to β_j for $j = 0, 1, \cdots, P$, setting the expression equal to zero, and solving:

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{Y})}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \frac{\exp(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}) x_{ij}}{1 + \exp(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})} = 0$$

$$= \sum_{i=1}^{N} x_{ij}(y_i - \pi(\mathbf{x}_i)) = 0$$

$$= \mathbf{D}^T(\boldsymbol{\beta}) \mathbf{V}^{-1}(\boldsymbol{\beta})(\mathbf{Y} - \mu(\boldsymbol{\beta})) = \mathbf{X}^T(\mathbf{Y} - \mu(\boldsymbol{\beta})) = 0$$

This expression cannot be solved for $\boldsymbol{\beta}$ analytically because there exists no closed-form solution. As an alternative, the estimating equations are fit by a computational method such as the Gauss-Newton algorithm.

2.3 Estimating Risk Among Study Population

Though the inclusion of more parameters in a typical logistic regression model makes use of more of a patient's covariate profile, this does not necessarily imply that a more complex model is best. To demonstrate this, let us compare a logistic regression where one key covariate is used (Model 1) to one with five key covariates (Model 2). The covariates are selected from Table 5.1. The covariate used in Model 1 is age, measured in months. The covariates used in Model 2 are age, the sex of the child, whether the child has a history of asthma, the duration of the child's current illness measured in days, and whether the child exhibits cyanosis. We use the MMIE outcome previously defined in Section 2.1 for both models. For this demonstration, we will only consider the complete-case subjects.

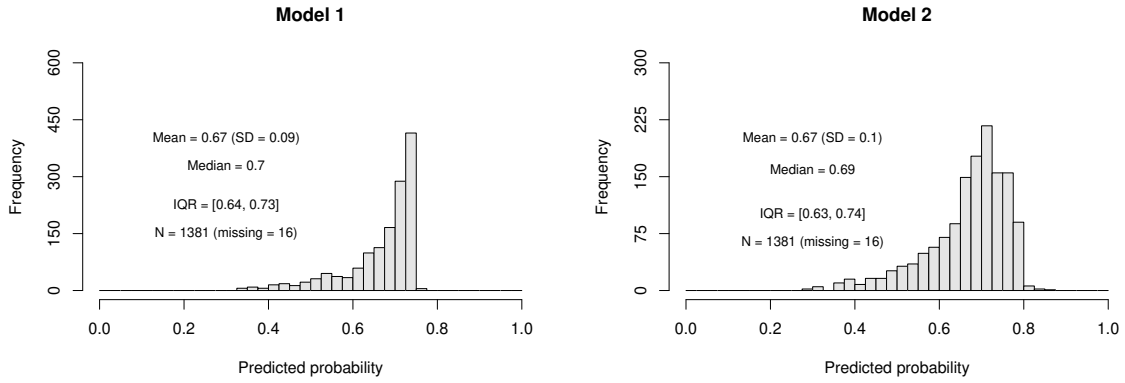

Figure 2.4 | **Histograms for logistic regression.** Though Model 2 makes use of more of a patient's covariate profile than Model 1, we wish to ensure that the superior predictive performance of the more comprehensive model is directly related to its ability to better predict the outcome, not that it is merely a better fit to the data used to generate it. To mitigate this issue, we will invoke a ridge penalty when specifying the model.

It is apparent from the histograms in Figure 2.4 that the Model 2 approach results in increased heterogeneity in the risk score distribution compared to the Model 1 approach,

suggesting that the use of additional covariates is in general more desirable. However, we wish to ensure the superior predictive performance of the more comprehensive Model 2 is directly related to its ability to better predict the outcome rather than the model merely being a better fit to the data used to generate it. The addition of more covariates is not guaranteed to yield a better model in terms of prediction, and we wish to avoid the dangers of overfitting. An alternative model-building approach that will help to mitigate the aforementioned challenges is known as penalized regression.

Penalized regression is a type of regression analysis where each of the model coefficient estimates undergoes a certain degree of shrinkage toward zero, thereby reducing the overall variance for and improving the out-of-sample predictive ability of the model. The mechanism by which this is accomplished is related to the well-known bias-variance tradeoff, whereby a small increase in bias is sacrificed for a significant reduction in variance to minimize the overall prediction error. Ridge regression is a form of penalized regression where, instead of estimating the coefficients by minimizing the sum of the squared residuals, the following expression is minimized:

$$\sum_{i=1}^{N} \left(y_i - \mathbf{X}_i^T \boldsymbol{\beta}\right)^2 + \lambda \sum_{p=1}^{P} \beta_p^2 = ||y - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda ||\boldsymbol{\beta}||^2$$

Notice that the above expression is similar to the expression minimized under traditional regression analysis. The difference is in the presence of the second term in the above expression, $\lambda||\boldsymbol{\beta}||^2$. This term acts as a shrinkage penalty; it causes the coefficients to be drawn closer to zero. The overall effect of the penalty term is modulated by the value chosen for λ. A value of zero for λ produces classical least squares and no shrinkage will be applied; as λ approaches ∞, the effect of the shrinkage penalty increases and all coefficients approach zero.

We will specify a logistic regression model with a ridge penalty. The objective function

for this model will take the following form:

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))] + \lambda||\boldsymbol{\beta}||^2$$

To determine the optimal value for λ, we employ M-fold cross-validation, where $M = 10$. In this method, all observations from the data are first placed into one of two sets: a training set and a test set. From there, the training set is split into M equally-sized folds. Then, for each fold, the model is fit using the observations **not** contained in that fold and an estimate for the validation error, CV_m, is calculated. A value for λ is chosen that will minimize the mean cross-validation error,

$$\overline{CV}_\lambda = \frac{1}{M}\sum_{m=1}^{M} CV_m$$

. We may also arrive at the optimal λ by minimizing the binomial deviance,

$$D = -[Y_i \log_{10}(E_i) + (1 - Y_i)\log_{10}(1 - E_i)]$$

, where Y_i represents the i^{th} patient's outcome and E_i represents the i^{th} patient's risk score.

Once the optimal value for λ is identified, it may be used in the penalty term of the regularized logistic regression to adjust model coefficients. The models will be fit using the command `cv.glmnet` from the R package `glmnet`.

Nearly all data-related research must deal with the problem of missingness; this case is no different. One approach to handling missing data is to conduct a complete case analysis where subjects with missingness are excluded completely. A drawback to this method, however, is that potentially useful information is ignored; subjects with partial missingness are still relevant to the research question and should not be dismissed out of hand. A second limitation of complete-case analysis is that the results can be biased if the missingness pattern is informative. Another approach called multiple imputation (MI) is an improvement over complete case analysis in that it makes use of all available information in the data. This

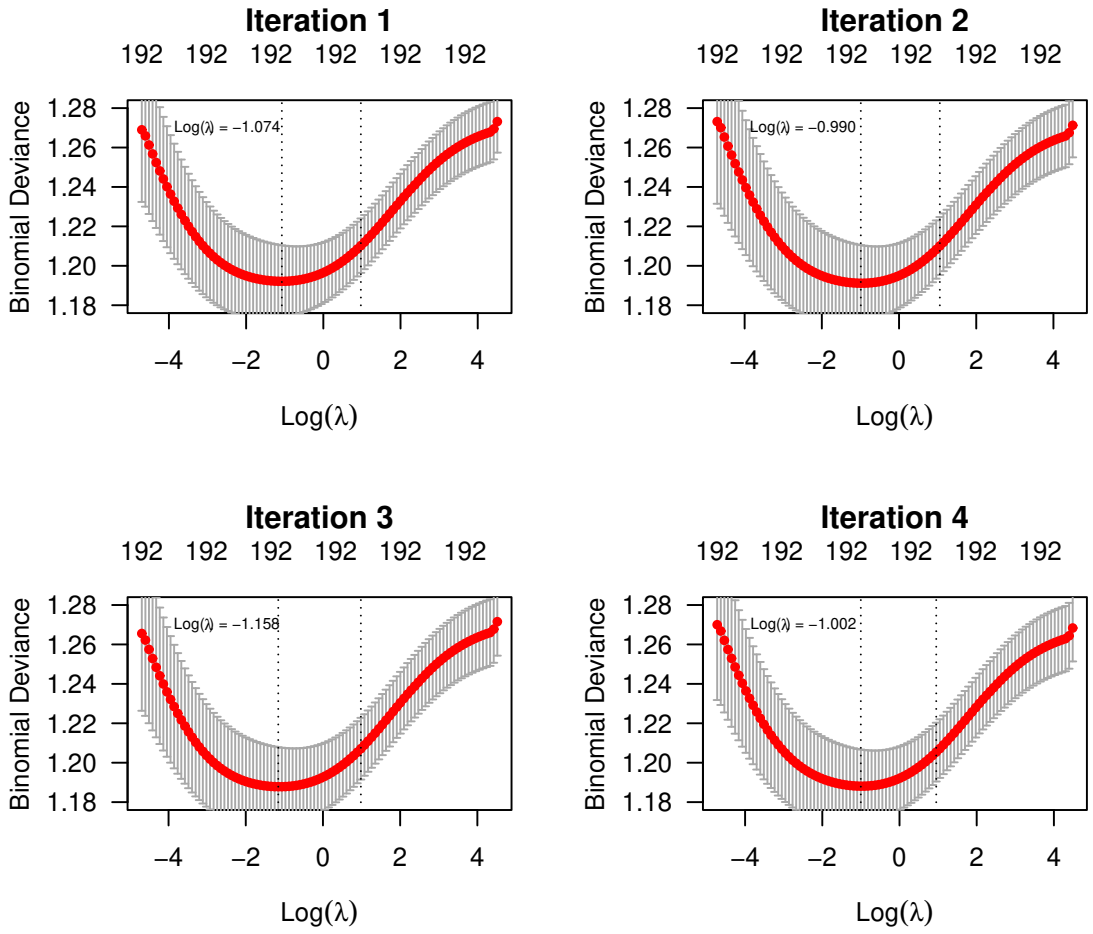

Figure 2.5 | **Finding optimal value for λ by binomial deviance minimization.** Depicted here are the binomial deviances as a function of the natural logarithm of the λ values for the first four of one hundred iterations of multiple imputation. The dashed vertical line on the left indicates the optimal value for λ that minimizes the prediction error of the model.

method generates multiple plausible copies of the entire dataset where missing values are replaced with imputed values informed by the rest of the data. The `mice` package in R implements multiple imputation by chained equations and will be used to accommodate the missing data for this research. This procedure assumes that the data are missing at random (MAR). It consists of a four step process; first, every missing value in the dataset is imputed using a simple imputation. Next, the imputed values for one variable are reverted back to missing. Then, a regression model is fit for the observed values of that variable predicted

by the other data variables with missingness. Lastly, the missing values for the variable of interest are replaced with predictions from the regression model. This is repeated for each variable in the dataset, and that entire process is repeated [14]. Reference Section 5.2 in the Appendix for pseudocode detailing the multiple imputation procedure.

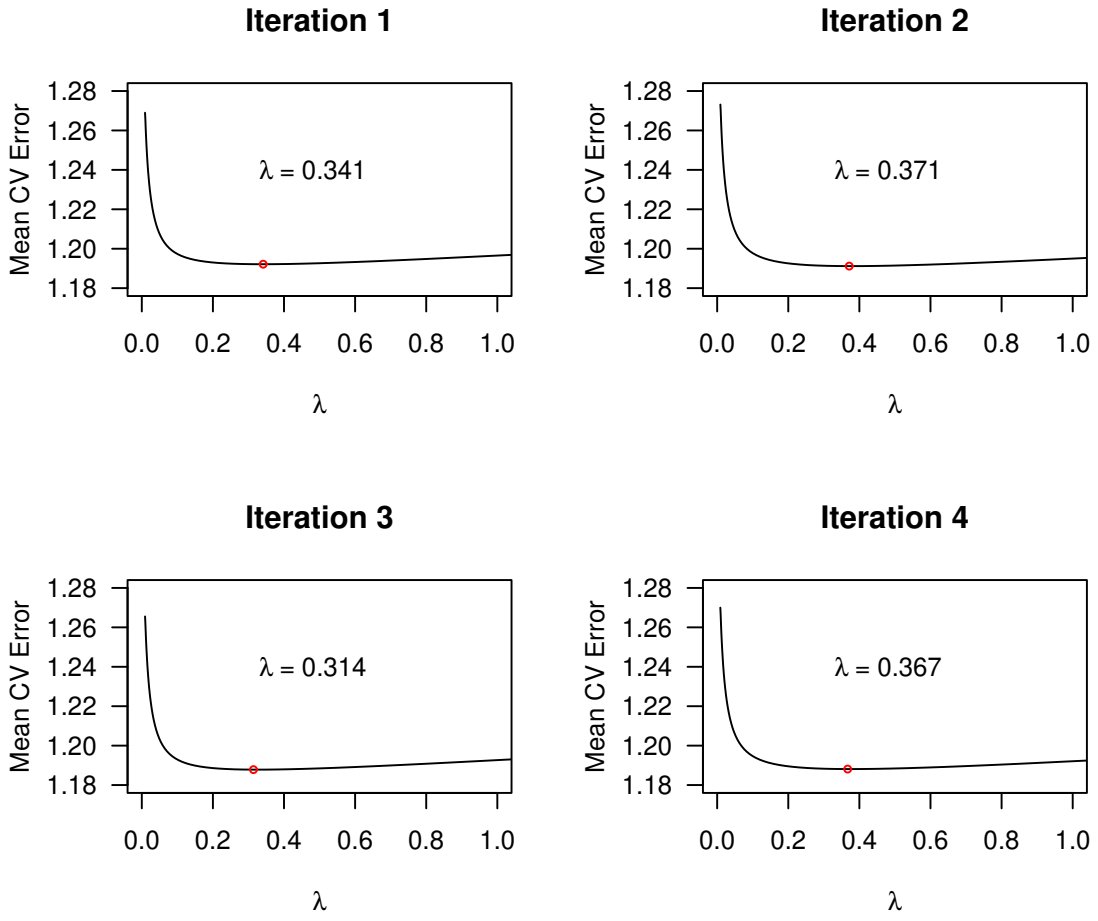

Figure 2.6 | **Finding optimal value for λ by mean cross-validation (CV) error minimization.** Depicted here are the λ values and average errors for the CV procedures corresponding to the first four of one hundred iterations of multiple imputation. In each plot, the value for λ that minimizes the mean CV error is labeled by the red dot.

We conducted one hundred rounds of multiple imputation on our data to fill in plausible values where information is unknown. This process yielded one hundred copies of our data. We next fit penalized logistic regression one hundred times to each of the data copies, where

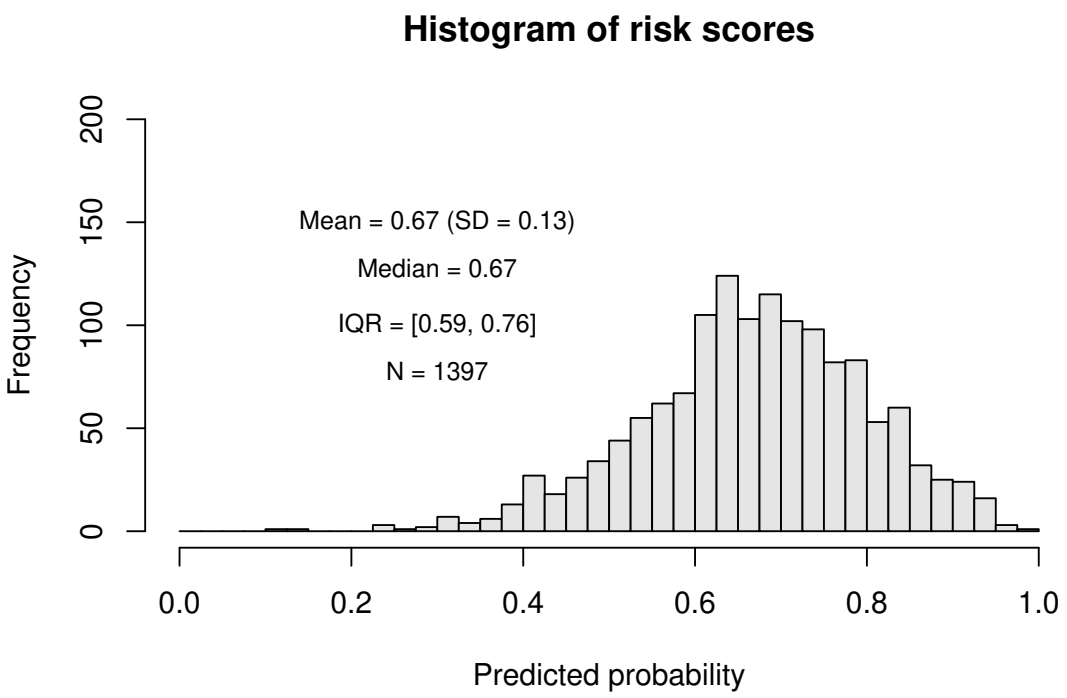

Figure 2.7 | **Histogram of average risk scores.** This histogram summarizes the average risk scores across one hundred iterations of multiple imputation for the 1,397 subjects in the study population. The average patient from this population has a risk of experiencing an MMIE of approximately $\frac{2}{3}$. The figure suggests that the collected covariate information does not indicate that unnecessary hospitalizations occurred frequently in this population.

M-fold cross-validation was employed for each iteration to determine the optimal value for λ that minimizes the prediction error of the model. Figure 2.5 demonstrates how we can find the natural logarithm for the optimal λ by minimizing the binomial deviance. The first four of one hundred iterations of multiple imputation are shown here. The optimal value is where the left vertical dashed line intersects with the x-axis. Similarly, Figure 2.6 shows how we can also derive the λ values by minimizing the mean cross-validation error. Again the first four iterations of one hundred are depicted. The red dot identifies the value for λ that minimizes the error. These λ values regulate the amount of shrinkage to be applied to the coefficients in the resulting one hundred model fits. A naive approach might then take the coefficients from each of the fits and average them to form a single model. However,

since each of the fits are a function of the random imputation procedure, it becomes unclear for patients with missingness what data to plug in to generate the risk scores. Instead we avoided this approach and, rather than aggregate to a single model, plugged the covariate information from each data copy into the corresponding model. This process returned one hundred risk scores corresponding to each of the 1,397 patients in our study population. Last, we calculated the average risk score for each subject across the one hundred iterations. These average scores are plotted in Figure 2.7.

2.4 Applying Model to New Subjects with Missingness

In real-world applications of this model, the scenario where a portion of a patient's covariate information is incomplete should not be rare. However, it is unclear how to best characterize risk of MMIE for such a patient. When a patient's covariate profile is complete, a single risk score is returned and its interpretation is straight-forward: the probability of this patient experiencing an MMIE given his/her covariate information. When dealing with partial covariate information, we can generate a distribution of plausible risk scores by leveraging multiple imputation to use the patient's existing covariate profile to inform his/her unavailable information. We will explore this mechanism by predicting risk scores for pseudo-subjects, patients from our study population for whom we have access to complete covariate information that have been altered to have different types of missingness. For this demonstration, three subjects were selected. Each of these subjects were predicted by the model to be at either low (subject A), medium (subject B), or high (subject C) risk for experiencing an MMIE prior to alteration.

We will first observe when more information is missing from patient covariate profiles relative to less. Figure 2.8 depicts four plots lettered A-D; the densities for three pseudo-subjects are shown in each plot. The variable k represents the amount of covariates each subject is missing in the corresponding plot. As we go from plot A to plot D, the amount of covariates that are missing from the patients' covariate profiles increases. In plot A, the

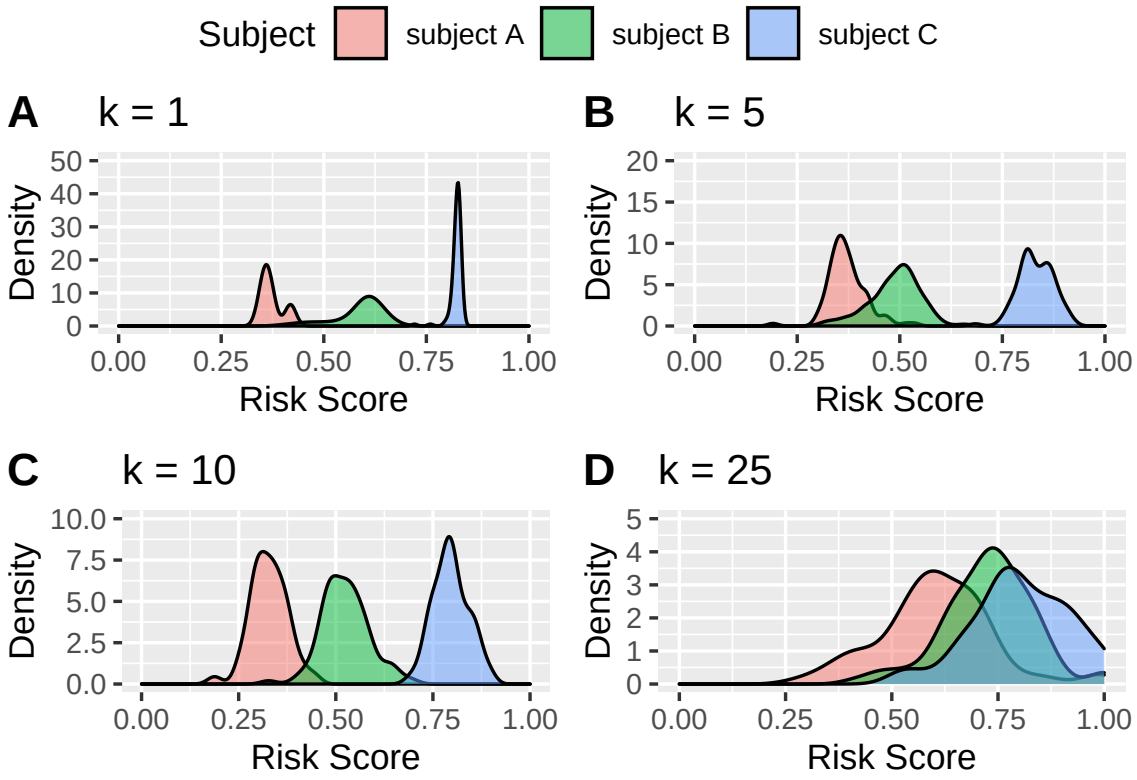

Figure 2.8 | **Density plots evaluating effect of missingness on risk prediction: amount of missing covariates.** Each plot lettered A-D depicts three densities corresponding to three pseudo-subjects. The variable k represents the number of missing covariates in each subject's covariate profile. The amount of missingness increases as we go from plot A to plot D.

age variable is missing for the three subjects. For plot B, the duration of the child's current illness in days, heart rate, body temperature, and respiratory rate variables are also missing. Additional missing covariates for plot C include whether there is a smoker in the household, gestational age of the child, whether child exhibited wheezing, whether child has a history of rhinitis, and whether the child experienced shortness of breath. For plot D, all covariates are missing with the exceptions of number of household members, whether child had an underlying heart, lung, neuromuscular, immune, or cancerous condition, whether there is a family history of asthma, and oxygen saturation level range.

We also wish to evaluate this mechanism when important variables are missing relative to less important ones. "Important" variables are those that existing scientific literature has

identified as closely related to disease severity. Three variables considered "important" are age, whether the child was born prematurely, and the gestational age of the child at birth [1, 16, 17]. We will compare the scenario where these variables are missing to when variables comparatively less related to disease severity are missing, such as whether the child's mother has a history of asthma, whether there is a smoker in the household, and the child's heart rate. Figure 2.9 contrasts these situations with two plots lettered A & B. In plot A the less important variables have been removed; in plot B the important variables are missing.

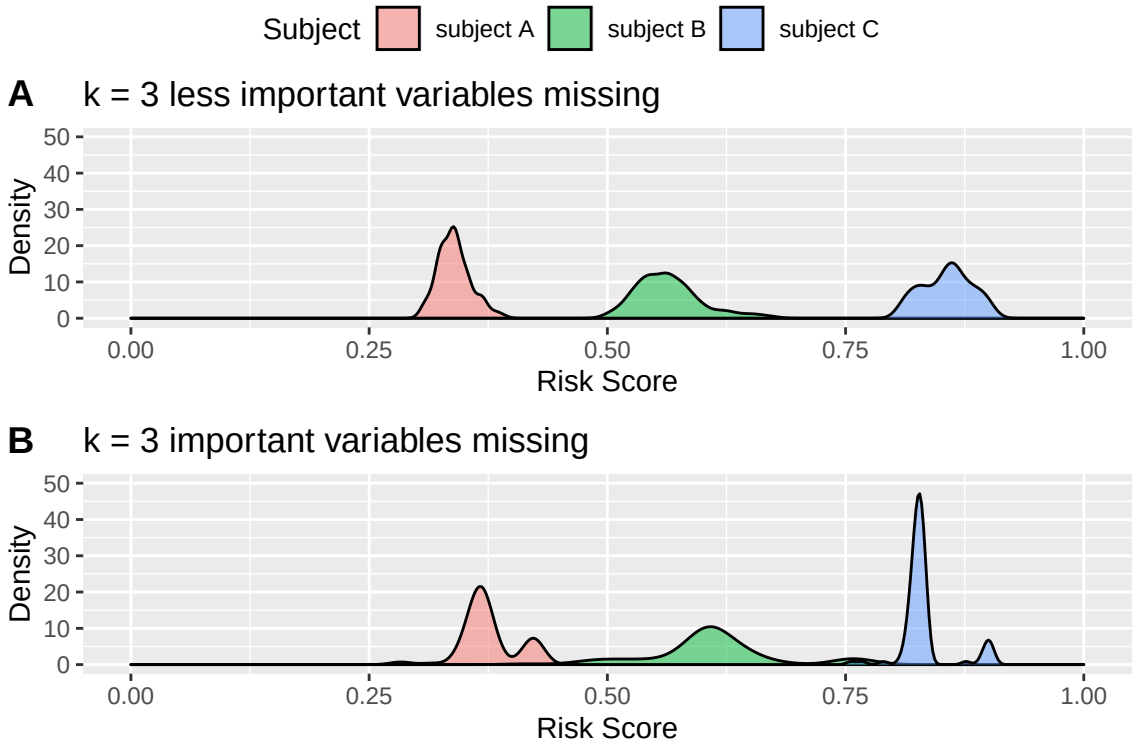

Figure 2.9 | **Density plots of pseudo-subjects with missingness in non-important and important variables.** Each plot lettered A & B depicts three densities corresponding to three pseudo-subjects. The variable k represents the number of missing covariates in each subject's covariate profile. In plot A, three variables considered to be less related to disease severity relative to the ones in plot B are missing. In plot B, three variables that have been identified by existing scientific literature as related to disease severity are missing.

Chapter 3

DISCUSSION

The 3,168 hospitalized children in the three-year cohort were tested for 16 different respiratory viruses. RSV was detected the most frequently. This finding is consistent with prior literature and demonstrates the high burden of RSV-associated ARI for this particular age group [17]. Several existing epidemiologic studies concerning RSV-associated ARI in young children have focused on comparing characteristics among subgroups defined by hospitalization status [6, 8, 18]. But hospitalization only serves as a marker of disease severity given that the hospitalization is necessary, which in our study is reflected by development of MMIE. Of the RSV-associated ARI hospitalizations in this cohort for which the data were available, approximately two-thirds ($\frac{923}{1381}$) experienced an MMIE and approximately one-third ($\frac{458}{1381}$) did not. This finding seems to challenge the notion that all hospitalizations in this cohort were necessary.

With that being said, the average risk score in Figure 2.7 was roughly $\frac{2}{3}$ and the majority of patients in this cohort had high risk scores. A naive interpretation of this figure could be that patients with a risk score under 0.50 should not have been hospitalized and those with a risk score north of 0.50 should have been hospitalized. However, committing the error of keeping a patient at the hospital who could have been discharged is significantly less harmful than the error of discharging a patient who should have remained hospitalized. This suggests a conservative application of the risk scores is most appropriate. We would thus conclude by Figure 2.7 that the collected covariate information does not indicate that unnecessary hospitalizations occur often in this study population.

As discussed in Section 2.1, key considerations were made as to what length of hospital stay rises to a meaningful level of severity. Figure 3.1 presents histograms that summarize

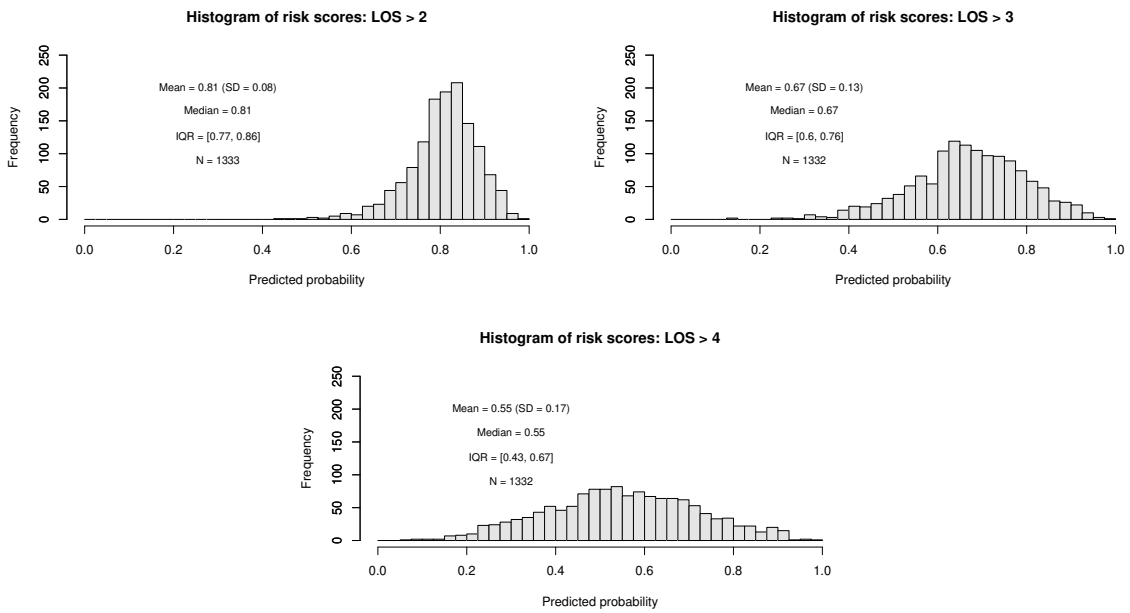

Figure 3.1 | **Risk score histograms for each considered threshold for hospital length of stay.** We wish to determine the threshold for hospital length of stay that appropriately captures disease severity for the configuration of the model outcome. The histograms in this figure summarize the distribution of risk scores for the complete-case patients. The histograms correspond to each of the potential configurations for defining the severe outcome by length of stay: length of stay exceeding two days (top left), three days (top right), and four days (bottom middle). It was determined that a length of hospital stay exceeding three days captures disease severity appropriately.

the distribution of the risk scores corresponding to the 1,333 patients. The three histograms correspond to the different thresholds being considered: lengths exceeding 2, 3, and 4 days. The difference in sample size is due to a subject who remained in the hospital for three days whose mechanical ventilation status is unknown. Referencing the figure, it is evident that the distribution of risk scores is heavily left skewed when a threshold of 2 days is chosen (top left). This phenomenon is a consequence of too many subjects meeting the severity requirement. On the other hand, when the threshold is 4 days we arrive at a much flatter distribution (bottom middle). Under this configuration those who remain in the hospital long enough to reach the threshold are likely to be considered severe on account of one of the other markers, and so the inclusion of length of stay starts to border on redundancy. It

is for the 3 day threshold that a balance is found (top right). Here, enough subjects qualify as severe that the variable's inclusion in the model is useful, but not too many such that the distribution of the risk scores becomes inflated. Thus a length of stay greater than three days was determined to appropriately constitute severity for the purposes of this study.

In Section 2.4 we were interested in determining how to best quantify MMIE risk for subjects whose covariate profiles have various forms of missingness. First, we juxtaposed the situations where more of a patient's covariate profile is unknown relative to less. Referencing Figure 2.8, when the patients' covariate profiles are more complete, we are more confident as to the patients' true risk of experiencing an MMIE. This confidence is illustrated by the reduced spread in the densities where more complete information is available relative to subsequent plots. When the extent of missingness is greater, there is less certainty regarding the subjects' true risk scores. This results in increased variability in the densities where fewer covariates are available in comparison to those in the other plots. Another noticeable phenomenon is that in situations with more complete covariate information, the central tendencies of the densities are more distinct from one another (an artifact of the original risk profiles for the selected subjects). In contrast, when more of the patients' covariate profiles are unknown, the densities appear to be merging together. They are converging to a shape we are familiar with: that of the histogram of risk scores from Figure 2.8. This trend reveals that in cases of severe absence of information, this procedure will rely on population-level tendencies for forecasting patient risk.

In addition, we investigated missingness in terms of variable relatedness to disease severity. In Figure 2.9, it is apparent that there is generally more variability in the densities in plot B where important variables are removed relative to plot A when less important variables are missing. This indicates that, in line with intuition, missingness in variables closely related to disease severity is associated with a greater degree of uncertainty regarding true patient risk relative to missingness in other variables. The reverse also holds: in situations where covariates associated with severity are available, we are more certain regarding the

true risk of MMIE experience relative to when other covariates are available. We also see bimodal densities for subjects A & C in plot B. These are an artifact of a binary variable that was made to be missing: birth prior to 37 gestational weeks (i.e. premature birth). It has been documented in existing scientific literature that a child being born prematurely is associated with experiencing severe forms of RSV. The covariate profiles for the subjects appear to indicate that they most likely did not experience premature birth as evidenced by the higher peaks at lower risk values. However, for both subjects the imputation procedure is willing to allow for the possibility of premature birth judging from the lower peaks at higher risk values, albeit at a reduced probability. In comparing the heights of the peaks associated with lower risk of MMIE experience, it appears that the covariate profile for subject A is more indicative of a person born prematurely relative to subject C.

There are some limitations to this study that should be addressed. First, this analysis is heavily dependent on how the MMIE outcome is defined, as it is used as a proxy for hospitalization requirement and an indicator of disease severity. Components related to severity such as length of stay in the hospital, use of a ventilator, a trip to the intensive care unit, and death were incorporated into the outcome configuration. It is crucial to the efficacy of the procedure that the model outcome is specified such that it appropriately emulates RSV disease severity and hospitalization need. If the specification of the study endpoint inadequately captures this, the findings of this research will prove to be invalid. Another limitation is that the data are solely comprised of patients who were hospitalized. We do not have access to data for non-hospitalized patients and thus have no way to compare those in our study population to them. The ability to compare these two subgroups would be valuable, as it is unclear whether hospitalization itself has an effect on disease severity. Thus the proposed index is not properly calibrated to a population of non-hospitalized subjects and is more reasonably applied to patients once the decision to hospitalize them has been made.

Future directions beyond this work could include validating the proposed risk index on

a separate study population. Another angle could involve the collection of two samples of hospitalized and non-hospitalized patients to then contrast them and evaluate whether hospitalization has an impact on disease severity. The risk index conceived in this paper can be used to quantitatively characterize RSV patient risk of disease severity at hospital admission. Its use could provide unique insights regarding the risk of experiencing an MMIE among hospitalized patients and assist in alleviating the global burden of RSV in children.

Chapter 4

REFERENCES

1. Piedimonte G & Perez MK. Respiratory Syncytial Virus Infection and Bronchiolitis. *Pediatr Rev.* 2014;35(12):519-530.

2. Nam HH, Ison MG. Respiratory syncytial virus infection in adults. *BMJ.* 2019;366:l5021.

3. Griffin MP, Yuan Y, Takas T, et al. Single-Dose Nirsevimab for Prevention of RSV in Preterm Infants. *N Engl J Med.* 2020;383(5):415-25.

4. Centers for Disease Control and Prevention. Transmission and Prevention of RSV (Respiratory Syncytial Virus). 2020. https://www.cdc.gov/rsv/about/transmission.html

5. Nair H, Nokes DJ, Gessner BD, et al. Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. *Lancet.* 2010;375(9725):1545-55.

6. Hall CB, Weinberg GA, Blumkin AK, et al. Respiratory syncytial virus-associated hospitalizations among children less than 24 months of age. *Pediatrics.* 2013;132(2):e341-8.

7. El Saleeby CM, Bush AJ, Harrison LM, Aitken JA, Devincenzo JP. Respiratory syncytial virus load, viral dynamics, and disease severity in previously healthy naturally infected children. *J Infect Dis.* 2011;204(7):996-1002.

8. Halasa N, Williams J, Faouri S, et al. Natural history and epidemiology of respiratory syncytial virus infection in the Middle East: Hospital surveillance for children under

age two in Jordan. *Vaccine*. 2015;33(47):6479-6487.

9. Bierman CW & Person WE. The Pharmacologic Management of Status Asthmaticus in Children. *Pediatrics*. 1974;54(2):245-247.

10. Tal A, Bavilski C, Yohai D, et al. Dexamethasone and Salbutamol in the Treatment of Acute Wheezing in Infants. *Pediatrics*. 1983;71(13):13-18.

11. Caserta MT, Qiu X, Tesini B, et al. Development of a Global Respiratory Severity Score for Respiratory Syncytial Virus Infection in Infants. *The Journal of Infectious Diseases*. 2017;215:750-756.

12. Kubota J, Hirano D, Okabe S, et al. Utility of the Global Respiratory Severity Score for predicting the need for respiratory support in infants with respiratory syncytial virus infection. *PLoS ONE*. 2021;16(7).

13. Howard LM, Rankin DA, Spieker AJ. Clinical features of parainfluenza infections among young children hospitalized for acute respiratory illness in Amman, Jordan. *BMC Infect Dis*. 2021;21(323).

14. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*. 2011;20(1):40-49.

15. Acero-Bedoya S, Wozniak PS, Sánchez PJ, et al. Recent Trends in RSV Immunoprophylaxis: Clinical Implications for the Infant. *American Journal of Perinatology*. 2019;36(suppl S2):S63-S67.

16. Garcia CG, Bhore R, Soriano-Fallas A, et al. Risk Factors in Children Hospitalized with RSV Bronchiolitis Versus Non-RSV Bronchiolitis. *Pediatrics*. 2010;126(6):e1453-e1460.

17. Haddadin Z, Spieker AJ, Resser JJ, et al. A risk index for major medical interventions and events in young children hospitalized with respiratory syncytial virus-associated acute respiratory infections.

18. Caserta MT, Yang H, Bandyopadhyay S, et al. Measuring the Severity of Respiratory Illness in the First 2 Years of Life in Preterm and Term Infants. *The Journal of Pediatrics*. 2019;214:12-19.e3.

Chapter 5

APPENDIX

5.1 Table of Model Predictors

Table 5.1 Model Predictors

Variable name	Description	Details
age_months_revised	Age	Months (continuous)
sex_child	Sex of child	0 = Female, 1 = Male
delivery_child	Type of child's delivery at birth	0 = CS, 1 = NSVD
gestage_week	Gestation age at child's birth	Weeks (continuous)
premature	Child born before 37 gestational weeks	0 = No, 1 = Yes
breastfeeding	History of breastfeeding	0 = No, 1 = Yes
breastfeeding_time	Total time of breastfeeding	Months (continuous)
no_siblings	Number of siblings in same household	Simple count (continuous)
no_household	Number of household members	Simple count (continuous)
daycare	Does the child attend daycare?	0 = No, 1 = Yes
smoke	Does anyone in the household smoke?	0 = No, 1 = Yes
child_asthma	Patient hx of RAD?	0 = No, 1 = Yes
mother_asthma	Mother hx of asthma?	0 = No, 1 = Yes
hx	Patient has underlying heart, lung, neuromuscular, immune, cancerous condition?	0 = No, 1 = Yes
child_eczema	Patient hx of eczema?	0 = No, 1 = Yes
child_allergies	Patient hx of rhinitis?	0 = No, 1 = Yes
mom_allergies	Mother hx of rhinitis?	0 = No, 1 = Yes
fhx_asthma	Family hx of asthma?	0 = No, 1 = Yes
days_symptoms	Duration of child's current illness	Days (continuous)
app_sx	Loss of appetite	0 = No, 1 = Yes
sob_sx	Shortness of breath	0 = No, 1 = Yes
apnea_sx	Apnea/cessation of breathing	0 = No, 1 = Yes
pulse	Heart rate	Beats per minute (continuous)
temp	Body temperature	Celsius (continuous)
respiratory_rate	Respiratory rate	Breaths per minute (continuous)
o2_sat_on_room_air_range	Oxygen saturation level range	0 = $\geq$ 95, 1 = 90-94, 2 = 85-89, 3 = $<$ 85
flaring	Flaring/retractions/accessory muscle use	0 = None, 1 = Mild (flaring only), 2 = Moderate (retractions), 3 = Severe (accessory muscle use)
wheezing_yn	Wheezing	0 = No, 1 = Yes
cyanosis_yn	Cyanosis	0 = No, 1 = Yes

5.2 Multiple Imputation Pseudocode

A multiple imputation procedure has three steps:

1. Impute: fill in plausible values for missing data, resulting in multiple "complete" datasets

2. Analyze: conduct analysis procedure on each version of the complete dataset

3. Pool: aggregate analysis results

To begin we set a random seed for reproducibility, loaded relevant libraries including `mice` and `glmnet` as well as the data itself, and defined the number of iterations of multiple imputation we wished to conduct (in this case, 100). We defined the outcome as described in Section 2.1.

We made the object `pdat` a data frame containing all model variables and the outcome. We used the command `mice` to create 100 copies of `pdat` where the unknown data were replaced with plausible values informed by the known data. We then pulled the data from the $j^{th} + 1$ imputation and stacked it under that from the first to j^{th} imputation in a data frame called `impdat`.

From there we defined a 1,397 x 100 matrix called `p.res` to hold the generated risk scores and looped through `impdat` for one hundred iterations, each time selecting the data for the corresponding iteration and assigning it to an object, `kdat`. We then initialized a matrix `Xmat` and filled it with two-way interactions for the demographic/social, history, and clinical data. A final data frame, `mdat`, was derived that held the first and second order terms without place-holder, duplicated, or linearly dependent columns.

From this data frame we found that iteration's imputed design and outcome matrices `X.EN` and `Y.EN`, respectively. These matrices were used as arguments in the `cv.glmnet` command, the output of which was saved to an object `zz`. The `predict` function nested within a previously encoded `expit` function was used to return the predicted probabilities of

MMIE experience for that iteration, which were saved in the corresponding column of `p.res`. This process was repeated over 100 iterations, and the imputed scores were aggregated to a vector `pr` using `rowMeans`.