Statistical Methods for Predicting Dengue Diagnosis using Clinical and LC-MS Data

By

Carolyn Louise Cotterman

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy
in
Biostatistics
in the
Graduate Division
of the
University of California, Berkeley

Committee in Charge:
Professor Alan E. Hubbard, Chair
Professor Eva Harris
Professor Mark J. van der Laan

Fall 2015

Abstract

Statistical Methods for Predicting Dengue Diagnosis using Clinical and LC-MS Data

by

Carolyn Louise Cotterman

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Alan E. Hubbard, Chair

Dengue virus is the most widespread arthropod-borne virus affecting humans, with as many as 528 million annual infections each year. Of particular concern are the subset of cases which develop into life-threatening dengue hemorrhagic fever, and those which further progress into dengue shock syndrome. Non-invasive tools that accurately differentiate dengue and its subtypes from other viral infections early in the disease progression are vital for timely therapeutic intervention and supportive care. Unfortunately, such tools are sorely lacking. Using liquid chromatography-mass spectrometry (LC-MS), we detect tens of thousands of molecular features in serum, saliva, and urine of suspected dengue patients in Nicaragua. We then use machine-learning methods to help identify candidate small molecule biomarkers which, along with easily obtainable clinical data, predict dengue diagnosis and prognosis. Our findings should aid in developing a low-cost diagnostic tool for use in the field.

Acknowledgements

This dissertation would not have been possible if it weren't for my amazing community of advisors, collaborators, friends and family.

In terms of advisors, Alan Hubbard was an incredible source of wisdom both within and outside of the realm of statistics. I also feel extremely fortunate to have received guidance from Eva Harris, whose positive energy is at least as infectious as the diseases she studies. Finally, I'm grateful for Mark van der Laan, who took the time to keep my theory section in check despite his multitude of other obligations related to ridding the world of incorrect statistical inference.

Thank you also to my official collaborators in Nicaragua and Colorado and to my unofficial collaborators at Berkeley. In particular, thank you to Natalia Voge, Barb Andre, Kristof Webb, Rushika Perera, and Barry Beaty for providing the LC-MS data, and to Lionel Gresh and Douglass Javier for providing the clinical data. Lionel also deserves a big thank you for generously answering my annoyingly long list of detailed questions with much patience. Thank you also to Ben Bowen, who gave me some crucial pointers early on, and to Sue Celniker for facilitating much of my work. Finally, thank you to Aubree Gordon and Hope Biswas for introducing me to this project and to Sam Lendle for sharing his "SuPy learner" code.

I additionally owe a large debt of gratitude to the brilliant, dedicated members of UC Berkeley who were not directly involved in my dissertation but whom I credit for making me both a better statistician and a better person. Working with David Levine was an enormously enriching experience and one that I will continue to consider a highlight of my life eighty years from now. Kris Madsen, for whom I have boundless respect and admiration, has also been an incredible role model. Finally, thank you to my officemates, Marla Johnson and Boriska Toth, for keeping me [relatively] sane, to Maureen Lahiff and Molly Davies for their consistent encouragement, to Sharon Norris, who has superhuman powers for everything bureaucracy-related, and to Jon McAuliffe, Kari Kaufman, and Philip Stark, who were instrumental in developing my conceptual understanding of statistics.

And, of course, an enormous thank you to Nathan Boley, for his continual love and support. I look forward to our post-PhD life together!

Contents

Introduction

0.1 Dengue background

The dengue fever virus (DENV) is an arthropod-borne RNA virus transmitted between humans by mosquitoes. It is a leading cause of serious illness and death among children in some Latin American and Asian countries [71], having rapidly expanded its global footprint over the last 50 years in the absence of an effective licensed vaccine and vector control strategy [54].

Symptoms and classifications: The World Health Organization (WHO) has traditionally classified dengue cases into three groups: dengue fever (DF), dengue hemorrhagic fever (DHF), and dengue shock syndrome (DSS). Most dengue cases are non-severe, generally involving a high fever, muscle pain, and rash. However, some cases are more serious, involving hemorrhagic fever and plasma leakage in the case of dengue hemorrhagic fever, and possibly additionally involving hypotension or narrow pulse pressure in the case of dengue shock syndrome. As many as 5% of severe dengue cases (DHF or DSS) are fatal, though with proper medical care, this rate can be reduced to less than 1% [5]. Different signs and symptoms appear at different times over the course of the illness, in some cases disappearing and then re-emerging (Figure 1).

Global prevalence and transmission: In the past 50 years, the prevalence of dengue has increased 30-fold; it is now endemic in 112 countries [39]. In 2010 there were an estimated 390 million dengue infections (95% credible interval $284 - 528$ million). Ninety-six million of these infections were "apparent cases" (95% credible interval $67 - 136$ million), meaning sufficiently severe to modify a person's regular schedule, such as attending school [3]. There are an estimated $500,000$ hospitalizations for DHF/DSS each year [19], and an estimated $20,000$ deaths [40]. Dengue is spread primarily by *Aedes aegypti* mosquitoes, which primarily live in tropical and subtropical climates; thus, dengue's presence is mostly restricted to such regions (Figure 2). Mosquitoes become infected when they feed on a human host during the viraemia period. It then takes about ten days for the virus to pass from the mosquito's

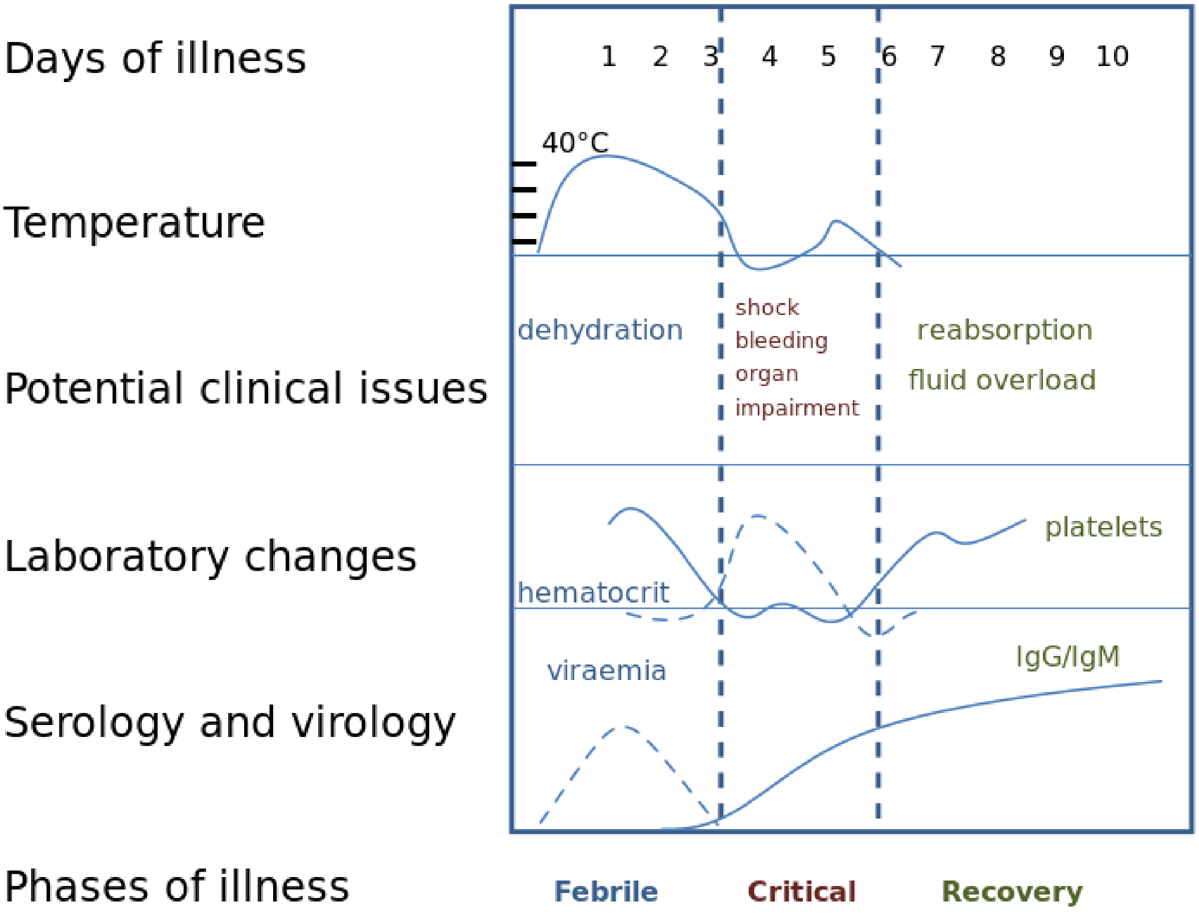

Figure 1: **Clinical progression of illness** (WHO, 2009 [75]).

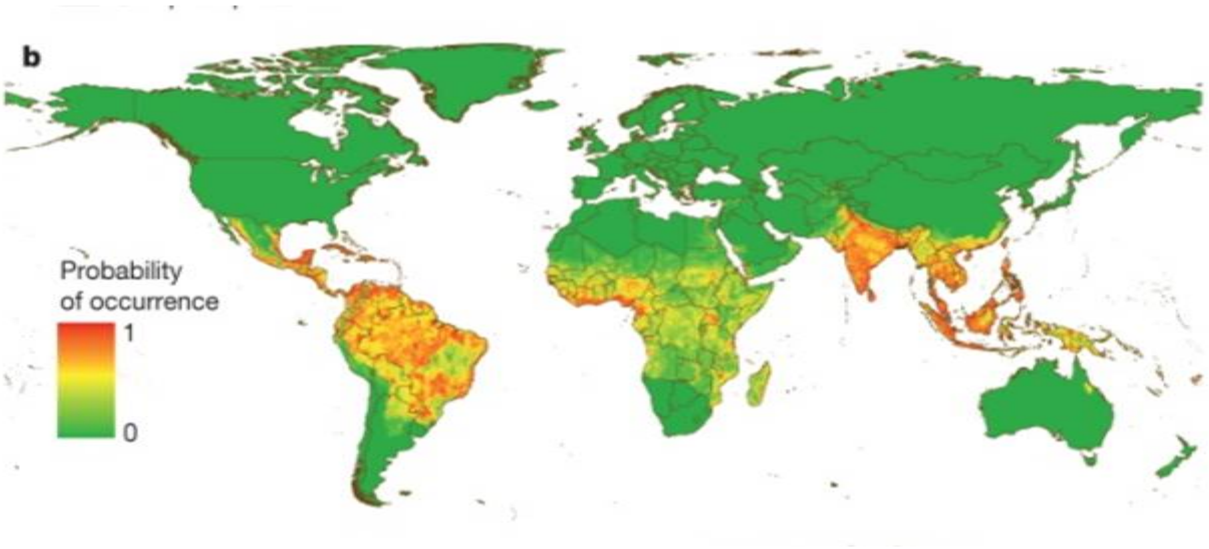

Figure 2: **Probability of occurrence of dengue infection within each 5 km x 5 km square globally** (Bhatt, 2013 [3]).

intestinal tract to its salivary glands, at which point the mosquito may spread the virus to another human [19]. Human-to-human transmission of dengue can occur through shared blood products, with other modes of transmission being unusual [59].

Virology: The dengue virus belongs to the *flavivirus* genus. Other members of this genus include the yellow fever, West Nile, and Japanese encephalitis viruses. The dengue virus encodes three structural proteins and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5). There are four dengue serotypes, which share about 65% of their genome – approximately the same amount of genetic overlap that the Japanese encephalitis virus has with the West Nile virus. Despite these genetic differences, the dengue serotypes share a nearly identical pathogenesis [19]. While infection with one serotype provides increased future immunity from that serotype, it simultaneously increases the likelihood that a subsequent infection with one of the other four serotypes will result in severe symptoms through a process known as antibody dependent enhancement [39].

Prevention: Efforts to reduce the prevalence of dengue have focused on (a) the elimination and incapacitation of the Aedes aegypti through genetic modification and habitat reduction, (b) transmission reduction through the use of bed nets and insecticides, and (c) vaccine development. No vaccine for dengue is yet commercially available, though progress in recent years has accelerated dramatically, and there are now several candidates in clinical trials. In the most advanced stage of development is a vaccine by Sanofi Pasteur that is a tetravalent mixture of live attenuated viruses representing each serotype. This vaccine, which requires three immunizations over a 12 month period, achieved moderate efficacy (56%) in phase III clinical trials. Efficacy across serotypes varied widely, ranging from 35% for serotype 2 to 78% for serotype 3 [8]. While vaccines will surely be part of the solution going forward, it is clear that an effective defense against dengue will additionally require other tactics, such as vector control.

Treatment: There is no cure for dengue, but we can dramatically improve health outcomes through supportive care, such as oral or intravenous re-hydration, or blood transfusion and oxygen for more severe cases [39]. It is thus critical to identify the patients who are likely to benefit from hospital care.

Diagnostics: Commonly used tests for dengue all involve the use of blood (or plasma) samples. In short, a diagnosis can be made by (a) detecting viral nucleic acid, (b) detecting proteins encoded by the virus, or (c) detecting dengue-specific antibodies produced by the host. The sensitivity of each approach is influenced by the duration of the patient's illness (Figure 3) as well as the patient's history with dengue.

During the febrile phase, detection of viral nucleic acid in serum by means of reverse-transcriptase–polymerase-chain-reaction (RT-PCR) is sufficient for a confirmatory diagnosis, though such a test is rarely available in rural areas. RT-PCR works by transcribing RNA into cDNA, which is then amplified by PCR. Amplification with PCR works by first physically separating the double helix of the sample DNA through the application of heat, and then using DNA polymerase to selectively amplify the target DNA by using primers (short DNA fragments) containing sequences complementary to the target region.

Also during the febrile phase, a diagnosis can be achieved by detecting the virus-expressed soluble nonstructural protein 1 (NS1). Though the sensitivity of diagnostic tests based on NS1 detection can exceed 90% for primary infections, they are significantly lower (60 to 80%) in secondary infections, reflecting an anamnestic serologic response due to a previous dengue virus or related flavivirus infection [54]. NS1 detection can be done using an enzyme-linked immunosorbent assay (ELISA) or the lateral-flow rapid test. ELISA involves attaching the sample to a surface over which antigen-specific antibodies are applied, which will bind to the antigen of interest; if the antibodies do not match with antigens in the sample, then they will be washed away during a rinsing step. The added antibodies are linked to enzymes and in the final step, a substance containing the enzyme's substrate is added to produce a visible signal, thus indicating the quantity of the antigens of interest in the sample.

Finally, a confirmatory diagnosis can be achieved via the detection of dengue-specific anti-bodies (IgM or IgG) using ELISA or a lateral-flow rapid test. While these tests are more widely available than RT-PCR, they require paired acute and convalescent samples for a confirmatory finding: detection of IgM or IgG may be achieved as early as 4 days after fever onset, but without a paired convalescent sample (with which changes in antibody concentration over time are determined), antibody detection yields only a presumptive diagnosis [54].

No existing diagnostic tool is perfect when considering availability, sensitivity, and functionality early in the disease progression.

0.2 Dissertation overview

This project was motivated by the quest for an accurate, inexpensive test for dengue fever that could be easily implemented in a resource-limited setting early in a patient's disease progression. The ideal test would be non-invasive and would rapidly determine whether a patient with flu-like symptoms is inflicted with the dengue virus, and whether s/he is likely to experience dengue hemorrhagic fever (DHF) or dengue shock syndrome (DSS). Such a test would be highly useful, as patients likely to develop severe dengue (DHF or DSS) should be kept in a clinical setting to reduce their risk of fatality, while patients with an ordinary flu need not absorb limited hospital resources.

In the following chapters, we will explore the potential of clinical, lab, and liquid chromatography - mass spectrometry (LC-MS) data with the above goal in mind. Chapter 1 introduces the general statistical framework and methods that will be used throughout the text. In Chapter 2, we explore the potential of clinical information for diagnosing dengue and predicting severe dengue. Chapter 3 discusses the use of LC-MS data to distinguish dengue patients from patients with non-dengue febrile illnesses, and explores methods for extracting features from LC-MS data that correspond to known metabolites. Chapter 4 discusses methods for selecting a subset of features that maximize predictive power. These methods bring us closer to developing a low-cost diagnostic test by identifying which features should be targeted when obtaining all features would be prohibitively expensive or logistically infeasible.

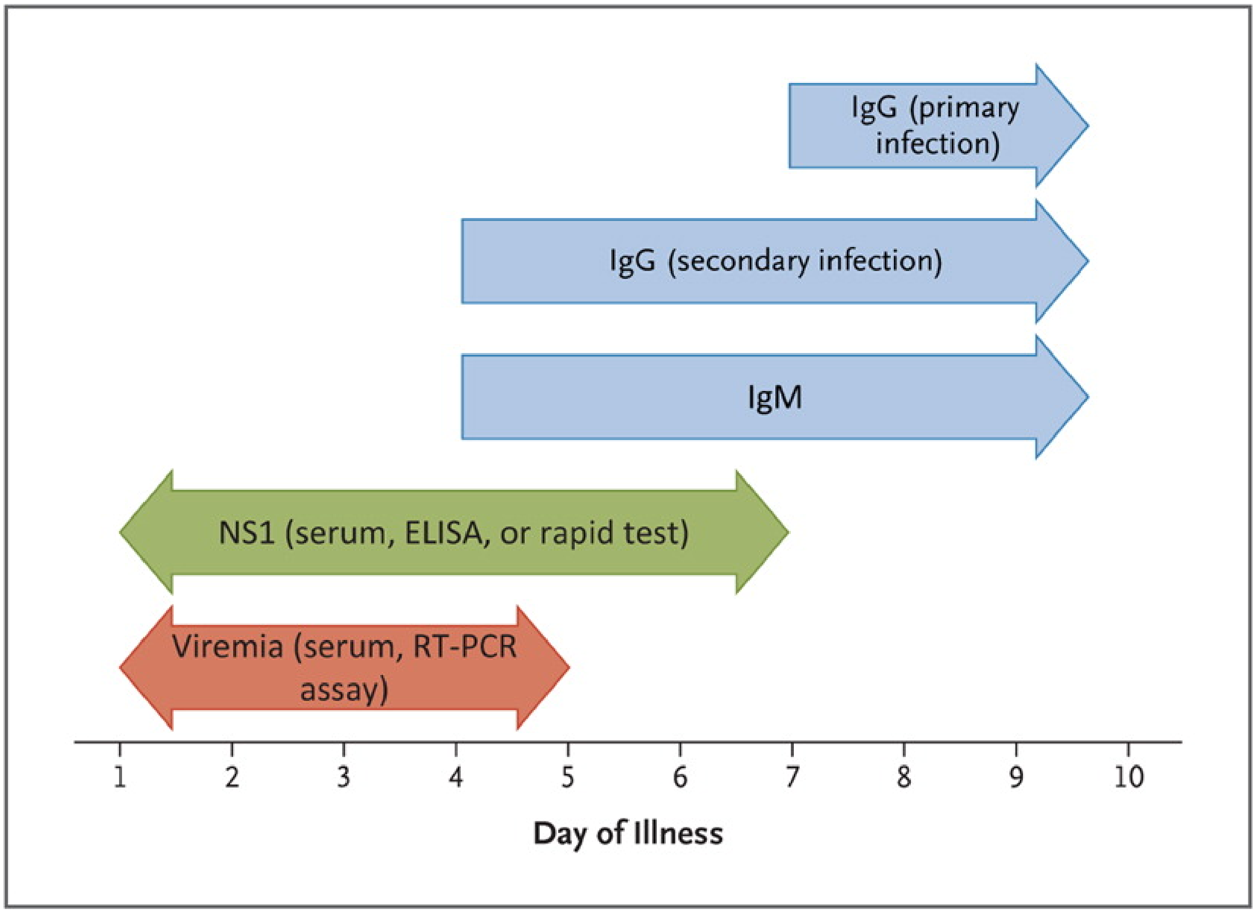
IgG (primary infection)
IgG (secondary infection)
IgM
NS1 (serum, ELISA, or rapid test)
Viremia (serum, RT-PCR assay)
1
2
3
4
5
6
7
8
9
10
Day of Illness

Figure 3: **Laboratory diagnostic options for dengue** (Simmons, 2012 [54])**.** A confirmatory diagnosis can be achieved during the febrile phase via the detection of dengue viral components using RT-PCR, an often unavailable procedure. NS1 detection early in the disease progression will also provide a confirmatory diagnosis, though with reduced sensitivity particularly for secondary infections. IgM or IgG seroconversion between paired acute and convalescent samples is considered a confirmatory finding while detection of IgM or IgG in a single specimen is less conclusive. Patients with secondary infections mount rapid anamnestic antibody responses in which dengue virus–reactive IgG may predominate over IgM.

Chapter 1

Statistical Framework

This section defines terms and establishes the estimation framework that will be used throughout the text.

1.1 Framework for loss-based estimation

Let $X_1, ...X_n$ be independent and identically distributed random variables with probability distribution P_0 such that X_i is the data collected on the ith patient, and is itself a vector of random variables consisting of a J-dimensional set of covariates, W_i, and a binary outcome Y_i, which equals 1 if patient i has diagnosable dengue fever and equals 0 otherwise. The probability distribution P_0 is an element of a statistical model, M, which is a collection of possible probability distributions P of X. That is, $X = (W, Y) \sim P_0 \in M$ and we have n observations of X, denoted by $X_1, ...X_n$.

A parameter Ψ is a map from a probability distribution or a family of probability distributions to the parameter space. Our *true target parameter* of interest, Ψ_0, is a mapping from the true probability distribution, P_0, to the *true target parameter value*, ψ_0. The true target parameter can be expressed as the minimizer of the expected loss over the true data generating distribution:

$$\psi_0 = \Psi_0(W) = \underset{\psi}{\operatorname{argmin}}\, E_{P_0}[L(X, \psi(W))]$$

where the *loss function*, L, is a real-valued function of a parameter value ψ and an observation X. *Risk* is defined as the expected value of $L(X, \psi)$ with respect to probability distribution P:

	Loss function, $L(x, \psi)$
0-1 loss (i.e., misclassification)	$I(\hat{y} \neq y)$
Support vector machine loss (i.e., hinge loss)	$[1 - t\psi]_+$
Absolute loss	$\lvert\psi - y\rvert$
Squared error (i.e., quadratic)	$(\psi - y)^2$
Exponential loss	$exp(-t\psi)$
Binomial negative log likelihood loss (i.e., deviance, cross-entropy)	$-(ylog\psi + (1-y)log(1-\psi))$
"Huberized" square hinge loss	$-4t\psi$ if $t\psi < -1$ $[1 - t\psi]_+^2$ otherwise

Table 1.1: Useful loss functions for an estimate Ψ of a binary outcome Y.

$$\Theta(\psi, P) \equiv \int \text{L(x,}\psi\text{)dP(x)} = E_P[L(X, \psi)],$$

so ψ_0 is the risk minimizer for a given loss function with respect to P_0.

We estimate Ψ_0 with the *estimator* $\hat{\Psi}$, which is a mapping from the empirical distribution, P_n, to our estimate of ψ_0, denoted by ψ_n. In the context of our binary classification problem, ψ_0 and ψ_n will generally represent a vector of the patient risks of disease while y is a vector of the actual patient disease outcomes and $\hat{y}$ is the vector of predicted disease outcome, generated by mapping ψ_n to $\{0, 1\}$, generally using some threshold value (so ψ_0,$\psi_n \in (0, 1)$ while y,$\hat{y} \in \{0, 1\}$). For a binary outcome Y, the conditional expected value, $E[Y|W]$, of Y given covariates W, is the population risk minimizer for the quadratic (i.e., L_2), exponential, hinge, and negative log likelihood loss functions. Other useful loss functions for binary outcomes are listed in Table 1.1. Here, we also introduce the random variable T, defined as $2Y - 1$. Thus, T represents the patient outcome with 1 corresponding to a positive diagnosis and -1 corresponding to a negative diagnosis. (Similarly, t and $\hat{t}$ are rescaled versions of y and $\hat{y}$, respectively.)

It is important to note that although the exponential loss, quadratic loss, and binomial deviance yield the same solution when applied to the population joint distribution, they yield different results for finite data sets. For binary classification, squared error loss is a particularly poor choice, as it penalizes algorithms for assigning very large risks (greater than

1) to observations with positive diagnoses and similarly penalizes the assignment of very small risks to negative diagnoses. It also allows outliers undue influence on its overall fit. The exponential loss, while attractive for computational purposes in the context of additive modeling, also concentrates influence on outliers (e.g., observations with mis-measurements), though to a lesser extent than does the squared error loss. The binomial deviance, support vector loss (to be discussed), and Huberized square hinge loss are all good options for generating an estimator that is robust in noisy settings where the Bayes error rate is not close to zero.

A good estimate of ψ_0 should minimize the expected loss over an independent test sample, also known as the *generalization error*. However, since we do not observe the true data generating distribution, we will have to estimate the generalization error. We will do so using cross validation, described below.

1.2 Cross-validation for risk estimation and estimator selection

Cross-validation consists of dividing the available data (i.e., the *learning* set) into two sets: a *training* set and a *validation* set. The main idea is that the training set is used to fit an estimator while the validation set is used to assess (*validate*) the estimator's performance.

There are several flavors of cross-validation. To describe them, we introduce some additional notation:

- Let B_n denote a split vector, $(B_n(i) : i = 1, ..., n) \in \{0, 1\}^n$, to describe which observations of the learning set belong to the training set, and which belong to the validation set. A realization of this vector assigns a value of 1 to $B_n(i)$ if observation X_i is in the validation set, and a value of 0 to $B_n(i)$ if observation X_i is in the training set.
- Let $P^0_{n,B_n}$and P^1_{n,B_n}denote, respectively, the empirical distributions of the training and validation sets.
- Let $\hat{\Psi}(P_{n,B^0_n})$ denote the estimator of the parameter ψ based only on the training set.
- Let $p_n \equiv \sum_i B_n(i)/n$ denote the proportion of observations in the validation sets.

The cross-validated risk estimator for the estimator $\hat{\Psi}$ is:

$$\hat{\theta}_{p_n,n} \equiv E_{B_n} \Theta(\hat{\Psi}(P^0_{n,B_n}), P^1_{n,B_n})$$

$$= E_{B_n} \int L(x, \hat{\Psi}(P^0_{n,B_n}))dP^1_{n,B_n}(x)$$

$$= E_{B_n} \frac{1}{np_n} \sum_{\{i:B_n(i)=1\}} L(X_i, \hat{\Psi}(P_{n,B^0_n})) \qquad (1.1)$$

The estimator, $\hat{\Psi}$, with the lowest cross-validated risk is called the *cross-validation selector*, denoted by $\hat{k}_{p_n,n}$. Thus, if we consider a collection of K learners $\hat{\Psi}_k, k = 1, ...K$ in the parameter space $\mathbf{\Psi}$, then the cross-validation selector is defined as

$$\hat{k}_{p_n,n} \equiv \underset{k}{\operatorname{argmin}}\, \hat{\theta}_{p_n,n}(\hat{\Psi}_k). \qquad (1.2)$$

The estimator that chooses this cross-validation selector is denoted $\hat{\Psi}_{\hat{k}_{p_n,n}}$. The different flavors of cross-validation differ according to the particular distribution of the split vector B_n. We will use V-fold cross-validation, which means that the learning set is randomly partitioned into V mutually exclusive and exhaustive sets of approximately equal size. Each set is used, in turn, as the validation set while the remaining $V - 1$ folds are used as the training set. The cross-validated risk, then, is the average of the risks calculated using the V validation sets.

Properties of the cross-validated risk

Before discussing the properties of the cross-validated risk, let us introduce some additional terms and notation: For a given loss function, define the *optimal risk* θ as follows:

$$\theta \equiv min_{\psi} \int L(x, \psi)dP_0(x). \qquad (1.3)$$

Define the *conditional risk*, denoted $\tilde{\theta}_n$, as the risk of $\hat{\Psi}(P_n)$ with respect to true distribution P_0. That is:

$$\tilde{\theta}_n \equiv \Theta(\hat{\Psi}(P_n), P_0) = \int L(x, \hat{\Psi}(P_n))dP_0(x). \qquad (1.4)$$

Note that the conditional risk depends on the data, $X_1, ...X_n$ on which ψ_n is based, and is therefore a random parameter. The conditional risk for the estimator mapping $\hat{\Psi}$ applied to the cross-validation sets of size $n(1 - p_n)$ is:

$$\tilde{\theta}_{p_n,n} \equiv E_{B_n} \int L(x, \hat{\Psi}(P^0_{n,B_n}))dP_0(x) \tag{1.5}$$

Dudoit and van der Laan derive the consistency and asymptotic linearity result for the cross-validated estimator $\hat{\theta}_{p_n,n}$ of the conditional risk $\tilde{\theta}_{p_n,n}$ for estimators based on cross-validation [17]. That is, they prove the following:

$$\hat{\theta}_{p_n,n} - \tilde{\theta}_{p_n,n} = \frac{1}{n}\sum_{i=1}^{n}(L(X_i, \Psi(P)) - \theta) + o_P(1/\sqrt{n}) \tag{1.6}$$

This asymptotic linearity result allows us to apply the central limit theorem to derive asymptotic confidence intervals for the conditional risk $\tilde{\theta}_{p_n,n}$. Specifically, we define the influence curve $IC(X|P)$ as $L(X, \Psi(P)) - \theta$ with expectation 0 and variance $\sigma^2 = Var[IC(X|P)] = \int IC^2(x|P)dP(x)$. Then $\sqrt{n}(\hat{\theta}_{p_n,n} - \tilde{\theta}_{p_n,n})/\sigma$ converges in distribution to a standard normal random variable as $n \to \infty$. To obtain an approximation of σ, we can use the following resubstitution estimator:

$$\sigma_n^2 \equiv \int IC^2(x|P_n)dP_n(x) \tag{1.7}$$

where

$$IC(X|P_n) \equiv L(X, \hat{\Psi}(P_n)) - \int L(x, \hat{\Psi}(P_n))dP_n(x) \tag{1.8}$$

Thus, an approximate asymptotic $(1-\alpha)100\%$ confidence interval for the conditional risk is given by

$$\hat{\theta}_{p_n,n} \pm z_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}}, \tag{1.9}$$

where $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$, for the standard normal cumulative distribution function $\Phi()$.

Performance of the cross-validation selector

Define the "oracle" selector $\tilde{k}_n$ as the conditional risk minimizer under the true data-generating distribution, P_0, (i.e., the minimizer of $\tilde{\theta}_n$) among the K candidate learners.

Dudoit and van der Laan (2005 [17]) derive finite sample risk bounds and use them to establish that the cross-validation selector (i.e., the minimizer of $\hat{\theta}_{p_n,n}$) performs asymptotically as well (in terms of risk) as $\tilde{k}_n$. Van der Laan, Dudoit, and Keles (2004) [65] provide stronger convergence results for likelihood-based loss functions, which include the negative log density loss and quadratic loss. These asymptotic results are derived under the assumption that the size of the validation sets converges to infinity and therefore do not cover leave-one-out cross-validation.

More specifically, the finite sample “oracle inequality” and corresponding asymptotic implications from Van der Laan, Dudoit, and Keles (2004) are established by the following theorem:

Theorem 1. Suppose that there exists $\epsilon > 0$ and $L < \infty$ so that $\epsilon < f_k(X|P_n) < L$ a.s. for all $k \in 1, ..., K$. Let $M_1 = 2\log(L/\epsilon)$ and $M_2 = 4L/\epsilon$.

For any $\delta > 0$ we have

$$E\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta \leq (1+2\delta)\left\{E(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta\right\} + 2c(M_1, M_2, \delta)\frac{1+\log(K)}{np_n}$$

where

$$c((M_1, M_2, \delta) = 2(1+\delta)^2\left(\frac{M_1}{3} + \frac{M_2}{\delta}\right).$$

This finite sample result has the following asymptotic implications: If

$$\frac{\log(K)}{(np_n)\{E\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta\}} \to 0 \quad \text{for} \quad \text{n} \to \infty$$

then

$$\frac{E\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta}{E\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta} \to 1 \quad for \quad n \to \infty.$$

Similarly, if

$$\frac{\log(K)}{(np_n)\{E\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta\}} \to 0 \quad \text{in probability for} \quad \text{n} \to \infty$$

then

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta}{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta} \to 1 \quad \text{in probability for} \quad \text{n} \to \infty.$$

Theorem 1 establishes that the cross-validation selector $\hat{k}_{p_n,n}$ performs asymptotically as well as the optimal benchmark selector $\tilde{k}_{p_n,n}$.

Van der Laan, Dudoit and Keles further show, through the following corollary, that if p_n converges to zero when the sample size n converges to infinity, then under given an additional mild condition, it follows that $\hat{k}_{p_n,n}$ also performs asymptotically as well as the benchmark selector, $\tilde{k}_n$.

Corollary 1. Suppose that there exists $\epsilon > 0$ and $L < \infty$ so that $\epsilon < f_k(X|P_n) < L$ a.s. for all $k \in 1, ..., K$.

If $p_n \to 0$ holds, and for $n \to \infty$

$$\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta_{opt}} \to 1 \quad \text{in probability} \tag{1.10}$$

then

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta_{opt}}{\tilde{\theta}_n(\tilde{k}_{p_n,n}) - \theta_{opt}} \to 1 \quad \text{in probability} \tag{1.11}$$

A sufficient condition for 1.10 to hold is that

$$\left(n^{\gamma}\left(\tilde{\theta}_n(\tilde{k}_n) - \theta\right), (n(1-p_n))^{\gamma}\left(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta\right)\right) \overset{D}{\Rightarrow} (Z, Z)$$

for some $\gamma > 0$ and random variable Z with $P(Z > a) = 1$ for some $a > 0$. In particular, if we use single split cross-validation, then it suffices to assume $n^{\gamma}\left(\tilde{\theta}_n(\tilde{k}_n) - \theta\right) \overset{D}{\Rightarrow} Z$ for some $\gamma > 0$ and $P(Z > a) = 1$ for some $a > 0$.

Theorem 1 establishes finite sample bounds for the expected value of the *predictive loss*, i.e., the expected value of the difference between the conditional risks for the cross-validated and oracle selectors, $\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})$, for likelihood-based cross-validation. These bounds imply convergence to zero in expectation and in probability of the predictive loss at rate $O(\log(K)/np_n)$. Theorem 2 of Dudoit and van der Laan (2005) establish that for general loss functions, convergence is achieved at the slower rate $O(\log(K)/\sqrt{np_n})$.

1.3 The super learner

In the above sections, we established oracle results for the cross-validation selector. Now, rather than restricting ourselves to choosing an estimator among our library of candidates, we formulate a weighted combination of our candidates. This estimator is referred to as the super learner and, as shown below, it performs asymptotically as well as the best possible weighted combination of candidate learners [66].

Formulation of the super learner algorithm

In the process of finding the cross-validation selector, we calculated cross-validated predicted values of our $n \times 1$ outcome vector Y for each candidate estimator, k, using the estimator $\hat{\Psi}_k(P_{n,B_n^0})$ to predict values in the corresponding validation set. We denote these cross-validated predicted values using the $n \times K$ matrix Z. We can now formulate a loss-minimization problem as we did in the beginning of this chapter, but here we will use Z in place of the original observed covariates, W. In other words, we wish to estimate a function, denote it by $\tilde{\Psi}$, that gives us estimated values for Y using Z as input; if our loss function is quadratic, then this amounts to estimating $E[Y|Z]$. Our function $\tilde{\Psi}$ could be formulated a number of different ways, from using simple regression methods to using cross-validation with a library of machine learning algorithms to find the cross-validation selector.

In this paper, I limit the functional form of $\tilde{\Psi}$ to $\tilde{\Psi}(Z) = \alpha Z$ where α is a vector of weights, $\alpha = \{\alpha_1, ..., \alpha_K\}$, such that the super learner is simply a convex combination of candidate estimators. We estimate α by minimizing a chosen loss function over our observed data, subject to the convexity constraint:

$$\alpha_n \equiv \underset{\alpha}{\text{argmin}} \sum_{i=1}^{n} L(Y_i - \alpha Z_i) \text{ such that } \sum_{k=1}^{K} \alpha_k = 1. \tag{1.12}$$

Thus, our estimate of $\tilde{\Psi}$ is $\tilde{\Psi}_n(Z) = \alpha_n(Z)$.

Next, we re-train our K candidate estimators using the full data sample, $X_1, ...X_n$, to obtain estimators $\hat{\Psi}_k(P_n)$ for $k = 1, ...K$. The super learner for new data X^* based on the observed data (i.e., P_n) is:

$$\hat{\Psi}(P_n)(X^*) \equiv \tilde{\Psi}_n(\hat{\Psi}_k(P_n)(X^*), k = 1, ..., K) \tag{1.13}$$

While restricting $\tilde{\Psi}$ to a convex combination of the candidate learners is not required, the oracle results for the super learner require a bounded loss function. Taking the convex combination of learners is therefore appealing, as we are guaranteed to have a bounded loss function so long as each algorithm in the library is bounded.

Finite sample result and asymptotics for the super learner

Define the *marginal risk* of the estimator $\hat{\Psi}(P_n)$ as the expectation of its conditional risk, $\tilde{\theta}_n$, where the expectation is with respect to the true data generating distribution, P_0, and thus averages over the possible empirical distributions that can be drawn (recall that $\tilde{\theta}_n$ is affected by P_n through ψ_n's reliance on P_n). The *expected risk difference* of $\hat{\Psi}(P_n)$ is the marginal risk minus the minimal risk:

$$E_{P_0} d_0(\hat{\Psi}(P_n), \psi_0) = E_{P_0} \int (L(X, \hat{\Psi}_n(P_n)) - L(X, \psi_0)) dP_0.$$

Judging performance using the expected risk difference, the super learner performs as well as the oracle selector, up to a typically second order term. More specifically, as long as the number of candidate learners is polynomial in sample size, the super learner has the following properties [66]:

- In the typical situation in which the library of candidate learners does not contain a correctly specified parametric model, then the super learner performs asymptotically as well (up till the constant) as the oracle selector in terms of the risk difference.
- If a correctly specified parametric model is among the candidate learners, then the super learner converges at rate $O(\frac{log(n)}{n})$.

These results are a direct consequence of theorem 3.1 from van der Laan, Dudoit, and van der Vaart (2006) [31], which establishes the oracle inequality for the cross-validation selector. Polley, van der Laan, and Hubbard (2007) extend this oracle inequality to the super learner with the following theorem [66], which uses the squared error loss function. Essentially, the proof works by conceptualizing the selection of α analogously to the selection of any other candidate estimator; by allowing the super learner to be an convex combination of the candidates in our original library, we effectively just enlarge the library of candidates under consideration and preserve the finite sample and asymptotic results of the cross-validation selector.

Theorem 2. Assume $P((Y,X) \in \mathcal{Y}\times\mathcal{X}) = 1$, where $\mathcal{Y}$ is a bounded set in $\mathbb{R}$, and $\mathcal{X}$ is a bounded Euclidean set. Assume that the candidate estimators map into $\mathcal{Y}: \ P(\hat{\Psi}_j(P_n) \in \mathcal{Y},\, j = 1, ..., J) = 1$.

Let $v \in 1, ...V$ index a sample split into a validation sample $V(v) \subset \{1, ..., n\}$ and corresponding training sample $T(v) \subset 1, ..., n$ (complement of $V(v)$), where $V(v) \cup T(v) = \{1, ..., n\}$, and $\cup_{v=1}^{V} V(v) = \{1, ...n\}$. For each $v \in \{1, ...V\}$, let $\psi_{njv} \equiv \hat{\Psi}_j(P_{nT(v)}),\ \mathcal{X} \to \mathcal{Y}$, be the realization of the jth estimator $\hat{\Psi}_j$ when applied to the training sample $T(v)$.

For an observation i let $v(i)$ be the validation sample observation i belongs to, $i = 1, ...n$. Construct a new data set of n observations defined as: (Y_i, Z_i), where $Z_i \equiv (\psi_{njv(i)}(X_i)\ :\ j = 1, ..., J) \in \mathcal{Y}^J$ is the J-dimensional vector consisting of the J predicted values according to the J estimators trained on the training sample $T(v(i))$, $i = 1, ...n$.

Consider a regression model $z \to m(z|\alpha)$ for $E(Y|Z)$ indexed by a $\alpha \in \mathcal{A}$ representing a set of functions from $\mathcal{Y}^J$ into $\mathcal{Y}$. Consider a grid (or any finite subset)

$\mathcal{A}_n$ of α-values in the parameter space $\mathcal{A}$. Let $K_n = |\mathcal{A}_n|$ be the number of grid points which grows at most at a polynomial rate in n : $K_n < n^q$ for some $q < \infty$. Let

$$\alpha_n \equiv \underset{\alpha \in \mathcal{A}_n}{\operatorname{argmin}} \sum_{n=1}^{n} (Y_i - m(Z_i|\alpha))^2.$$

Consider the regression estimator $\psi_n : \mathcal{X} \to \mathcal{Y}$ defined as

$$\psi_n(x) \equiv m((\psi_{jn}(x) : j = 1, ..., J)|\alpha_n).$$

For each $\alpha \in \mathcal{A}$, define the candidate estimator $\hat{\Psi}_\alpha(P_n) \equiv m((\hat{\Psi}_j(P_n) : j = 1, ..., J)|\alpha)$: i.e.,

$$\hat{\Psi}_\alpha(P_n)(x) = m((\hat{\Psi}_j(P_n)(x) : j = 1, ...J)|\alpha).$$

Consider the oracle selector of α:

$$\tilde{\alpha}_n \equiv \underset{\alpha \in \mathcal{A}_n}{\operatorname{argmin}} \frac{1}{V} \sum_{v=1}^{V} d(\hat{\Psi}_\alpha(P_{nT(v)}), \psi_0),$$

where

$$d(\psi, \psi_0) = E_0(L(X, \psi) - L(X, \psi_0)) = E_0(\psi(X) - \psi_0(X))^2.$$

For each $\delta > 0$ we have that there exists a $C(\delta) > \infty$ such that

$$\frac{1}{V} \sum_{v=1}^{V} E d(\hat{\Psi}_{\alpha_n}(P_{nT(v)}), \psi_0) \leq (1+\delta) E \min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^{V} d(\hat{\Psi}_\alpha(P_{nT(v)}), \psi_0) + C(\delta) \frac{V \log n}{n}.$$

Thus, if

$$\frac{E \min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^{V} d(\hat{\Psi}_\alpha(P_{nT(v)}), \psi_0)}{\frac{\log n}{n}} \to 0 \text{ as } n \to \infty, \tag{1.14}$$

then it follows that the estimator $\hat{\Psi}_{\alpha_n}$ is asymptotically equivalent with the oracle estimator $\hat{\Psi}_{\tilde{\alpha}_n}$ when applied to samples of size $(1 - 1/V)n$:

$$\frac{\frac{1}{V}\sum_{v=1}^{V} Ed(\hat{\Psi}_{\alpha_n}(P_{nT(v)}), \psi_0)}{E\min_{\alpha \in \mathcal{A}_n} \frac{1}{V}\sum_{v=1}^{V} d(\hat{\Psi}_{\alpha}(P_{nT(v)}), \psi_0)} \to 1 \text{ as } n \to \infty.$$

If 1.14 does not hold, then it follows that $\hat{\Psi}_{\alpha_n}$ achieves the $(\log n)/n$ rate:

$$\frac{1}{V}\sum_{v=1}^{V} Ed(\hat{\Psi}_{\alpha_n}(P_{nT(v)}), \psi_0) = O\left(\frac{\log n}{n}\right).$$

The discrete approximation $\mathcal{A}_n$of $\mathcal{A}$ used in the above theorem can be chosen such that minimizing over $\mathcal{A}_n$ results in an asymptotically equivalent procedure to minimizing over the whole set $\mathcal{A}$. For example, if α is a Euclidean parameter and $||m(\cdot|\alpha_1) - m(\cdot|\alpha_2)||_\infty < C||\alpha_1 - \alpha_2||$ for some $C < \infty$, where $||\cdot||_\infty$ denotes the supremum norm, then it follows that for each $\delta > 0$ we have that there exists a $C(\delta) < \infty$ such that

$$\frac{1}{V}\sum_{v=1}^{V} Ed(\hat{\Psi}_{\alpha_n}(P_{nT(v)}), \psi_0) \leq (1+\delta)E\min_{\alpha \in \mathcal{A}} \frac{1}{V}\sum_{v=1}^{V} d(\hat{\Psi}_{\alpha}(P_{nT(v)}), \psi_0) + C(\delta)\frac{\log n}{n},$$

where $\alpha_n = \text{argmin}_{\alpha \in \mathcal{A}} \sum_{n=1}^{n}(Y_i - m(Z_i|\alpha))^2$. The other conclusions of the theorem 2 now also apply [66].

By the argument presented in van der Vaart, Dudoit, and van der Laan (2006) [64], it follows that by letting the number of cross-validation folds, V, converge to infinity at a slow enough rate relative to n, then $\hat{\Psi}_{\alpha_n}(P_n)$ performs asymptotically as well (up till a constant) as the oracle applied to the full data sample, or it achieves the parametric rate of convergence up till the $\log n$ factor.

1.4 Assessing prediction accuracy

The previous sections describe the construction of a so-called super learner to obtain predicted values for an outcome of interest. This section is concerned with measures of the super learner's accuracy and the construction of confidence intervals around these measures of prediction accuracy. Properties of the super learner and of cross-validated risk estimators, established above, will prove critical for confidence interval construction.

1.4.1 Measures of prediction accuracy

We will use cross-validation on the super learner itself to estimate super learner's performance using various measures for binary outcomes, including the following:

1. Area under the ROC curve (i.e., AUC)[1]: $P(\hat{\Psi}(w_i) > \hat{\Psi}(w_h)|y_i = 1, y_h = 0)$

2. Sensitivity (i.e., true positive rate): $P(I(\hat{\Psi}(w_i) > c)|y_i = 1)$

3. Specificity (i.e., 1-false positive rate): $P(I(\hat{\Psi}(w_i) \leq c)|y_i = 0)$

4. Error rate (i.e., misclassification rate): $P(I(\hat{\Psi}(w_i) > c) \neq y_i)$

5. Positive predictive value (PPV): $P(y_i = 1|I(\hat{\Psi}(w_i) > c))$

6. Negative predictive value (NPV): $P(y_i = 0|I(\hat{\Psi}(w_i) \leq c))$

7. Brier score (i.e., risk under quadratic loss): $E(\hat{\Psi}(w_i) - y_i)^2)$

Note that the sensitivity, specificity, error rate, positive predictive value and negative predictive value all depend on a threshold, c, to classify observations as 0 or 1 according to the predicted risk of having a positive outcome. That is, if $\hat{\Psi}(w_i)$ is greater than c, then observation i is predicted to have a positive diagnosis, while if $\hat{\Psi}(w_i)$ is less than c, then the diagnosis is categorized as negative (i.e., our predicted value for y_i is 0). The chosen value for c will depend on one's preference for balancing sensitivity with specificity, and the estimated error rate will depend on both the value of c and the fraction of the validation data that has a positive diagnosis.[2] While the sensitivity and specificity of diagnostic tests are the most widely reported, the positive predictive value (PPV) and negative predictive value (NPV) are more useful when the goal is to inform medical decision-making: patients are generally most interested in the probability that they have dengue, given their test result.

The area under the ROC curve (AUC) is an appealing performance measure, as it does not require a specified diagnostic threshold, but indicates how well the estimator separates those with a negative outcome from those with a positive outcome. If all of the cases have higher predicted risk than the non-cases, then the AUC, which is equivalent to the Mann-Whitney U statistic, will reach its maximum value of 1, indicating perfect discrimination. Note, though, that AUC does not capture how well a model is *calibrated*. That is, AUC only

[1] ROC stands for receiver operating characteristic. The ROC curve is produced by plotting the false positive rate on the x-axis and the true positive rate on the y-axis.

[2] If it is indeed worse to miss a true positive than it is to miss a true negative (or vice versa), then it would be best to use a loss function to reflect such a preference when building your prediction algorithm. This can be done by giving a larger weight to the dengue positive patients (or dengue negative patients).

cares whether patients are properly *ranked*, not whether their predicted risks are accurate beyond their ranking. In fact, all the other measures listed aside form the Bier score have this weakness. Thus, we will also report the Brier score as an indication of how well our models assign patient risk, particularly when models perform similarly by other measures [11].

1.4.2 Confidence intervals for prediction accuracy estimates

In many estimation situations, we can quantify the uncertainty of our performance estimates using (a) the non-parametric bootstrap and (b) influence-curve based methods. However, because the super learner uses cross-validation to estimate Ψ_0, the non-parametric bootstrap will give incorrect inference. This is because the bootstrap involves re-sampling with replacement to create each so-called bootstrap sample. Most bootstrap samples, then, will contain some repeated observations. Thus, for most draws of B_n, the training set will contain some of the same observations as the validation set. As a result, the learners which over-fit the data will be unfairly favored, as their cross-validated risk will be overly optimistic and our confidence bounds for our risk estimate will thus also be optimistic.

Meanwhile, as shown in the preceding section, the cross-validated risk estimator is consistent and asymptotically linear for the conditional risk, so we can obtain asymptotic confidence intervals if we can derive the estimator's influence curve. Here, we illustrate this process using cross-validated AUC as the target parameter. These results were established by LeDell, Peterson, and van der Laan [32].

As introduced in section 1.4.1, the area under the ROC curve can be defined as follows: $P(\hat{\Psi}(w_i) > \hat{\Psi}(w_h) | y_i = 1, y_h = 0)$

Or, equivalently

$$AUC(P_0, \psi) = P_0(\psi(W_1) > \psi(W_2) | Y_1 = 1, Y_2 = 0) \tag{1.15}$$

where (W_1, Y_1) and (W_2, Y_2) are i.i.d. samples from P_0. The empirical AUC is:

$$\begin{aligned} AUC(P_n, \psi) &= \frac{1}{n_0 n_1} \sum_{i=1}^{n} \sum_{h=1}^{n} I(\psi(W_h) > \psi(W_i)) I(Y_i = 0, Y_h = 1) && (1.16) \\ &= \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{h=1}^{n_1} I(\psi(W_h) > \psi(W_i)) && (1.17) \end{aligned}$$

where n_0, n_1 are the number of observations with $Y = 0$ and $Y = 1$, respectively.

We will assume V-fold cross-validation for notational simplicity, though other types of cross-validation will work as well (so long as the size of the validation set converges to infinity). Drawing from the notation introduced previously, we assume that the split vector, B_n, places mass $1/V$ on each of V random vectors, $B_n^v = (B_n^v(i) : i = 1, ..., n)$, $v = 1, ..., V$ such that $\sum_i B_n^v(i) \approx \frac{n}{V} \forall v$ and $\sum_v B_n^v(i) = 1 \forall i$. For each B_n^v, we define $\psi_{B_n^v} = \hat{\Psi}(P^0_{n,B_n^v})$, where P^0_{n,B_n^v} is the empirical distribution of the observations contained in the v^{th} training set. The "true" cross-validated AUC, denoted $\tilde{R}(\hat{\Psi}, P_n)$, is thus

$$\begin{aligned}
\tilde{R}(\hat{\Psi}, P_n) &= E_{B_n} AUC(P_0, \psi_{B_n}) \\
&= \frac{1}{V} \sum_{v=1}^{V} AUC(P_0, \psi_{B_n^v}) \\
&= \frac{1}{V} \sum_{v=1}^{V} P_0(\psi_{B_n^v}(W_1) > \psi_{B_n^v}(W_2) | Y_1 = 1, Y_2 = 0) \qquad (1.18)
\end{aligned}$$

where (W_1, Y_1) and (W_2, Y_2) are i.i.d. samples from P_0. This is our target parameter, which we will estimate with the V-fold cross-validated AUC estimator, denoted by $\hat{R}(\hat{\Psi}, P_n)$ and defined as:

$$\begin{aligned}
\hat{R}(\hat{\Psi}, P_n) &= E_{B_n} AUC(P^1_{n,B_n}, \psi_{B_n}) \\
&= \frac{1}{V} \sum_{v=1}^{V} AUC(P^1_{n,B_n^v}, \psi_{B_n^v}) \\
&= \frac{1}{V} \sum_{v=1}^{V} \frac{1}{n_0^v n_1^v} \sum_{i=1}^{n} \sum_{h=1}^{n} I(\psi_{B_n^v}(W_h) > \psi_{B_n^v}(W_i)) \times I(Y_i = 0, Y_h = 1) I(B_n^v(i) = B_n^v(h) = 1) \\
& \qquad (1.19)
\end{aligned}$$

where n_1^v and n_0^v are the number of positive and negative samples in the v^{th} validation fold, respectively. So, whereas the "true" cross-validated AUC evaluates the performance of each $\psi_{B_n^v}$ over the true probability distribution P_0, our estimate of the cross-validated AUC evaluates the performance of $\psi_{B_n^v}$ over P^1_{n,B_n^v} – the empirical validation set of the cross-validation fold v.

LeDell, Peterson, and van der Laan show that the influence curve for $AUC(P^1_{n,B_n^v}, \psi_{B_n^v})$ is

$$
\begin{aligned}
IC_{AUC}(P^1_{n,B^v_n}, \psi_{B^v_n})(X_i) = {} & \frac{I(Y_i=1)}{P_n(Y=1)} P^1_{n,B^v_n}(\psi_{B^v_n}(W) < x|Y=0)|_{x=\psi_{B^v_n(W_i)}} \\
& + \frac{I(Y_i=0)}{P_n(Y=0)} P^1_{n,B^v_n}(\psi_{B^v_n}(W) > x|Y=1)|_{x=\psi_{B^v_n(W_i)}} \\
& - \left\{ \frac{I(Y_i=1)}{P_n(Y=1)} + \frac{I(Y_i=0)}{P_n(Y=0)} \right\} AUC(P^1_{n,B^v_n}, \psi_{B^v_n}) \qquad (1.20)
\end{aligned}
$$

Therefore, an asymptotic .95 confidence interval for $\tilde{R}(\hat{\Psi}, P_n)$ is $\hat{R}(\hat{\Psi}, P_n) \pm 1.96\frac{\sigma_n}{\sqrt{n}}$ where

$$
\begin{aligned}
\sigma^2_n &= E_{B_n} P^1_{n,B_n} \left\{ IC_{AUC}\left(P^1_{n,B_n}, \psi_{B_n}\right) \right\}^2 \qquad (1.21) \\
&= \frac{1}{V} \sum_{v=1}^{V} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[IC_{AUC}(P^1_{n,B^v_n}, \psi_{B^v_n})(X_i) \right]^2 I\left(B^v_n(i) = 1\right) \right\}
\end{aligned}
$$

Note that equation 1.20 follows from 1.8 and 1.21 follows from equation 1.7.

1.5 Candidate Estimators

As discussed, the super learner requires a finite collection of candidate estimators for the parameter of interest. In this section, we provide brief descriptions of some of the candidate estimators that we will use as input into the super learner. These include classification trees, penalty-based estimators, support vector machines, and neural networks.[3] Note that many of these so-called "machine-learning" algorithms themselves strike a balance between variance and bias by tuning a "complexity" parameter via cross-validation.

Nearest-neighbor classifiers

One of the most conceptually simple prediction methods is known as the *k-nearest-neighbor* method. To implement it, a distance measure is required, often chosen to be the euclidean distance. We then calculate the distance between each observation in the test set and each observation in the training set, and give the test set observation the average of the outcome values from its k nearest neighbors in the training set. The value of k is a complexity parameter that can be fit using cross-validation.

[3] For more information on these methods, please see Hastie, Tibshirani and Friedman's excellent text, *The Elements of Statistical Learning* [21].

The *nearest centroid classifier* also relies on a distance measure. Here, we simply calculate the centroid of each class using the training data, where the centroid consists of the mean values of each covariate. We then label each test observation according to the class of the centroid to which it is closest.

Classification trees

Tree-based prediction methods segment the feature space into M rectangles, $R_1, ..., R_M$, and fit a simple prediction model in each rectangle (e.g., assign a constant value, c_m, to observations in rectangle R_m). In the case of our binary prediction problem, c_m would simply be the value of the majority class (0 or 1) in rectangle m. Three popular tree-based methods are called CART, CHAID, and C4.5, though the term Classification and Regression Trees (CART) is sometimes used more broadly refer to all three of these tree-based methods. We will first describe CART (using the more restrictive definition of the term), and will then describe CHAID and C4.5 by contrasting them to CART.

CART grows prediction trees using recursive binary splitting and then prunes them using a complexity penalty chosen through cross-validation. To grow the tree, we first split the feature space into two regions using a splitting variable j (i.e., the jth column of covariate matrix W), and a split value s (chosen among all values contained in the jth column of W.) We choose j and s by searching across all possible values for them and selecting the j, s combination that produces the largest drop in some measure of tree impurity. *Node* impurity can be defined using any of the following measures, where p is the proportion of observations in node m that have a positive diagnosis.

Misclassification error: $1 - max(p, 1 - p)$
Gini index: $2p(1 - p)$
Cross-entropy or deviance: $-(p \log p) - (1 - p) \log(1 - p)$

Tree impurity is then the sum of the impurities of its nodes, weighted by the number of observations contained in each node.[4] Next, one or both of the regions created by the first split are themselves split into two more regions. This process is continued until the criteria of some stopping rule is met, such as when each node reaches a minimum size.

Once a tree is fully grown, it is time to prune it. We do so by successively collapsing the internal node that produces the smallest increase in tree impurity, and continue until all branches have been stripped. The full tree and each subtree produced during the pruning process will now be analyzed using cross-validation and the subtree which minimizes the

[4] The Gini index and cross-entropy are preferable to the misclassification rate for growing trees, as they better prioritize the production of pure nodes [21].

cross-validated risk (for example, the misclassification rate) is the final CART tree.

The main differences between CHAID and CART are: (1) CHAID uses a p-value for from a significance test to determine the desirability of a split (while CART uses the reduction of an impurity measure), (2) CHAID searches for multiway splits (while CART splits each node into just two groups at each stage), and (3) CHAID uses a statistical stopping rule (while CART grows a full tree and then prunes it).

The main differences between C4.5 and CART are: (1) C4.5 searches for multiway splits (while CART splits each node into just two groups at each stage), and (2) C4.5 prunes trees using a single-pass algorithm based on the binomial distribution (while CART bases its pruning decisions using a holdout dataset).

Random forests

Bagging, also known as bootstrap aggregation, is a technique used to reduce the variance of an estimated prediction function by averaging predictions over a collection of bootstrap samples. (For classification, a *committee* of trees each cast a vote for the predicted class.) Random forests, developed by Leo Breiman [7], uses this technique of averaging over bootstrap samples with regression trees, which are grown until the minimum node size is reached (and are not pruned). But, importantly, as the trees are grown, only a random subset of predictors are considered for each split. This reduces the correlation between trees and thus improves the variance reduction relative to the variance reduction achieved with bagging, making random forests considerably more successful than bagging trees in most data situations. Tuning parameters for random forests include (1) the fraction of variables selected for consideration at each split, (2) the number of trees grown, and (3) the minimum node size. Note that if the number of relevant variables is small relative to the total number of variables, then one should choose a reasonably large number for parameter (1). This consideration, though, must be balanced by the reduction in correlation between trees that comes from choosing a small value for this parameter.

Gradient boosted models (GBM)

Boosting, like bagging, involves combining the outputs from many "weak" classifiers into a committee (or *ensemble*). Boosting, though, differs from bagging in some important ways that ultimately make boosting a generally more effective technique for prediction. Essentially, boosting involves fitting a simple classification algorithm[5] (referred to as the *base* classifier)

[5] Boosting is not limited to classification problems, though classification is our focus.

to repeatedly modified versions of the data. Specifically, in each iteration, the observations that were misclassified by the classifier in the previous step are given greater weight, leading to a classifier in the current step that prioritizes classifying these observations correctly. The final prediction algorithm is a weighted combination of the fitted classifiers from each iteration, where classifier weights are a function of the classifier's error (based on the weighted data used to fit each classifier), with more accurate algorithms receiving greater weight.

Boosting can be applied using various base classifiers and loss functions but is most often used with trees. When the base classifier is a tree, the size of each tree becomes a meta-parameter for the entire boosting procedure that can be estimated to optimize performance, or chosen based on subject knowledge: note that the number of splits limits the number of interaction effects contained in the model, with a single split stump covering main effects only. Another meta-parameter is the number of boosting iterations. This is a complexity parameter that, if set too large, can lead to over-fitting.

Penalized regression

Penalized regressions can improve prediction accuracy by reducing the variance of regression prediction estimates by shrinking or setting some regression coefficients to zero. As discussed, more "complex" methods tend to over-fit the data (high variance, low bias), while methods that are too smooth will have reduced variance at the cost of increased bias. Here, we will see how with penalized regressions we can tune a complexity parameter to strike an optimal balance between variance and bias (within the regression framework) based on the cross-validation prediction error.

Consider maximizing the log-likelihood of our data (or minimizing some loss function), subject to a constraint on the size of the regression coefficients. Or, equivalently, our objective is to maximize an objective function that is the log-likelihood minus a penalty where the penalty grows with the size of the regression coefficients. While we can apply such penalties to any linear regression model, we will illustrate their use with the logistic regression model

$$log\left(\frac{p(Y=1|W=w)}{1-p(Y=1|W=w)}\right)=\beta_0+\beta^T w \tag{1.22}$$

where β_0 is the intercept parameter and β is a J-dimensional vector of parameters to estimate. The log-likelihood of this logistic model can be written

$$\ell(\beta)=\sum_{i=1}^{N}[y_i \log p(Y=1|W=w_i;\beta_0,\beta)+(1-y_i)\log(1-p(Y=1|W=w_i;\beta_0,\beta]$$

$$\ell(\beta) = \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T w_i) - \log(1 + e^{\beta_0 + \beta^T w_i}) \right]. \tag{1.23}$$

In regular logistic regression, we would estimate β_0 and β by finding the values of these parameters that maximize this likelihood equation. (Note that if $J < N$, then there is no unique solution and regularization is not only a good idea for variance reduction purposes, but necessary for estimation purposes.)

For the *lasso*, we instead solve

$$\underset{\beta_0,\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T w_i) - \log(1 + e^{\beta_0 + \beta^T w_i}) \right] - \lambda \sum_{j=1}^{J} |\beta_j| \right\} \tag{1.24}$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of *shrinkage*, which we can choose to minimize estimated prediction error using cross-validation.

We can generalize the lasso using the generic "L_q" penalty

$$\underset{\beta_0,\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T w_i) - \log(1 + e^{\beta_0 + \beta^T w_i}) \right] - \lambda \sum_{j=1}^{J} |\beta_j|^q \right\} \tag{1.25}$$

where $q = 0$ corresponds to regression using the "best subset" of predictors, as the penalty simply counts the number of nonzero parameters: $q = 1$ corresponds to the lasso, and $q = 2$ corresponds to *ridge* regression. Note that $q = 1$ is the smallest q such that the constraint region is convex; using $q < 1$ makes the optimization problem much more difficult (see chapter on methods for "best subset selection", below). The L_1 penalty of the lasso forces coefficients to be zero for large values of λ, whereas ridge regression's L_2 penalty merely shrinks the size of the regression coefficients towards zero for large values of λ without eliminating predictors from the final prediction algorithm.

Rather than choosing between the lasso and ridge, we can incorporate both an L_1 and L_2 penalty, fitting an additional parameter that determines the relative weights given to each. This results in the so-called *elastic-net* regression, introduced by Zou and Hastie in 2005 [79]. The elastic-net selects variables like the lasso while shrinking together the coefficients of correlated predictors like ridge regression.

Discriminant analysis and nearest shrunken centroids

With a binary outcome, *linear discriminant analysis* (LDA) models the log-posterior odds between class 0 and class 1 as a linear function of w, just as we do for logistic regression (equation 1.22). However, the coefficients are estimated differently, resulting in different (though often similar) results. Specifically, LDA is fit by maximizing the full log-likelihood based on the joint density $Pr(W, Y = y)$, and relies on the assumption that the marginal distributions of W within each class are multivariate Gaussian with a common covariance matrix. In contrast, the logistic regression model leaves the marginal density of W as an arbitrary density function while fitting the parameters of $Pr(Y|W)$ by maximizing the conditional likelihood $P(Y = y|W)$. Thus, logistic regression has the appeal of relying on fewer assumptions and being more robust to outliers. However, if the assumptions of LDA are true, then it can estimate the parameters more efficiently (lower variance).

Relaxing LDA's assumption of a common covariance yields *quadratic discriminant analysis* (QDA). Since QDA needs to estimate separate covariance matrices for each class, it is more computationally intensive, with a dramatic increase in the number of parameters when J is large.

Regularization offers a compromise between LDA and QDA by shrinking the separate covariances of QDA toward a common covariance as in LDA. As in the case of logistic regression, regularization can overcome the singularity issue that LDA suffers from when $J > N$. Regularized discriminant analysis (RDA) can take a variety of forms, depending on the type of covariance shrinking and whether the centroids are also shrunk.

The simplest form of regularization assumes that the features are independent within each class (i.e., the within-class covariance matrix is diagonal). This method, known as diagonal LDA, is equivalent to the nearest centroid classifier (with appropriate standardization). By additionally shrinking the classwise mean toward the overall mean, for each feature separately, we automatically drop out features not contributing to the class predictions. This procedure, known as *nearest shrunken centroids* (NSC), is thus a powerful tool for classification as well as a valuable tool for variable selection.

Support vector machines

A support vector machine (SVM) produces nonlinear boundaries between two (or more) classes by constructing a linear boundary in a large, transformed version of the feature space (using basis expansions such as polynomials or splines). It does so by solving the following optimization problem:

$$argmin_{\beta_0,\beta} \sum_{i=1}^{N} [1 - y_i[h(w_i)^T \beta + \beta_0]_+ + \frac{\lambda}{2}||\beta||^2. \tag{1.26}$$

As equation 1.26 reveals, SVMs use a hinge loss function and a 2-norm regularizer. The hinge loss is an appealing loss function, as it penalizes more for estimates that generate an incorrect diagnosis, but also penalizes if the diagnosis is correct, but not by a large margin. SVMs have the potential to achieve excellent predictive power, though can have difficulty dealing with a large number of irrelevant inputs.

Neural networks

Neural networks extract linear combinations of the predictors as derived features, and then model the outcome as a nonlinear function of these features. Here we describe the most widely used neural net, known as the "single layer perceptron", as described in Hastie, Tibshironi, and Friedman (2009) [21]. Neural networks will be covered more extensively in Chapter 3's section on deep learning.

Neural networks apply to regression and to classification. For the sake of generality, we describe the method for a K-class classification problem, illustrated in Figure 1.1, where for regression we can set $K = 1$. For classification, each of the K output units $Y_1, ...Y_K$ represents the probability of class k. Derived features Z_m and the target Y_k are modeled as follows:

$$Z_m = \sigma(\sigma_{0m} + \alpha_m^T W), m = 1, ..., M \tag{1.27}$$

$$T_k = \beta_{0k} + \beta_k^T Z, k = 1, ..., K \tag{1.28}$$

$$Y_k = g_k(T), k = 1, ...K, \tag{1.29}$$

where M is the number of so-called *hidden neurons* or *units*, Z represents the vector $(Z_1, Z_2, ..., Z_M)$, T is the vector $(T_1, T_2, ..., T_K)$ and $\sigma(v)$ is the activation function, commonly chosen to be the sigmoid $\sigma(v) = 1/(1 + e^{-v})$. By including a constant "1" as an additional input feature in each layer, we capture the intercepts α_{0m} and β_{0k}. A final transformation of T is facilitated though the output function $g_k(T)$, typically chosen as the softmax function for K-class classification

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}$$

which guarantees positive estimates that sum to one. This is the same transformation function used in multilogit regression.

Thinking of the Z_m (also called *hidden units*) as a basis expansion of the original inputs, the neural network is then simply a standard linear model, or linear multilogit model, using

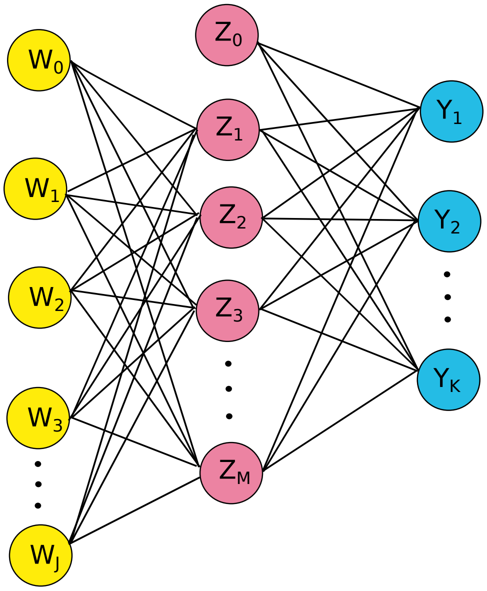

Figure 1.1: Schematic of a single hidden layer, feed-forward neural network with input layer in yellow, hidden layer in pink, and output layer in blue. W_0 and Z_0 represent bias units, corresponding to the intercept terms.

Z as input. The important twist is that the parameters of the basis functions (which we also refer to as *weights*), and potentially the number of hidden units, are learned from the data.

To fit a neural network with a fixed M, we must estimate the weights $\{\alpha_{0m,}\alpha_m;\ m = 1, 2, ..., M\}$ and $\{\beta_{0k}, \beta_k;\ k = 1, 2, ...K\}$, which contain $M(1 + J)$ and $K(M + 1)$ elements, respectively. We can estimate these weights by minimizing a chosen loss function over the training data. This can be done using gradient descent (called back-propagation in this setting), though other methods are faster. Of course, we know that minimizing our loss function over the training data leads to over-fitting. Regularization can be employed to combat over-fitting through the introduction of a penalty term to our objective function. For example, rather than minimizing the training set's empirical loss, we can minimize this loss plus $\lambda\left[\sum_{km} \beta_{km}^2 + \sum_{ml} \sigma_{ml}^2\right]$, where the tuning parameter λ is determined via cross-validation. Complicating the minimization problem is that the error function is non-convex, with different starting values leading to different local minima. One option is to take the solution that yields the lowest (penalized) error after trying a variety of starting configurations; a potentially better approach is to use the average predictions over the collections of networks.

It is most common to choose a reasonably large number for M and allow regularization to shrink hidden layer weights to avoid over-fitting. M is typically chosen somewhere in the range of 5 to 100. Of course, we can once again use cross-validation to select a value of M

among a set of options. Since we have two free parameters to select (M and λ), we would search over the unique combinations of M and λ, selecting the combination that minimizes the cross-validation test error.

Single-layer neural nets, as described above, are extraordinarily flexible.[6] In practice, however, adding layers often improves performance by increasing the ease at which nonlinear terms are captured. Neural nets with multiple layers are discussed in the section on Deep Learning in Chapter 3.

1.6 Summary of estimation procedure

In this chapter, we have described a procedure for estimating an ensemble learner and evaluating its performance. The procedure consists of the following steps:

- Step 1: Choose a full data loss function to represent the desired measure of performance and define the parameter of interest as the risk minimizer for this loss function. Map the full data loss function into an observed data loss function having the same expected value and leading to an efficient estimator of this risk. (See van der Laan and Robins [67] for details).
- Step 2: Create a finite collection of candidate estimators for the parameter of interest.
- Step 3: Use cross-validation to assess the performance of each candidate estimator.
- Step 4: Find the convex combination of the candidate estimators that minimizes cross-validated risk.
- Step 5: Fit each candidate learner to the full data. The super learner is the convex combination (established in step 4) of these learners.
- Step 6: Assess prediction accuracy of the super learner.

In the process of establishing our estimation framework, we have introduced the notation summarized in Table 1.2.

[6] In fact, single-layer neural nets can approximate any function arbitrarily well. That is, given some desired level of accuracy $\eta > 0$ and an arbitrary function $f(w)$, we can find a neural network with output $g(w)$ such that $|g(w) - f(w)| < \eta$ for all inputs w by choosing large enough M. (This universality theorem was proven in Cybenko, 1989 [13] with other groups proving closely related results.)

Symbol	Definition
X	Data matrix ($X = (W, Y)$)
W	Vector of J predictor variables (i.e., covariates)
Y	Binary outcome variable (i.e., diagnosis), coded $0, 1$
T	Binary outcome variable (i.e., diagnosis), coded $-1, +1$
P_0	True data distribution ($X \sim P_0$)
P_n	Empirical data distribution
Ψ_0	True target parameter (mapping from P_0 to ψ_0)
ψ_0	Value of true target parameter (in context of prediction, ψ_0 is a function while $\psi_0(x_i) \in (0, 1)$ is a number indicating patient i's risk of disease).
$\hat{\Psi}$	Estimator of the true target parameter (mapping from P_n to ψ_n)
ψ_n	Estimate of ψ_0 (in context of prediction, ψ_n is a function).
$\hat{y}$	Predicted diagnosis, $\hat{y} = I(\psi_n(x) > c) \in 0, 1$ for some constant c.
$\hat{t}$	Predicted diagnosis, $\hat{t} = 2\hat{y} - 1 \in -1, +1$
θ	Optimal risk (i.e., risk of ψ_0 with respect to P_0)
$\tilde{\theta}_n$	Conditional risk (i.e., risk of ψ_n with respect to P_0, where ψ_n is based on all n observations in training data).
$\tilde{\theta}_{p_n,n}$	Conditional risk (i.e., risk of ψ_n with respect to P_0, where ψ_n is based on cross-validation training sets of size $n(1 - p_n)$.)
$\hat{\theta}_{p_n,n}$	Cross-validated estimator of $\tilde{\theta}_{p_n,n}$(i.e., risk of ψ_n with respect to P_n, where ψ_n is based on cross-validation training sets of size $n(1 - p_n)$ and risk is calculated using corresponding validation sets of size np_n.)
$\tilde{k}_n$	The minimizer of $\tilde{\theta}_n$, among the candidate learners.
$\tilde{k}_{p_n,n}$	The minimizer of $\tilde{\theta}_{p_n,n}$, among the candidate learners.
$\hat{k}_{p_n,n}$	The "cross-validated selector," defined as the minimizer of $\hat{\theta}_{p_n,n}$, among the candidate learners.

Table 1.2: Notation

Chapter 2

Prediction using Clinical Data

2.1 Introduction and definition of terms

In this chapter, we will use the statistical foundation presented above to create prediction algorithms that target the diagnosis and prognosis of dengue using data from 3,578 pediatric patients in Nicaragua collected over a ten year period. As 90% of the world's severe dengue cases occur in children under age 15, it is an important population to examine [39]. We will present results for algorithms which aim to reduce statistical risk with no constraint on model complexity or variable costs, as well as algorithms which attempt to strike a balance between complexity, costs, and risk reduction.

Before proceeding, let us define our diagnostic terms and abbreviations, also presented in table 2.1. All patients in our sample are suspected dengue patients, which means that they had an acute fever (at least $37.5°C$) or history of fever less than 7 days and displayed at least one of the WHO's criteria for suspected dengue if hospitalized, or at least two of the WHO's criteria if not hospitalized. Additionally, suspected dengue patients must not have an initial alternative non-dengue diagnosis. A patient is considered dengue positive (DENV) if one or more of the following criteria were met: (a) dengue viral RNA was detected by RT-PCR, (b) dengue virus was isolated, (c) seroconversion of dengue virus-specific IgM was detected by MAC-ELISA in paired acute and convalescent samples, (d) there was at least a four-fold increase in antibody titer specific to dengue virus, measured using Inhibition ELISA in paired acute and convalescent samples. Patients who tested DENV-negative are classified as OFI (other febrile illness) patients while those who tested positive are further classified as having Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF), or Dengue Shock Syndome (DSS) in accordance with the 1997 ("traditional") WHO criteria, summarized in Table 2.1. A newer WHO classification scheme segments dengue-positive disease states into Dengue without

Warning Signs, Dengue with Warning Signs, and Severe Dengue, as defined in Table 2.1. While this newer classification scheme is more easily implemented by physicians, it is less indicative of the pathophysiology of the disease by not distinguishing, for example, patients with plasma leakage from those without [41]. Thus, our analysis will use the traditional WHO classification scheme.

Throughout this chapter, we categorize clinical predictors into three groups: basic, lab, and costly. The *basic* category includes clinical information that can be cheaply and easily obtained at virtually any care facility (e.g., presence of skin rash, cough, pulse, temperature); the *lab* category includes information obtained from a blood or urine analysis that requires no more than a microscope or other relatively inexpensive piece of equipment (e.g., cholesterol, white blood cell count, albumina concentration); the *costly* category includes information obtained via ultrasound, X-ray, or other expensive piece of equipment (e.g., RT-PCR) and is thus available in only resource-rich settings (e.g., interstitial fluid, spleen enlargement, gallbladder wall thickening).

2.2 Literature on dengue prediction with clinical features

While a number of studies have examined clinical features associated with dengue [48] and severe dengue [76], fewer have developed predictive models with validated performance measures for dengue diagnosis and prognosis, and no study has exploited the power of the super learner algorithm for such purposes. The list of covariates that we examine is also uniquely expansive, enabling us to address questions of variable importance not previously explored.

In tables 2.2 and 2.3, we summarize the predictive model literature, restricting to studies with a reasonably transparent methodology, at least 20 patients in each comparison group, and which use (or at least may have used) cross-validation or another method involving an independent test set to assess prediction accuracy. When a study included multiple approaches, we list the most successful among them. Disappointingly, only one study (Tuan, 2015) included confidence intervals around its performance estimates.

Clinical features-based diagnostic algorithms for DENV have been mainly developed using logistic regressions and decision trees (Table 2.2), though Tuan recently explored the potential of random forests, finding that its performance was not measurably different from that of logistic regression [62]. Using basic clinical features and lab measurements, logistic regression models have yielded cross-validated AUC estimates as high as .93 [49] while decision trees, which are more user-friendly in a clinical setting, have achieved similarly impressive

Term	Definition
Suspected dengue	Acute fever plus at least two of the following: (1) headache, (2) retro-orbital pain, (2) myalgia, (3) leukopenia, (4) arthralgia, (5) rash, (6) hemorrhagic manifestations, (7) hospital admission.
Dengue-positive (DENV)	Laboratory-confirmed dengue case
Other febrile illness (OFI)	Suspected dengue but not dengue-positive
Dengue Fever (DF)	Dengue-positive without DHF or DSS
Dengue hemorrhagic fever (DHF)	Dengue-positive plus *all* of the following: (1) fever of history of acute fever lasting 2 - 7 days, occasionally biphasic (2) hemorrhagic manifestations (positive tourniquet test; petechia, equimosis, purpura or bleeding from mucosa, gastrointestinal tract, injection sites or other locations; hematemesis/melena), (3) thrombocytopenia ($\leq 100,000$ platelets/mm^3), (4) evidence of plasma leakage due to increased vascular permeability.
Dengue shock syndrome (DSS)	DHF plus (1) hypotension for age or narrow pulse pressure (≤ 20 mmHg) and (2) one of: rapid and weak pulse; cold, clammy skin, restlessness.
Dengue without Warning Signs	Dengue-positive with 2 of the following: (1) nausea or vomiting, (2) Rash, (3) aches and pains, (4) leukopenia, (5) positive tourniquet test.
Dengue with Warning Signs	Dengue without Warning Signs plus any of the following: (1) abdominal pain or tenderness, (2) persistent vomiting, (3) clinical fluid accumulation, (4) mucosal bleeding, (5) lethargy, restlessness, (6) liver enlargement ¿ 2 cm, (7) increase in HCT concurrent with rapid decrease in platelet count (laboratory).
Severe Dengue	Dengue without Warning Signs plus at least one of the following: (1) severe plasma leakage leading to shock and/or plasma leakage leading to fluid accumulation with respiratory distress, (2) severe bleeding as evaluated by clinician, (3) severe organ involvement (Liver: AST or ALT ≥ 1000; impaired consciousness; failure of heart and other organs).

Table 2.1: Disease classifications

cross-validated AUC estimates [60]. (Or course, we cannot tell how future performances of these models on new data will compare using only the point estimates provided.) Tuan's 2015 study was the only one to consider day of illness for diagnosing DENV; surprisingly, it was not selected by his predictor selection methods, though he did not test its interaction with other covariates (aside from implicitly when implementing CART).

Studies that distinguish OFI and DF patients from DHF/DSS patients (Table 2.3) fall into two categories: those that only use data from patients who have not yet satisfied the DHF/DSS definition at the time of measurement (resulting in an algorithm that is *prognostic* in nature), and those which use data from all patients, even those which have already satisfied the DHF/DSS definition at the time of sample collection (resulting in an algorithm that is *diagnostic* in nature). Algorithms in the latter category are not useful for guiding clinical management, but can be useful for disease surveillance by offering an alternative to disease classification methods that rely on expensive and often unavailable tools (e.g., to precisely apply the WHO's DHF criteria, a chest X-ray is needed to ascertain pleural effusion). For diagnostics, Potts finds that he can distinguish between OFI/DF and DHF pediatric patients in Thailand with 77% sensitivity and 86% specificity using only three measurements (platelet count, aspartate aminotransferase, and hematocrit), selected from a set of 11 candidate predictors using forward step-wise additions in a logistic regression framework [49].

For the prognosis of dengue, Potts developed a decision tree model that predicted whether suspected dengue patients would go on to develop DSS, as opposed to being classified as OFI or DF, with 97% sensitivity and 48% specificity (where a misclassification cost ratio of 1:10 severe dengue vs. non-severe was used in order to prioritize sensitivity). This model used four laboratory features collected during the first 72 hours of fever onset: white blood cell count, percent monocytes, platelet count, and hemocrit level [47]. Lee similarly approached the prognostics problem by developing a decision tree, though he took the OFI patients out of the equation, instead focusing on distinguishing eventual DF from eventual DHF patients [33]. His prediction results were very similar to those of Potts, achieving 100% sensitivity and 46% specificity when applying the algorithm to his test data. Of the 38 basic and lab variables considered, Lee used the three found to be significant from multivariate analysis and which remained in the decision tree model after applying standard stopping and pruning criteria: history of bleeding, urea levels, and total protein levels. Thus, both Lee and Potts used variables that themselves define DHF/DSS in their prognostic model. (This is allowed since none of the patients yet qualified as DHF/DSS at the time of data collection.) Tanner also implemented a decision tree approach for predicting the development of severe dengue, though he used a simpler definition by classifying patients as having "severe dengue" if their platelet count was less than 50,000/mm^{3}on days 5 to 7 of illness. He also used a different set of candidate predictors by including, in addition to the standard clinical features, the crossover threshold value of real-time RT-PCR (Ct) and the presence of anti-dengue IgG antibodies using patient samples acquired during the first 72 hours of fever onset. These indicators were both selected by the decision tree's stopping and pruning criteria. Though Tanner achieved respectable results (78% sensitivity and 80% specificity) with a simple decision tree involving just three variables, real-time RT-PCR is not widely available, thus limiting the usefulness of his prognostic algorithm. Strikingly, none of the studies to the authors' knowledge that distinguish OFI and DF patients from DHF/DSS patients use day of illness as a candidate covariate. This is particularly surprising given that many dengue and severe dengue symptoms vary substantially by day of illness [5, 76], with day of illness therefore seeming to be an important interaction term to consider.

Method			Data			Performance			Reference
Statistical model and objective	Predictors considered	Predictors in final model	Patient location and age range	Time of sample collection (Days since fever onset)	Outcome groups and sample sizes	sensitivity (%)	specificity (%)	AUC	
Decision tree (C4.5) for diagnosis	15 basic + 24 lab + 1 costly	7 chosen via tree stopping and pruning criteria	Adults in Singapore and Vietnam	0 - 3	OFI(836), DENV(364)	71	90	.88	Tanner & Schreiber, 2008 [60]
Logistic regression for diagnosis	2 basic + 9 lab	5 chosen via forward step-wise additions	Children in Thailand	0 - 3	OFI(613), DENV(614)	82	91	.93	Potts, 2010a [49]
Logistic regression for diagnosis	36 basic	6 chosen via univariate analysis, backward and forward step-wise selection	Adults in Singapore	0 - 2	OFI(233), DENV(148)	74[a]	79[a]	.82[a]	Chadwick, 2006 [9]
Logistic regression for diagnosis	36 basic + 18 lab	6 chosen via univariate analysis, backward and forward step-wise selection	Adults in Singapore	0 - 2	OFI(233), DENV(148)	84[a]	85[a]	.92[a]	Chadwick, 2006 [9]
Logistic regression for diagnosis	13 basic + 6 lab	3 chosen via stability selection	Children in Vietnam	0 - 3	OFI(4015), DENV(1692)	75 [73 - 77]	76 [75 - 78]	.83	Tuan, 2015 [62]
Random Forests for diagnosis	13 basic + 6 lab	19 (no selection procedure)	Children in Vietnam	0 - 3	OFI(4015), DENV(1692)	87 [83 - 87]	81 [58 - 85]	.81 [.71 - .87]	Tuan, 2015 [62]

Table 2.2: DENV diagnosis results in the literature

[a] No mention of using a validation set to obtain performance measures – results may be optimistic.

Method			Data			Performance			Reference
Statistical model and objective	Predictors considered	Predictors in final model	Patient location and age range	Time of sample collection (Days since fever onset)	Outcome groups and sample sizes	sensitivity (%)	specificity (%)	AUC	
Decision tree (CHAID) for prognosis	24 basic + 14 lab	3 chosen via multivariate regressions and tree pruning	Adults in Singapore	3 - 7 (5th-95th percentile)	DF(1855), DHF(82)	100	46	NA	Lee, 2009 [33]
Logistic regression for prognosis	23 basic + 14 lab	4 chosen via multivariate regressions	Adults in Singapore	3 - 7 (5th-95th percentile)	DF(1855), DHF(82)	98[a]	60[a]	.89[a]	Lee, 2008 [34]
Logistic regression for diagnosis	2 basic + 9 lab	6 chosen via forward step-wise additions	Children in Thailand	0 - 3	DF(386), DHF(228)	77	81	.86	Potts, 2010a [49]
Decision tree (CART) for prognosis	2 basic + 9 lab	4 chosen via tree stopping rules	Children in Thailand	0 - 3	OFI/DF(1193), DSS(37)	97	48	NA	Potts, 2010 [47]
Logistic regression for diagnosis	2 basic + 9 lab	3 chosen via forward step-wise additions	Children in Thailand	0 - 3	OFI/DF(999), DHF(228)	77	86	.91	Potts, 2010a [49]
Decision tree (C4.5) for prognosis	15 basic + 24 lab + 1 costly	3 chosen via tree stopping and pruning criteria[b]	Adults in Singapore	0 - 3	"non-severe" (55), "severe" (106)[c]	78	80	.83	Tanner & Schreiber, 2008 [60]

Table 2.3: Severe dengue classification results in the literature

[a] No mention of using a validation set to obtain performance measures – results may be optimistic.

[b] Final criteria included (a) the crossover value (Ct) or the real-time RT-PCR for dengue viral RNA, (b) the presence of anti-dengue IgG antibodies, and (c) platelet count.

[c] Dengue cases were classified as "non-severe" if platelet count>50,000/mm^3 on days 5 to 7 of illness.

Infectious disease researchers have demonstrated considerable interest in developing better tools for dengue diagnosis and prognosis, and some have had significant success using easy or moderately easy to obtain clinical information. Still, day of illness as a covariate for the prognosis of severe dengue has not yet been examined and nor have indicators obtained via ultrasound and x-ray equipment. Ensemble learning methods have also not yet been fully utilized for prognostics nor for diagnostics[1], and there remains a dearth of studies conducted using data from outside of Asia. Our study, therefore, fills a significant gap in the literature.

2.3 Methods

2.3.1 Data description

Data collection overview

Our data contains repeated measurements for 3,578 suspected dengue patients in Managua, Nicaragua, collected from September 2004 to April 2015. This data comes from two main sources: (a) an ongoing prospective cohort study [29], and (b) a cross-sectional hospital-based study [41]. Both studies were approved by the Institutional Review Boards of the University of California at Berkeley, the Nicaraguan Ministry of Health, and the International Vaccine Institute in Seoul, Korea. Parents or legal guardians of all subjects provided written informed consent while subjects over the age of five provided verbal assent.

For the prospective cohort study, children aged 2 to 9 were recruited through home-to-home visits in District II of Managua, a low-to-middle income area with a population of approximately 62,500 [5]. Children were eligible to remain in the study until their 15th birthday, or until they moved from the study area, with additional two year-olds enrolled each year. Medical information was obtained from the Health Center Socretes Flores Vivas (HCSFV), a public health clinic that is the primary source of care for District II. All study participants were encouraged to seek medical care at HCSFV, free of charge, upon the first sign of illness. Those who met the WHO criteria for suspected dengue (acute fever plus at least 2 of items 1 - 6 listed in table 2.1) and lacked an alternative febrile-illness diagnosis were treated as possible dengue patients. At HCSFV, clinical information was obtained for suspected dengue patients on a daily basis, with complete blood counts (CBC) completed every 48 hours. If these patients displayed any sign of alarm, they were transfered to the Infectious Disease Ward of the Hospital Infantil Manual de Jesus Rivera (see hospital-based

[1] Some machine-learning algorithms for variable selection were explored by Ju and Brasier [24] for the prediction of DHF, though they only had 13 DHF patients in their sample.

study description, below).

The hospital-based study involves the collection of detailed clinical information of possible dengue patients who were admitted to the Hospital Infantil Manual de Jesus Rivera (HIMJR), the national reference children's hospital in Managua, Nicaragua. We used the data collected for children between 6 months and 15 years old who came to HIMJR with a fever or history of fever for fewer than 7 days and who additionally displayed at least one of the following: headache, arthralgia, myalgia, retro-orbital pain, positive tourniquet test, petechia, or signs of hemorrhaging. Additional exclusion criteria included: (a) episodes with an alternative febrile illness diagnosis, (b) children weighing less than 8 kg, (c) children 6 years of age or older displaying signs of altered consciousness at the time of recruitment, (d) episodes already captured by the clinic (cohort) study described above. Most clinical information was obtained for inpatients every 12 hours, with a Complete Blood Count, blood chemistry, x-ray and ultrasound data collected on a daily basis for a minimum of three days. Between 14 and 21 days after symptom onset, a blood sample was again taken and analyzed as part of a convalescent follow-up. For outpatients, clinical data was collected once per day using a less comprehensive form than that which was used for inpatients.

Data preparation

Data cleaning was done in `R` [50], with biologically infeasible values removed from both the descriptive statistics and all analyses presented below. Clinical values were standardized before conducting the machine-learning analyses, as the penalized regression methods are not equivariant under scaling. We consistently standardized variables using the full data (hospital plus cohort patients) so that algorithms are applied in a consistent manner to the training and test sets.

Variable descriptions

The health clinic recorded patient information using "Case Report Form A" of Appendix A (translated from Spanish to English); the more detailed information gathered for hospital inpatients was recorded using "Case Report Form B" of Appendix A (also translated). Nearly all information recorded for clinic patients was also recorded for hospital patients but not vice versa. Among the variables *not* collected at the clinic are indicators necessary for determining disease severity; as a result, initial disease severity diagnosis for the clinic patients is unknown and we will thus be unable to use these observations for our prognostic work.

As mentioned, repeated measurements were recorded for each patient-episode. However, since our goal is to develop a diagnostic and prognostic algorithm that is applicable to

	Binary	Categorical	Continuous	Total
Demographics	1	0	1	2
General signs and symptoms	22	1	7	30
General indicators of hemorrhaging	13	1	0	14
Blood and urine counts (lab)	0	1	8	9
Blood chemistry (lab)	0	0	17	17
Ultrasound and X-ray	7	0	3	10
Total	43	3	36	82

Table 2.4: Overview of clinical variables available for hospital patients.

patients upon arrival to a health facility, we consider only clinical information collected during the first clinical consultation (occurring within the first 12 hours of the patient's arrival) as input in our prediction methods.

Tables 1 through 5 of Appendix B describe all 82 variables extracted from these forms that we have available for our analyses. This set of variables excludes those which had ten or fewer non-missing values in the hospital data and those which had zero variation. Serotype is only applicable to DENV patients and is thus used only for the disease severity prediction analysis, not for OFI vs. DENV prediction; all other variables are used for both our diagnostic and prognostic analyses. Table 2.4 gives an overview of our variable types. We have easy-to-collect variables (those which require minimal time and equipment), moderately-difficult-to-collect variables (those which a resource-limited setting could still reasonably acquire), and difficult-to-collect variables (those that require special equipment unlikely to be available in a resource-limited setting). Demographics, general signs and symptoms, and general indicators of hemorrhaging are considered easy-to-collect; blood and urine lab variables are considered moderately-easy-to-collect; and ultrasound and X-ray variables are considered difficult-to-collect. For cohort patients, we have demographic information (age and gender), 25 of the 30 general signs and symptoms indicators, 8 of the 14 general indicators of hemorrhaging, all 9 blood and urine count variables (which also includes serotype), and none of the blood chemistry, ultrasound or x-ray information. (Detailed information on variable availability is in Appendix B.)

Descriptive statistics

The HILMR data contains 1,658 patient-episodes while the HCSFV contains 4,218. Unlike the HILMR data, the HCSFV contains unique person identifiers with which we can distinguish repeated visits from the same individual. These visits are generally separated by one or more years, thus representing distinct patient illness episodes. We will use the

Initial Diagnosis	Final Diagnosis				Overall
	OFI	DF	DHF	DSS	
Suspected DF	657	745	104	19	1525
Suspected DHF	15	0	79	15	109
Suspected DSS	1	0	0	1	24
Overall	673	745	183	57	1658

Table 2.5: Initial and final diagnoses of hospital patients.

4,218 patient-episodes from the clinic, which were generated by 2,609 patients, as a test set with which to evaluate our diagnostic algorithm's power to distinguish OFI from DENV in a different clinical setting. For the remainder of this paper, we will generally use "patient" in place of "patient-episode" to simplify wording.

Table 2.5 indicates our sample sizes for our diagnostic and prognostic analyses using hospital data. For diagnosing dengue, we use data from all 1,618 suspected dengue patients, 985 of which ended up with a confirmed dengue-positive diagnosis. Our analysis involving the prognosis of severe dengue uses only the 868 dengue-positive patients who were not already displaying severe dengue symptoms at the time of sample collection, 123 of which ended up with a severe dengue diagnosis. The clinic data is used solely as a test set for validating our diagnostic model for distinguishing OFI from DENV patients. For this purpose, we use all 4,218 patient-episodes, 582 of which resulted in a confirmed dengue diagnosis.

Initial data tended to be collected earlier in the disease progression for clinic patients than for hospital patients, with the hospital receiving on average sicker patients, some of whom were referred by other health facilities around Nicaragua. About a third of patients came to the hospital five or more days after fever onset (Figure 2.1). Thus, many of the patients with severe dengue were already displaying severe dengue symptoms upon arrival, with the likelihood of presenting with severe dengue symptoms increasing as a function of time since fever onset (Figure 2.2), as expected given the known clinical progression of severe dengue (Figure 1).

Of the confirmed dengue-positive cases with known serotype, approximately half tested as serotype 3 with the remaining cases equally divided between serotypes 1 and 2. The immune response test revealed that about half of dengue-positive cases were first-time dengue infections. Please see Figures 7 through 12 in Appendix B for distributions of all other predictors.

Figure 2.3 summarizes the number of variables available (i.e., non-missing) per patient. Clearly, we will need find an alternative to dropping patients with missing values (not a single patient has zero missing values). Figures 13 through 15 in Appendix B summarize

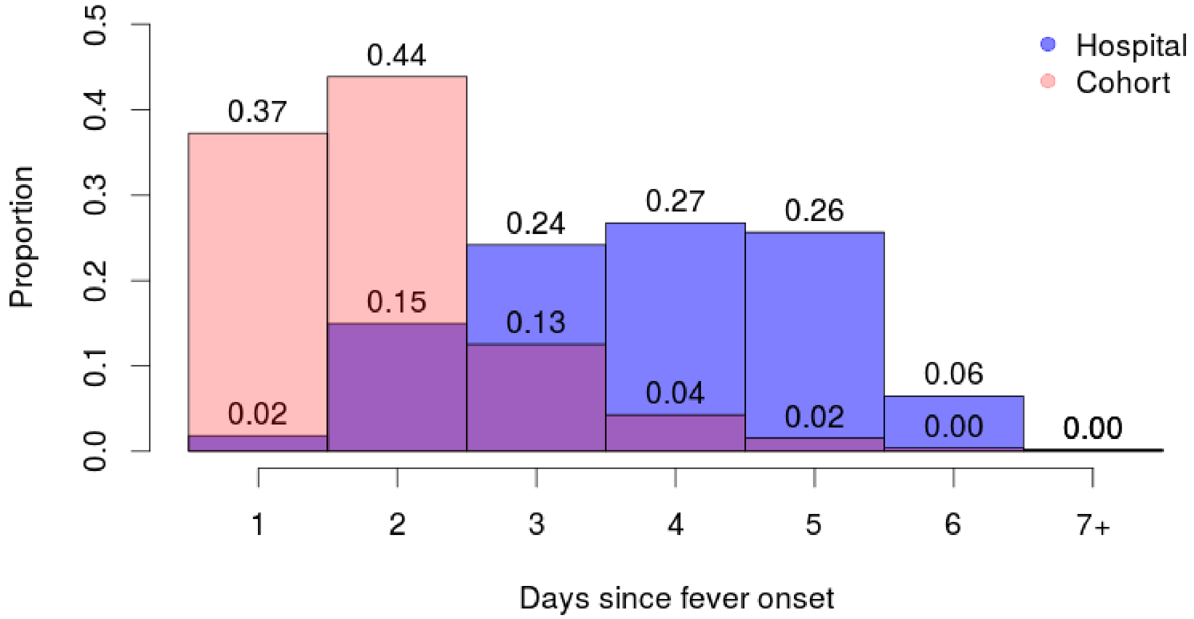

Figure 2.1: Timing of sample collection for cohort and hospital patients.

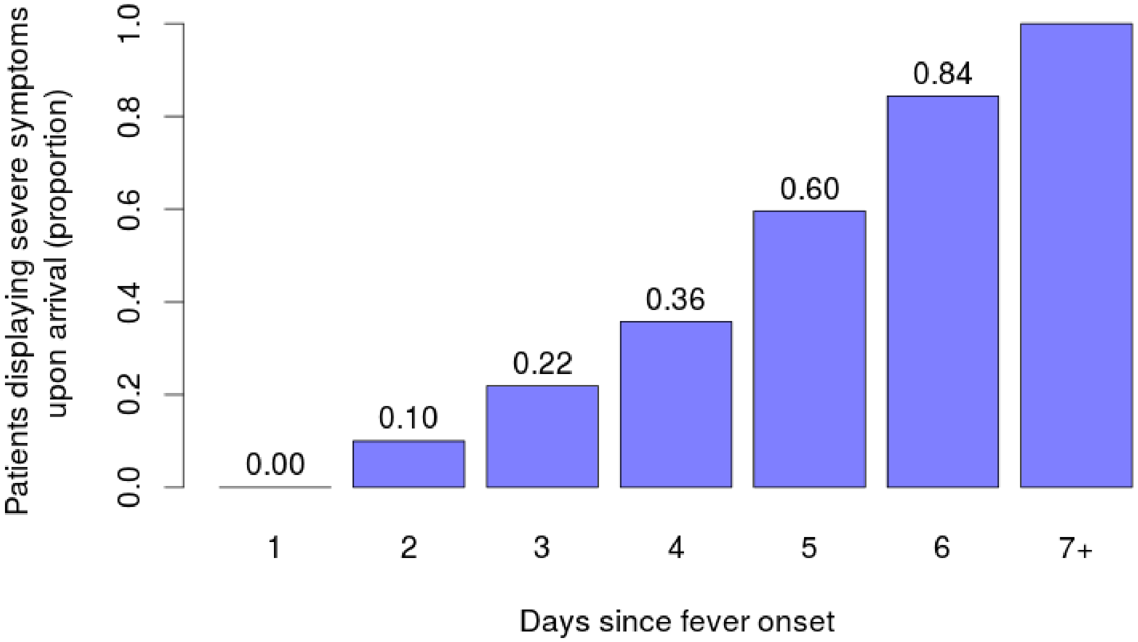

Figure 2.2: Presentation of hospital patients with eventual severe dengue diagnosis.

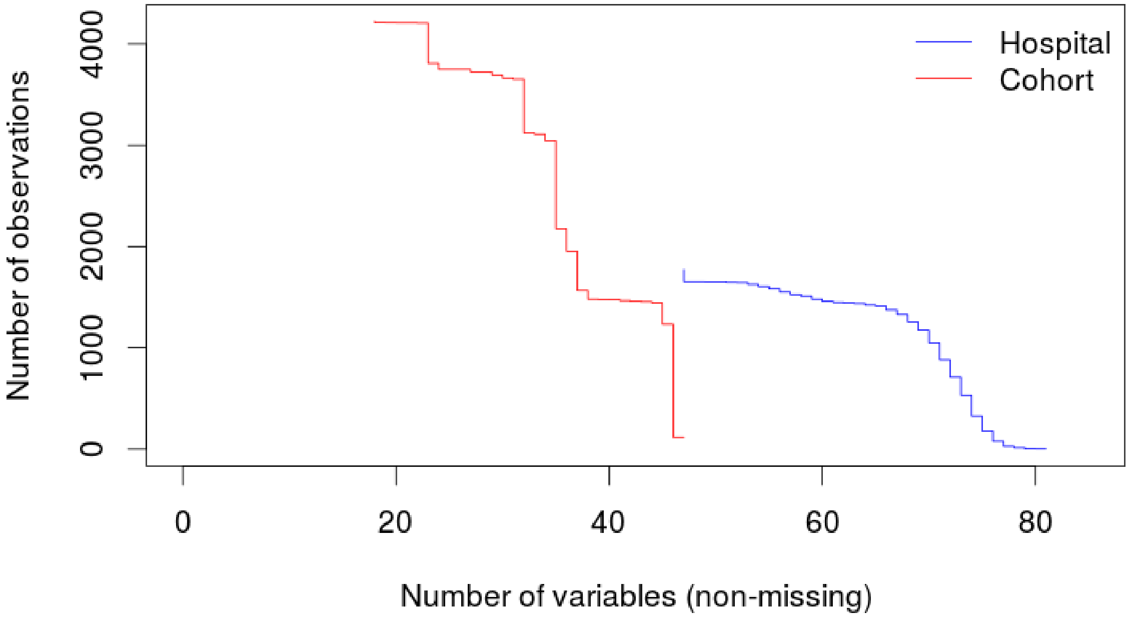

Figure 2.3: Availability of clinical variables per patient-episode

missingness per variable: 38 variables are never missing for any of the hospital patients while 10 variables are never missing across the combined hospital and clinic data. In the next section we will describe our imputation method that allows us to keep all patients and all variables in the analysis.

2.3.2 Prediction algorithms, missingness handling, variable importance, and statistical inference

Prediction algorithms

We employ prediction algorithms of various levels of intricacy, from the super learner (high intricacy, high accuracy) to classification trees (low intricacy, low accuracy). We refer the reader to Chapter 1 for details on these prediction algorithms and their properties. We used `Python's scikit-learn` module [44] to implement the algorithms in our super learner library, employing grid searches using 5-fold cross-validated AUC to choose optimal tuning parameters. Additional details on our implementation of each algorithm are provided in Table 2.6. Super learner's coefficients are estimated using sequential least squares programming

Name	Details
Mean	Assigns most frequent outcome to all observations.
CART	Decision tree classifier with maximum depth chosen via CV.
Centroids	Nearest centroid classifier using euclidean distance.
LDA+shrinkage	Linear discriminant analysis with optimal shrinkage determined using the Ledoi-Wolf lemma [53].
Gradient Boost	Gradient boosted trees in which maximum depth, minimum number of samples required for a split, and minimum number of samples required to be at a leaf node are determined with CV. We build 100 trees and consider all predictors for each split.
Adaboost	Implementation of the Adaboost-SAMME algorithm [20] with 500 decision tree classifiers.
Random Forests	Random forests with 500 trees built using Gini impurity to measure the quality of split. We consider p features for each possible split where p is determined using CV.
N.Neighbor	N-nearest neighbors using euclidean distance with number of neighbors and the weights they are given (uniform or inversely proportional to distance) chosen via CV.
Logit-L1	L_1-penalized logistic regression with the complexity parameter chosen via CV.
Logit-L2	L_2-penalized logistic regression with the complexity parameter chosen via CV.
Elastic Net	Regularized logistic regression using a combination of the L_1 and L_2 penalties. The weight given to each penalty term is determined via CV.
SVM-L2	Support vector machine using the squared hinge loss with the complexity parameter on the L_2 penalty term fit via CV.
Super Learner	Weighted combination of the above algorithms where weights are determined via CV with the negative log likelihood loss.

Table 2.6: Algorithm implementations

to minimize the negative log likelihood, also using 5-fold cross-validation.

Note that we use a total of three embedded cross-validation steps to obtain super learner prediction results: we cross-validate the super learner itself, and within each fold of this outermost cross-validation step we fit the super learner which involves running each algorithm inside of a cross validation step. Finally, the algorithms that have tuning parameters are run using cross-validation every time they are fit to a training set. So, using 5-fold cross-validation each time means that at the inner most level we are using $\frac{4}{5} \times \frac{4}{5} \times \frac{4}{5} = \frac{64}{125}$, or about half of all observations, to fit the data and $\frac{4}{5} \times \frac{4}{5} \times \frac{1}{5} = \frac{16}{125}$, or about 13% of all observations, to judge the performance of our tuning parameter values.

Missingness Handling

Missing values in our data typically occur for three reasons: (1) unavailable equipment, (2) different intake forms for inpatient and outpatients, (3) physician deemed the test to be irrelevant. Due to reasons (2) and (3), our missingness is therefore *not* random. While this fact complicates the causal interpretation of variable effects, our inference with regards to conditional prediction accuracy remains valid.

To examine the meaningfulness of the missingness mechanisms in our data, and to exploit these mechanisms for improved prediction accuracy, we employ an imputation indicator method as follows. Let us decompose the vector W_i into covariates W_{1i} and W_{2i} such that we can represent the full data structure as $X = (W_1, W_2, Y)$. Furthermore, suppose we do not have W_2 for a certain subset of patients. Let Δ_{2i} equal one if patient i is observed (non-missing) W_2 covariates, and equal to zero otherwise. Define W_2^* as being equal to W_2 when $\Delta_2 = 1$ and being equal to $E(W_2|W_1)$ when $\Delta_2 = 0$. We estimate $E(W_2|W_1)$ using `R's Random Forests package [36]` (the super learner would be a viable alternative) to get our so-called imputed values. Specifically, we train random forests using observations for which $\Delta_2 = 1$ (with W_1 serving as our predictors) and then use this fitted model to predict W_2 for the observations that lack an observed W_2 value. We can now proceed with our prediction analysis using our new predictor set consisting of W_1, Δ_2, and W_2^*, or can also go further by additionally incorporating an interaction between Δ_2 and W_2^*. Including the interaction term allows the effect of W_{2i}^* on Y_i to differ according to whether or not W_{2i}^* was imputed for patient i. Since our problem is a prediction problem (not a causal inference problem), including Δ_2 and the $\Delta_2 W_2^*$ interaction is a modeling decision that will not affect the validity of our inference with regards to the conditional risk, but could give us a superior prediction performance.

Variable Importance

In the context of statistical prediction, we can conceptualize the "importance" of the *jth* variable various ways. At the two extremes, we have the following importance measures:

1. The effectiveness of variable j at predicting the outcome *without using any other covariates.*

2. Prediction performance *with* variable j relative to prediction performance *without* variable j, keeping all other covariates in the equation.

Clearly, if various covariates are highly correlated with one another and also highly correlated with the outcome, then they will be classified as very important under the first measure and less important under the second. In contrast, a variable that is correlated with the outcome and not with other variables will appear to increase in relative importance under the second measure as compared to the first. Since each measure provides different and useful information, we provide both. Specifically, to measure the effectiveness of variable j at predicting the outcome in isolation of other covariates, we simply run Super Learner with variable j as the only predictor, and report the corresponding cvAUC. (To do this, we modify our library of algorithms to include those appropriate for univariate analysis: linear discriminant analysis, for example, is included *without* shrinkage, and CART is included while random forests is not.) For the second measure, we run Super Learner with and without variable j (using our full library of algorithms) and report the difference in cvAUC (full model minus restricted model such that larger positive values indicate greater significance). Finally, we provide two variable importance measures based on Random Forests, both of which give us *relative* importance. One is based on the decrease in prediction accuracy when vector j is randomly permuted in the out-of-bag samples (the "RF - Permutation" method) and the other is based on the gini index improvements across all tree splits involving variable j (the "RF - Gini" method). Both of these measures are somewhere between measures (1) and (2) in terms of measuring the importance of a variable in isolation of other predictors, and net of the predictive power of the other predictors. Details of these methods are found in Chapter 4.

2.4 Results

2.4.1 Distinguishing DENV from OFI

Findings based on hospital data

Using all available clinical features with imputed values in place of missings, super learner achieves a cross-validated area under the curve (cvAUC) of .87 [95% confidence interval: .86−.87], with the gradient boosting and random forest algorithms performing similarly (Table 2.7). The simple CART algorithm achieves a cvAUC of .67 [95% confidence interval:.66−.68] – substantially lower than the more sophisticated algorithms. Included in our algorithm list is also the simple mean function, which assigns the most dominant label to all observations regardless of covariate values. In our data, there are more DENV than OFI patients; thus, the mean function assigns DENV to everyone and therefore achieves a perfect sensitivity with specificity zero and an overall error rate of 41%. Super learner, meanwhile, achieves an overall error rate of 21% with a negative predictive value of 73% and positive predictive value of 84% when using the threshold value c (discussed in Chapter 1.4.1) which minimizes the error rate.

Method	NPV	PPV	Error Rate	Sensitivity	Specificity	MSE	AUC 95% CI		AUC
Super Learner	0.74	0.85	0.19	0.81	0.79	236	0.86	0.88	0.87
Gradient Boost	0.76	0.82	0.20	0.84	0.73	241	0.86	0.87	0.86
Random Forests	0.77	0.79	0.22	0.86	0.67	256	0.85	0.86	0.86
Logit-L2	0.76	0.79	0.22	0.85	0.67	313	0.83	0.85	0.84
Logit-L1	0.76	0.79	0.22	0.86	0.66	265	0.83	0.84	0.84
LDA+shrinkage	0.75	0.77	0.24	0.86	0.62	278	0.82	0.83	0.83
AdaBoost	0.70	0.77	0.26	0.81	0.64	412	0.80	0.82	0.81
SVM-L2	0.73	0.79	0.23	0.83	0.69	381	0.74	0.77	0.76
N.Neighbor	0.67	0.72	0.29	0.83	0.53	333	0.73	0.75	0.74
CART	0.71	0.73	0.27	0.85	0.55	326	0.72	0.74	0.73
Centroids	0.55	0.79	0.35	0.56	0.79	579	0.66	0.69	0.67
Mean	NA	0.59	0.41	1.00	0.00	673	0.48	0.52	0.50

Table 2.7: Cross-validated performance measures of algorithms using all available clinical features, OFI vs. DENV analysis. The positive predictive value (PPV), negative predictive value (NPV), error rate, sensitivity and specificity are based on the threshold value c (discussed in Chapter 1.4.1) which minimizes the error rate.

Including imputation indicators only marginally affects results (compare blue and yellow bars in Figure 2.4[2]), indicating that either missingness is uncorrelated with outcome, or that missingness provides information that is largely redundant with the information provided by other covariates. To get a sense for how informative missingness is, we ran super learner using only imputation indicators and no actual clinical information (gray bars in Figure 2.4). Doing so gave us a cross-validated AUC of .65 [.64 – .66] with super learner; missingness is not *completely* random, but neither is it a substitute for clinical information. Since the missingness mechanism in our particular hospital setting is unlikely to generalize to new datasets, we opt to exclude imputation indicators from the remaining analyses in order to develop a diagnostic algorithm that is more likely to be useful in other contexts. (And even if we instead care only about our algorithms' expected performance for diagnosing future patients in the HILMR facility where the missingness mechanism remains constant, our results reveal that little to nothing is to be gained by including the imputation indicators.)

Not all algorithms are fairly treated if we judge them purely by cvAUC, as our scikit-learn implementations of the support vector machine (SVM-L2) and nearest centroid methods give predicted class membership (dengue or not-dengue), as opposed to predicted probabilities of

[2] Figures presented throughout this section and in subsequent sections were produced using `Python's Matplotlib library` [23].

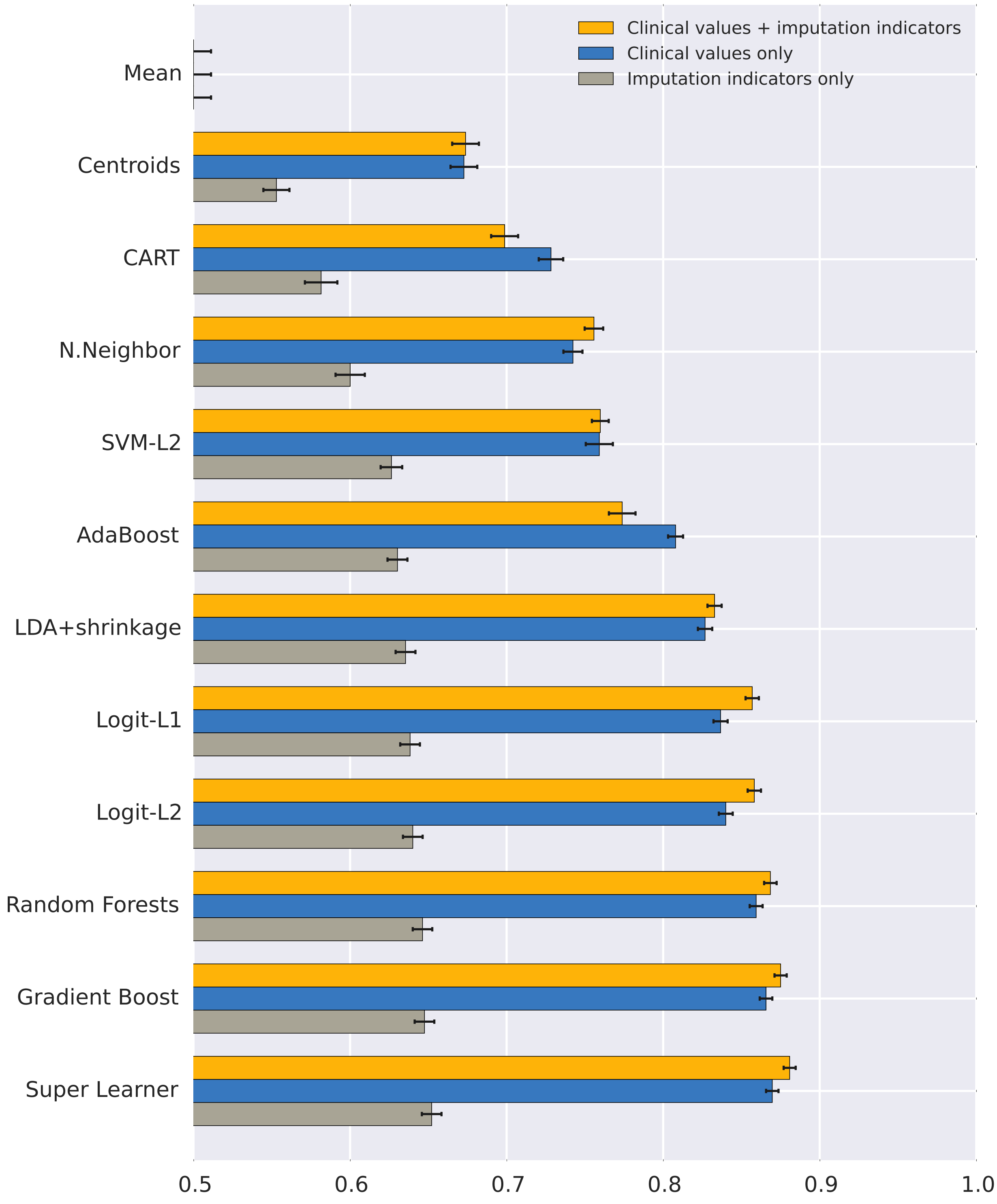

Figure 2.4: Cross-validated AUCs and corresponding 95% confidence intervals for various algorithms with and without missing value indicators, OFI vs. DENV analysis.

dengue. As a result, their cross-validated ROCs (and corresponding AUCs) are effectively based on only three points since all critical values between 0 and 1 give the same classifications and therefore the same true positive and false positive rates (Figure 2.5).[3] Indeed, we see that the support vector machine compare favorably to other well-performing algorithms when we look at the measures from the first five columns of Table 2.7, yielding an overall error rates of 23%.

Interestingly, the most costly variables (i.e., those obtained using ultrasound and x-rays) add negligible predictive value when used in conjunction with the less expensive variables (super learner's cvAUC remains at .87 after eliminating the costly variables, Figure 2.6). In fact, super learner's cvAUC fell only a small amount (down to .85) when we additionally eliminate the covariates that were not collected in at the health clinic. However, we do find that the moderately expensive lab variables collected in both the clinic and hospital settings contain useful information for prediction: super learner's cvAUC estimate drops to .77 when we remove all lab variables from our predictor set, using only our basic variables for prediction.

Within each of the three main variable categories (basic, lab, and costly), some variables are clearly more helpful for predicting dengue diagnosis than are others (Figure 2.7). Four lab variables stand out as being particularly important: white blood cell count, platelet count, Aspartate aminotransferase (AST), and Alanine aminotransferase (ALT). Indeed, using just these four lab variables in combination with the basic signs and symptoms gives us a cross-validated AUC of .86 $[.85-.86]$ using the super learner algorithm — results indistinguishable from those generated using all basic and lab variables (as well as all basic, lab, and costly variables). Note, though, that by selecting variables using importance measures based on the full data, we are guilty of over-fitting; more principled methods for obtaining a best subset of predictors within the cross-validation step will be covered in Chapter 4.

Results using cohort data

We now investigate the degree to which our fitted algorithms are applicable in a different patient setting: using the final fitted algorithms developed in the previous section with the covariates present in the cohort data, we predict dengue diagnosis for the cohort patients. If both samples contain patients drawn randomly from the same population, then the super learner performance should not suffer when fit to one dataset and tested on another. Neither should performance suffer with infinitely large samples drawn from different populations so

[3] Note that when all we have to work with is predicted class membership, then the AUC corresponds to the [non-weighted] mean of the true positive rate and the true negative rate.

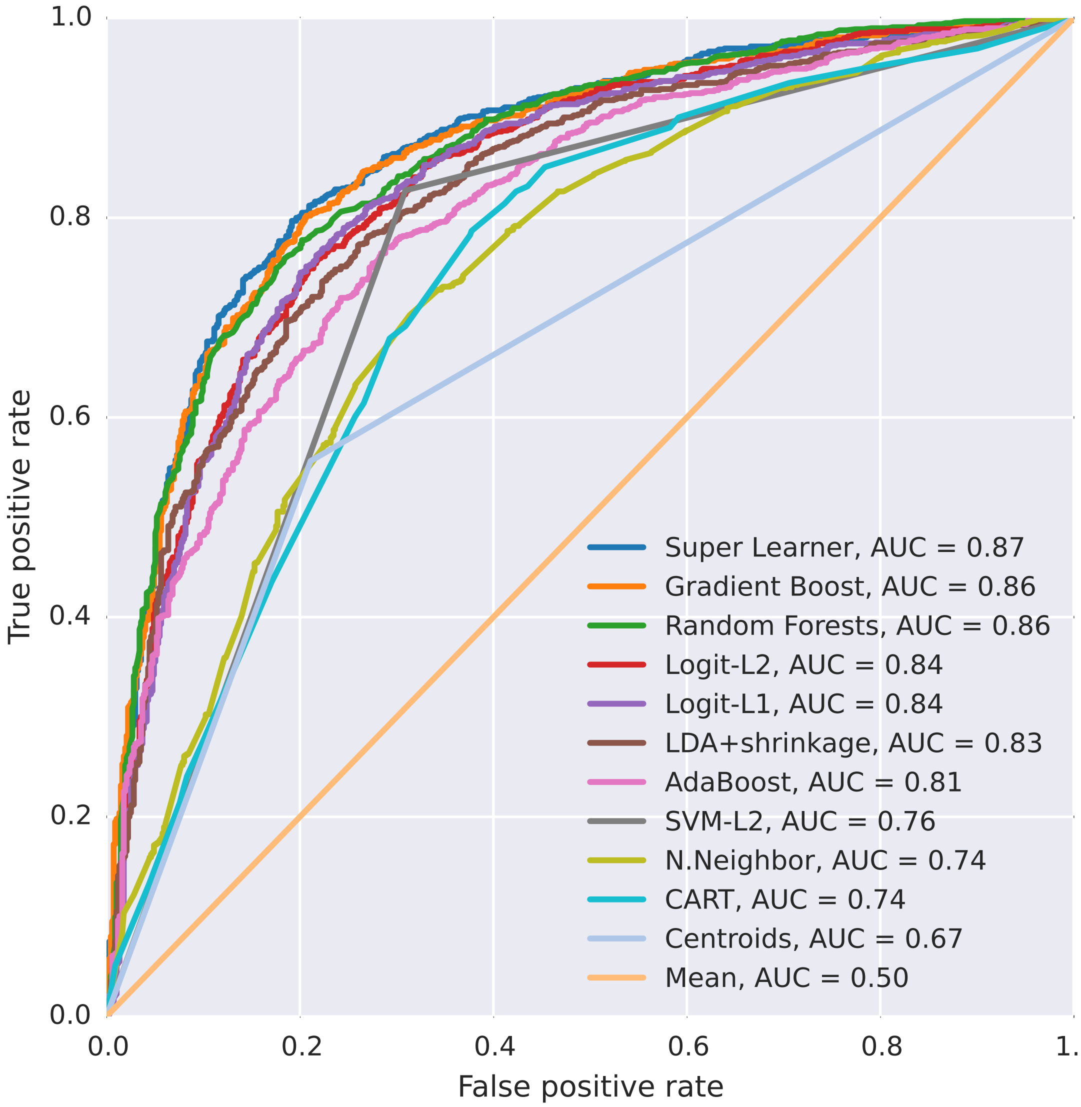

Figure 2.5: Cross-validated ROCs (and corresponding AUCs) of algorithms using all available clinical features, OFI vs. DENV analysis.

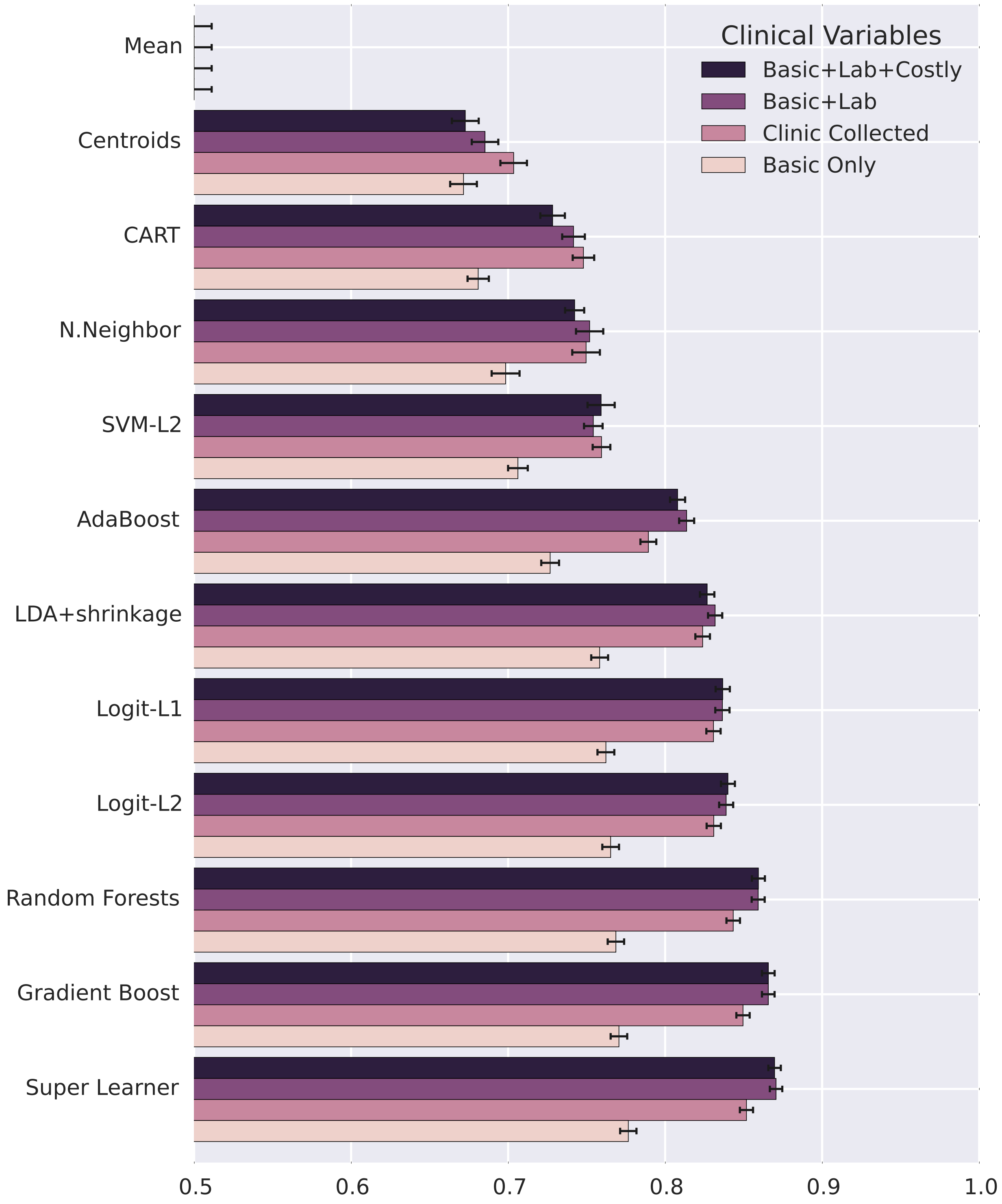

Figure 2.6: Cross-validated AUCs and corresponding 95% confidence intervals for various predictor subsets and algorithms, OFI vs. DENV analysis.

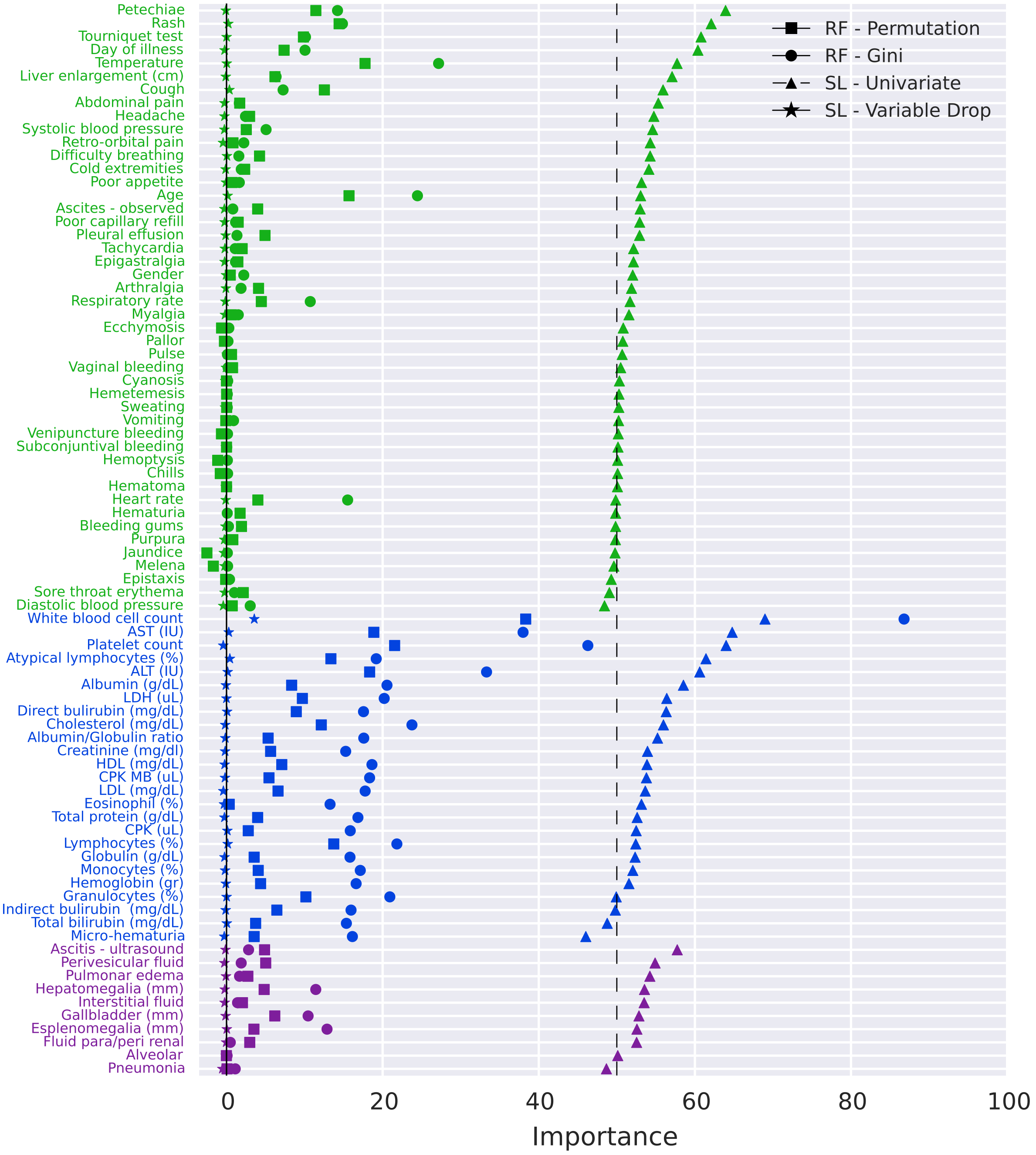

Figure 2.7: Variable importance measures for OFI vs. DENV analysis. Basic variables appear in green, lab variables in blue, and costly variables in purple. Cross-validated AUC values (for super learner methods) have been scaled by 100, with 50 representing "no importance" for the univariate analysis (dotted line) and 0 representing "no importance" for other methods (solid line).

long as the two patient samples contain overlapping patient characteristics and so long as the relationship between patient characteristics and outcome is the same in both samples. However, we have, of course, finite samples drawn from two different patient populations. In this situation we can therefore expect some loss of performance brought but an effectively smaller training set, as the effective size of the training set is conceptually related to the size of the relevant overlapping patient population.

Indeed, algorithms trained on the hospital data do significantly worse at predicting outcomes for the cohort data (pink bars in Figure 2.8) than they do at predicting outcomes for additional hospital patients using cross-validation (dark blue bars in Figure 2.8). For the same reason, cross-validated results based on only the cohort data are superior to the results obtained by fitting to the hospital data and testing on the cohort data (compare red bars to pink bars in Figure 2.8.)

Combining the hospital and cohort data together generates better cross-validated results than analyses based on either dataset alone; apparently sample size was a limiting factor for our machine-learning algorithms fit to the cohort and to the hospital data separately. We additionally find that including a sample indicator (cohort or hospital) has almost no effect on the prediction results of our high-performance estimators and in fact confuses some of our weaker learners; apparently the relationship between the covariates and outcome does not differ substantially between the cohort and hospital data (purple bars in Figure 2.8).

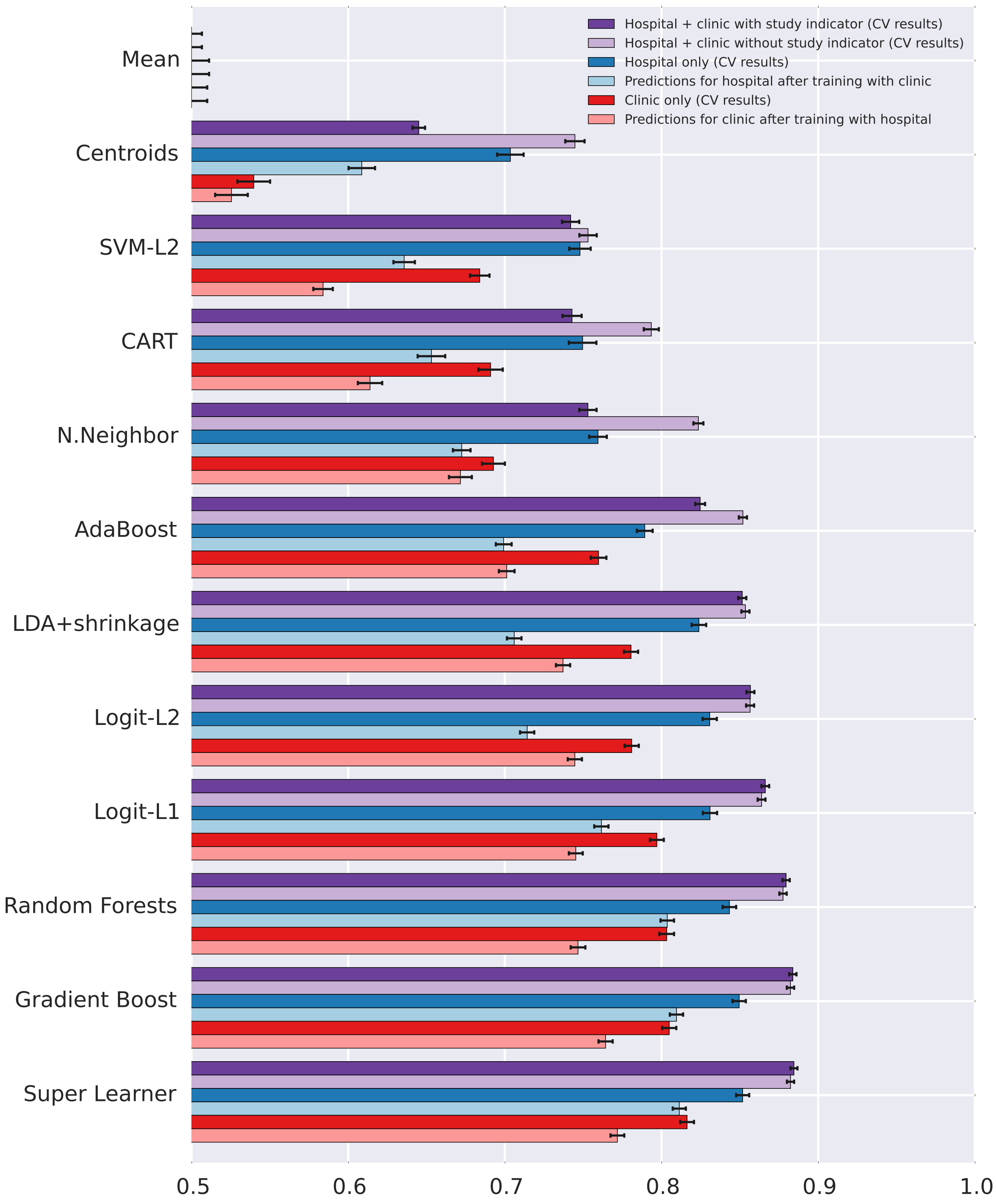

Figure 2.8: Performance comparisons (cvAUC) using different patient sample combinations for training and testing. All results are from using just the 42 covariates applicable to the OFI vs. DENV analysis that are collected in both the clinic and hospital settings.

2.4.2 Predicting DHF/DSS

While distinguishing dengue patients from those with a different febrile illness is important for a variety of reasons, it is arguably more important from a clinical care perspective to identify the patients that will develop DHF or DSS. Here, we present results from two approaches. In the first, we consider all 1,525 suspected dengue patients who did not display DHF/DSS symptoms during the first hospital consultation and try to determine who will develop DHF/DSS using the same predictor sets as we did for the OFI vs. DENV analysis. In the second approach, we use data from the 868 hospital patients who tested as dengue-positive but who did not display DHF/DSS symptoms during the first hospital consult, and predict which of these patients will develop DHF/DSS. In this latter analysis, we use the same predictors as we did for the OFI vs. DENV analysis, but with the addition of dengue serotype. This latter analysis would be applicable in care facilities that are already able to distinguish OFI from DENV patients, while the former assumes no prior knowledge of dengue status. In this section, we will refer to DHF and DSS as "severe dengue" (not to be confused with the updated WHO definition of severe dengue).

We find that our ability to distinguish severe dengue cases from other cases (OFI or DF) is approximately as good as our ability to distinguish OFI from DENV cases. Since only 8% of our patients developed severe dengue, many algorithms achieved their minimal error rate by assigning OFI/DF to all patients (Table 2.8). Using the full set of clinical predictors, super learner was in fact unable to get the misclassification rate lower than 8%. But misclassification rate does not tell the whole story: in looking at sensitivity and AUC, we find that super learner far outperforms many algorithms. Gradient boosting, L_1-penalized logistic regression, and random forests are not far behind.

Similar trends are found with our DF vs. DHF/DSS analysis in terms of the relative success of our various algorithms (Table 2.9). Our AUCs for this analysis are overall slightly lower than for our other two analyses, possibly because it was easier for our algorithms to distinguish OFI patients from DHF/DSS patients than it was to separate DF from DHF/DSS patients (or OFI from DF for that matter). In other words, the OFI vs. DENV and OFI/DF vs. DHF/DSS contained some low hanging fruit that the DF vs. DHF/DSS analysis did not get to pick. Indeed, the easily classified negatives appear to be driving the higher AUCs for the OFI/DF vs. DHF/DSS analysis, as this analysis actually provides lower sensitivity (17% vs. 26%) for a specificity that is only slightly higher (99% vs. 98%) compared to the DF vs. DHF/DSS analysis (this phenomena can also be seen through examination of the ROCs in Figures 2.11 and 2.12).

Including imputation indicators does not affect super learner's performance to any mea-

Method	NPV	PPV	Error Rate	Sensitivity	Specificity	MSE	AUC 95% CI		AUC
Super Learner	0.93	0.60	0.08	0.17	0.99	91	0.87	0.89	0.88
Gradient Boost	0.93	0.56	0.08	0.12	0.99	95	0.86	0.88	0.87
Logit-L1	0.94	0.52	0.08	0.28	0.98	94	0.85	0.87	0.86
Random Forests	0.92	0.64	0.08	0.06	1.00	95	0.85	0.87	0.86
Logit-L2	0.92	NA	0.08	0.00	1.00	421	0.84	0.86	0.85
AdaBoost	0.92	0.50	0.08	0.01	1.00	365	0.82	0.84	0.83
LDA+shrinkage	0.92	NA	0.08	0.00	1.00	111	0.79	0.82	0.80
CART	0.92	NA	0.08	0.00	1.00	102	0.72	0.77	0.74
Centroids	0.92	NA	0.08	0.00	1.00	324	0.69	0.75	0.72
N.Neighbor	0.92	0.50	0.08	0.02	1.00	112	0.66	0.71	0.69
SVM-L2	0.92	NA	0.08	0.00	1.00	129	0.59	0.67	0.63
Mean	0.92	NA	0.08	0.00	1.00	123	0.47	0.53	0.50

Table 2.8: Cross-validated performance measures of algorithms using all available clinical features, OFI/DF vs. DHF/DSS analysis. The positive predictive value (PPV), negative predictive value (NPV), error rate, sensitivity and specificity are based on the threshold value c (discussed in Chapter 1.4.1) which minimizes the error rate.

Method	NPV	PPV	Error Rate	Sensitivity	Specificity	MSE	AUC 95% CI		AUC
Super Learner	0.89	0.64	0.13	0.26	0.98	85	0.82	0.85	0.83
Logit-L1	0.89	0.59	0.13	0.31	0.97	87	0.80	0.83	0.82
Gradient Boost	0.88	0.69	0.13	0.16	0.99	89	0.80	0.83	0.81
Random Forests	0.88	0.67	0.13	0.18	0.99	88	0.80	0.82	0.81
LDA+shrinkage	0.89	0.52	0.14	0.26	0.96	100	0.76	0.79	0.78
AdaBoost	0.86	0.71	0.14	0.04	1.00	210	0.75	0.78	0.76
Logit-L2	0.89	0.51	0.14	0.25	0.96	225	0.74	0.78	0.76
CART	0.86	NA	0.14	0.00	1.00	94	0.70	0.75	0.72
Centroids	0.86	NA	0.14	0.00	1.00	211	0.67	0.73	0.70
N.Neighbor	0.86	0.50	0.14	0.01	1.00	102	0.66	0.71	0.69
SVM-L2	0.86	NA	0.14	0.00	1.00	128	0.60	0.68	0.64
Mean	0.86	NA	0.14	0.00	1.00	123	0.46	0.54	0.50

Table 2.9: Cross-validated performance measures of algorithms using all available clinical features, DF vs. DHF/DSS analysis. The positive predictive value (PPV), negative predictive value (NPV), error rate, sensitivity and specificity are based on the threshold value c (discussed in Chapter 1.4.1) which minimizes the error rate.

surable degree (compare blue and yellow bars in Figures 2.9 and 2.10), though we again find evidence that missingness is informative (gray bars in Figures 2.9 and 2.10). As with the OFI vs. DENV analysis, we exclude missing value indicators from subsequent analyses involving DHF/DSS prediction in order to improve the applicability of our results to other clinical settings.

Cross-validated ROCs again reveal additional information regarding the behavior of our algorithms (Figures 2.11 and 2.12). As before, our support vector machine and nearest centroid algorithms assign class labels rather than probabilities. We also note that CART's ROC contains few distinct kinks, indicating that many observations are assigned the same likelihood of DHF/DSS. That is, tree depth is relatively low. Indeed, our cross-validated CART procedure tends to choose a tree depth of 1 (after testing all depths from 1 to 20) for our severe dengue analyses.

We find that neither the most costly variables nor the other variables collected exclusively in the hospital improve predictions when used in conjunction with the less expensive variables (Figures 2.13 and 2.14). But once again, the moderately expensive lab variables collected in both the clinic and hospital settings contain useful information for prediction: super learner's cvAUC estimate drops to .80 and .75 when we remove all lab variables from our predictor set in the OFI/DF vs. DHF/DSS and in the DF vs. DHF/DSS analyses, respectively.

Our clinical variables rank similarly in their importance for predicting severe dengue as compared to their importance for distinguishing OFI from DENV patients (Figures 2.15 and 2.16): temperature, rash, the tourniquet test, liver enlargement, and petechiae are still among the most important general signs and symptoms, while platelet count, white blood cell count, AST, and albumin concentration continue to be among the most important lab variables. Our most remarkable finding is that using the super learner with platelet count alone generates an AUC of .80 and .76 for our OFI/DF vs. DHF/DSS and DF vs. DHF/DSS analyses, respectively – approximately the same AUC as we get when using all general signs and symptoms. Platelet count is also the only variable whose presence has a measurable impact on the performance of the super learner even when all other 80 variables are included in our analysis (see the "variable drop" measurements in Figures 2.15 and 2.16). In contrast, white blood cell count was the one variable that appeared to provide unique information for the OFI vs. DENV analysis beyond that provided by the other clinical indicators (Figure 2.7).

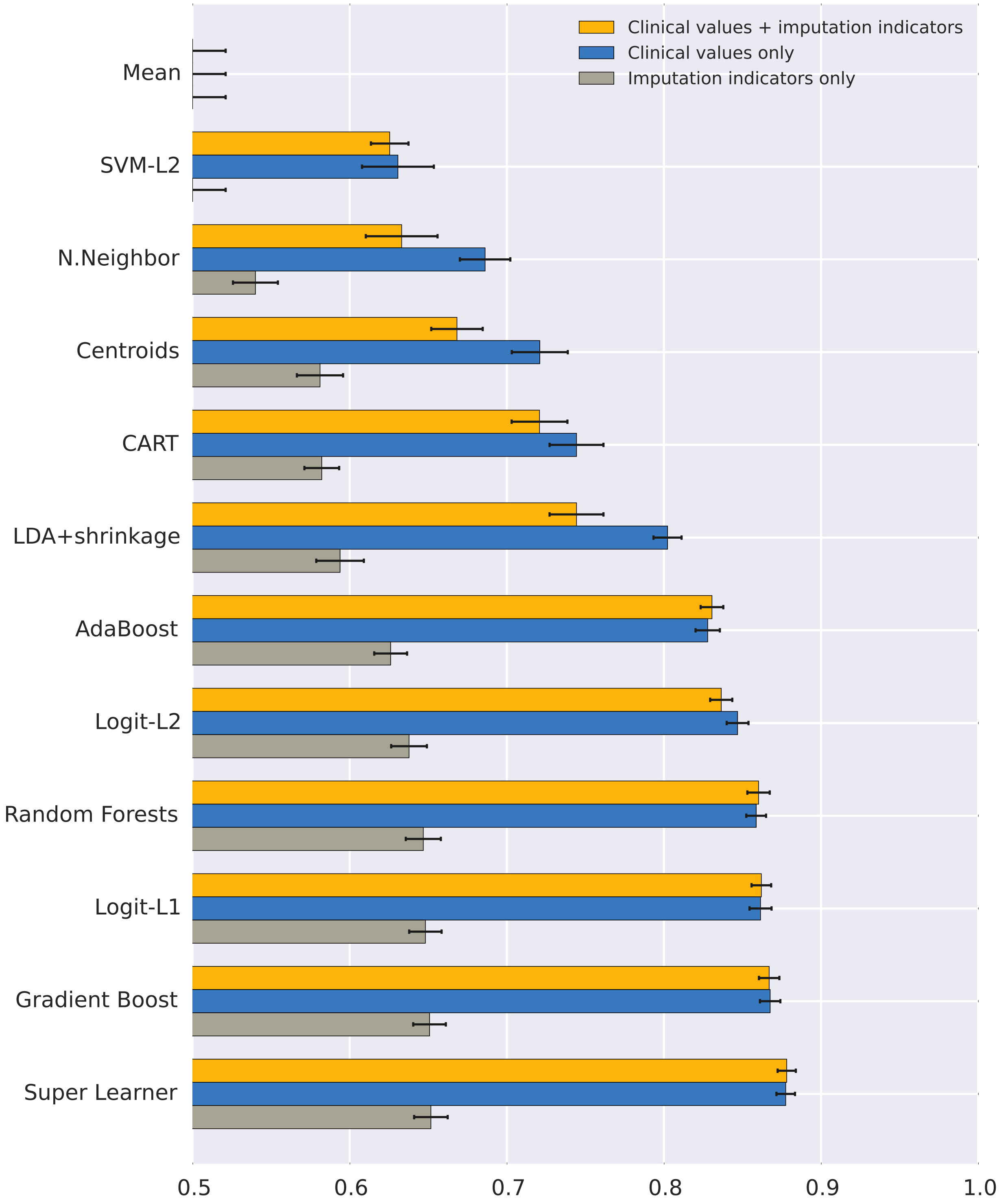

Figure 2.9: Cross-validated AUCs and corresponding 95% confidence intervals for various algorithms with and without missing value indicators, OFI/DF vs. DHF/DSS analysis.

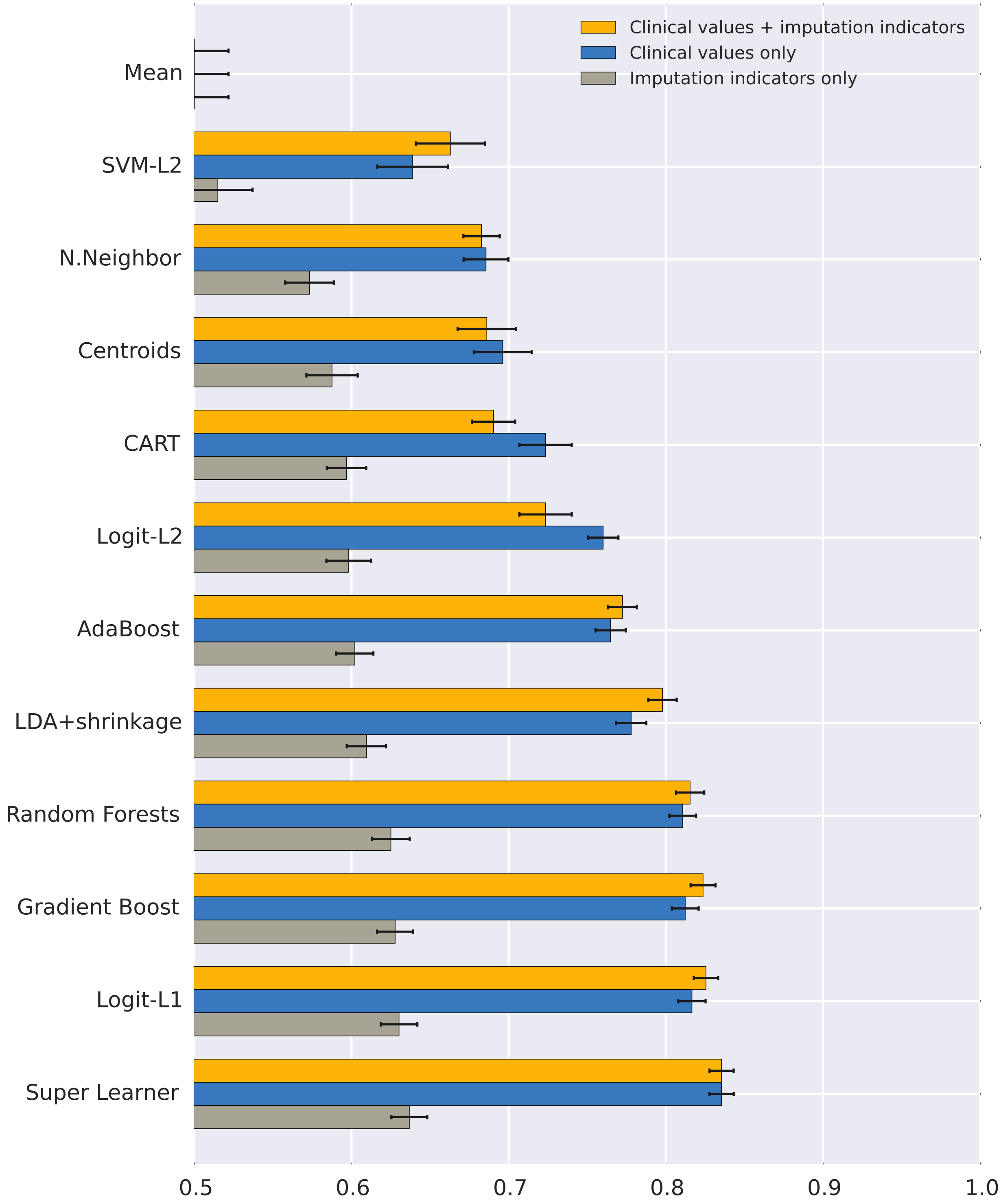

Figure 2.10: Cross-validated AUCs and corresponding 95% confidence intervals for various algorithms with and without missing value indicators, DF vs. DHF/DSS analysis.

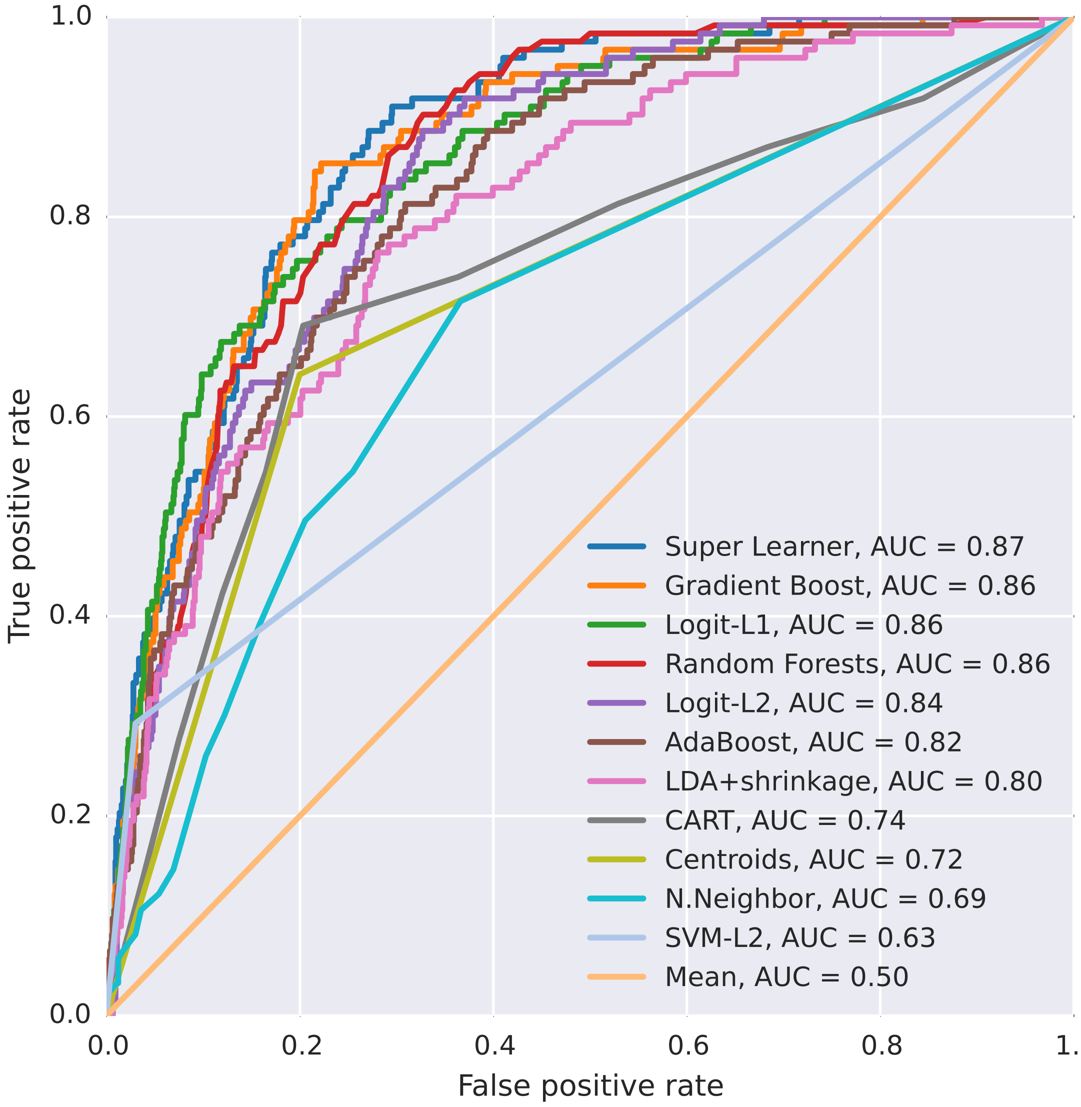

Figure 2.11: Cross-validated ROCs (and corresponding AUCs) of algorithms using all available clinical features, OFI/DF vs. DHF/DSS analysis.

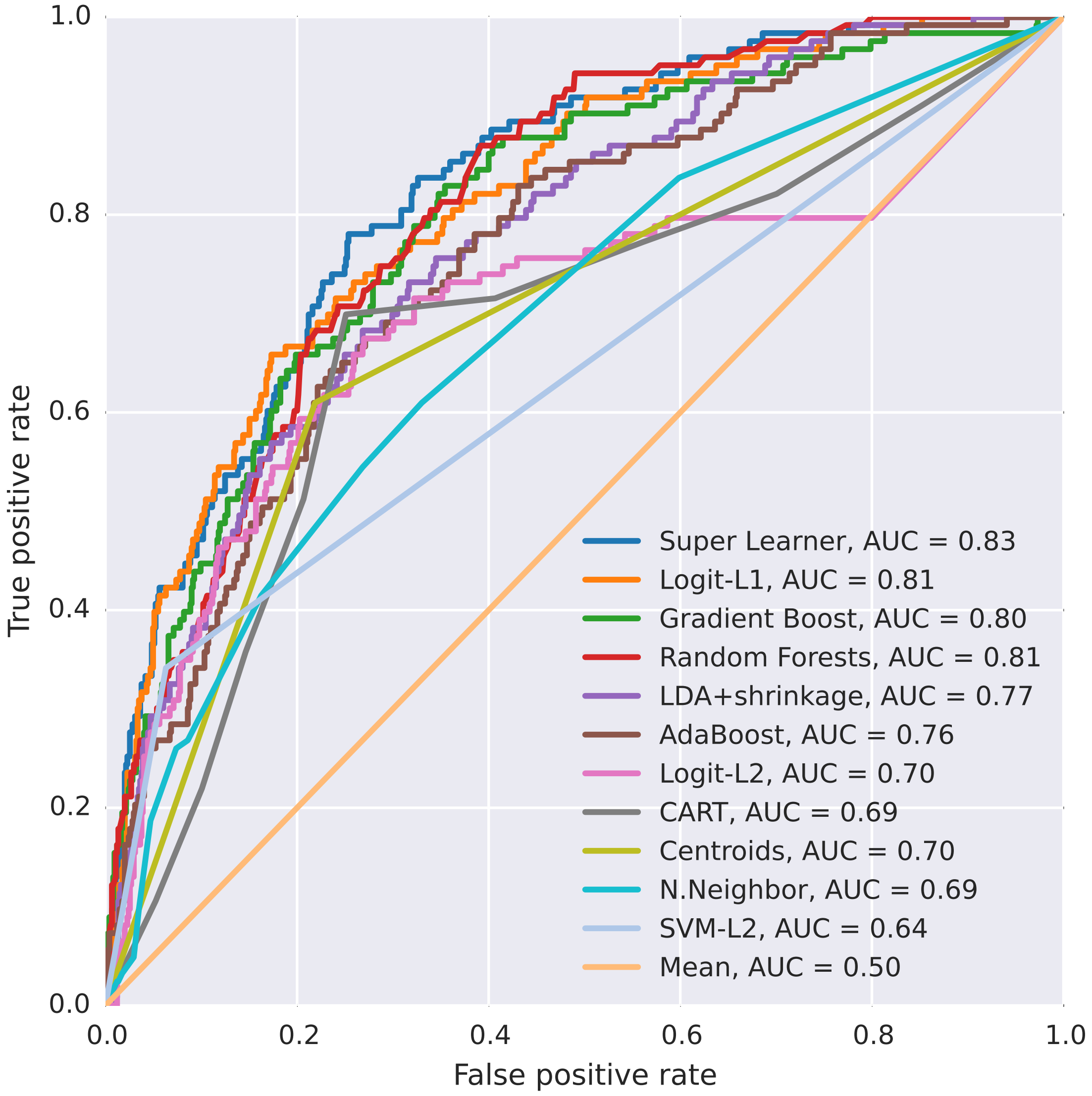

Figure 2.12: Cross-validated ROCs (and corresponding AUCs) of algorithms using all available clinical features, DF vs. DHF/DSS analysis.

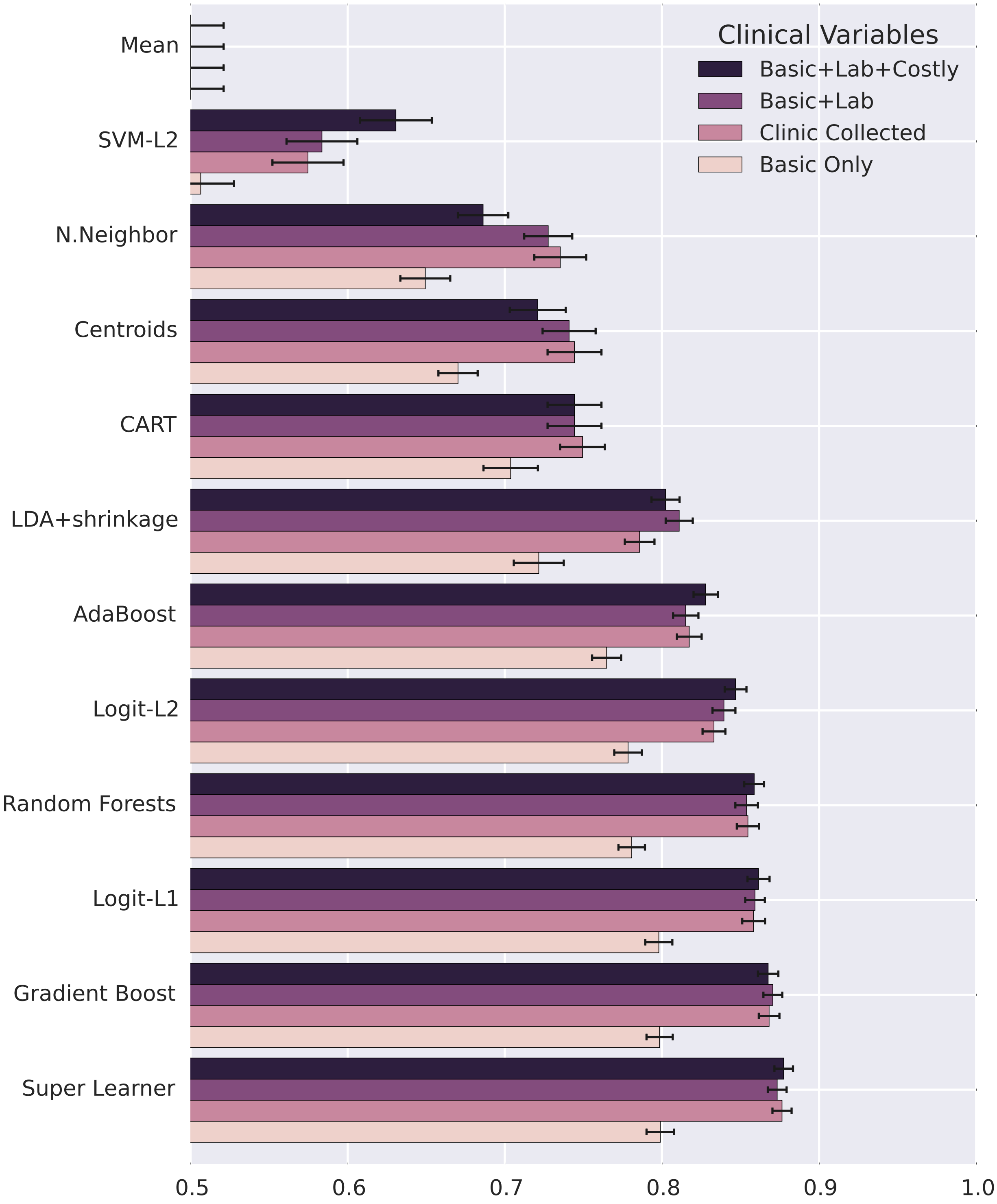

Figure 2.13: Cross-validated AUCs and corresponding 95% confidence intervals for various predictor subsets and algorithms, OFI/DF vs. DHF/DSS analysis.

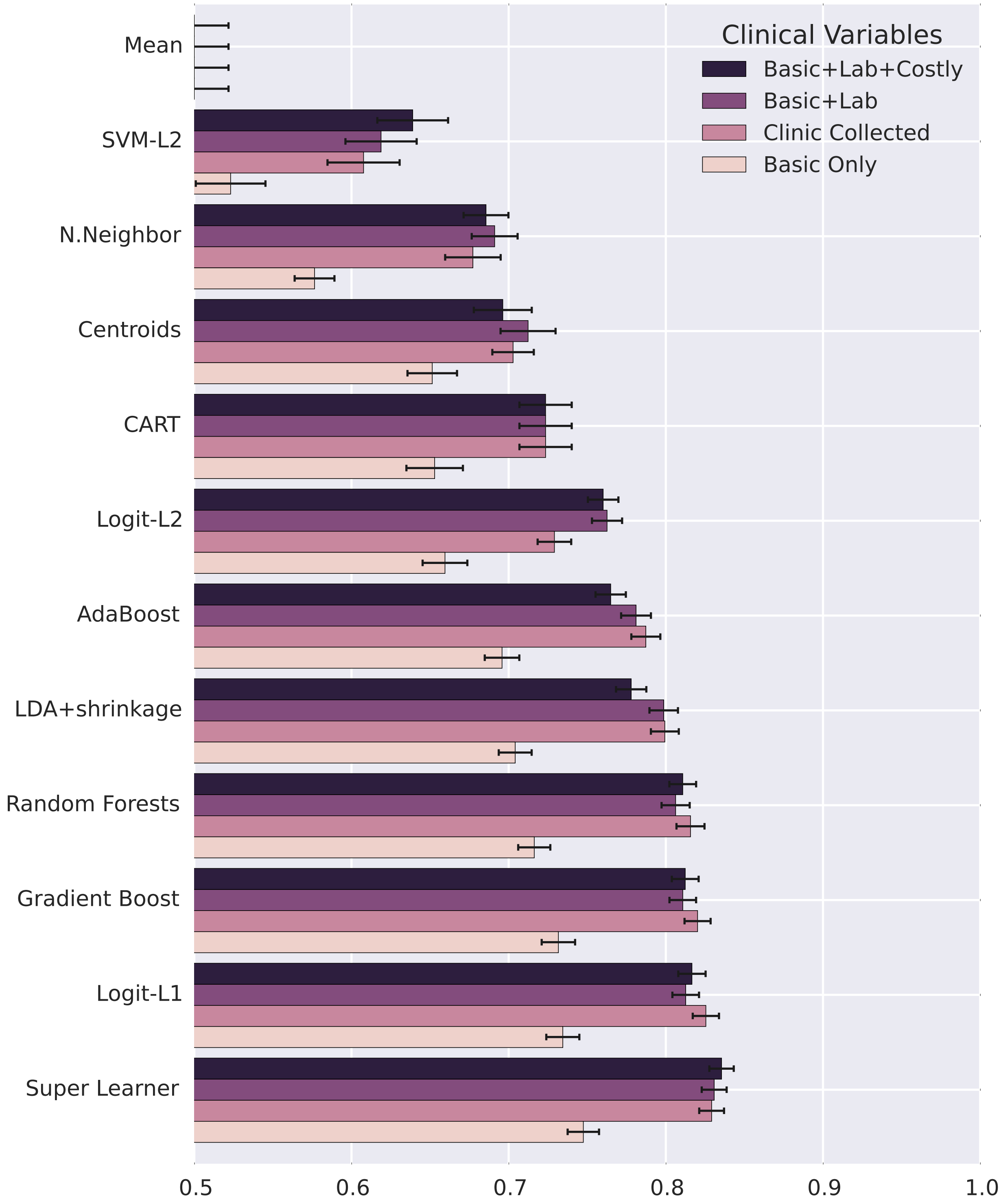

Figure 2.14: Cross-validated AUCs and corresponding 95% confidence intervals for various predictor subsets and algorithms, DF vs. DHF/DSS analysis.

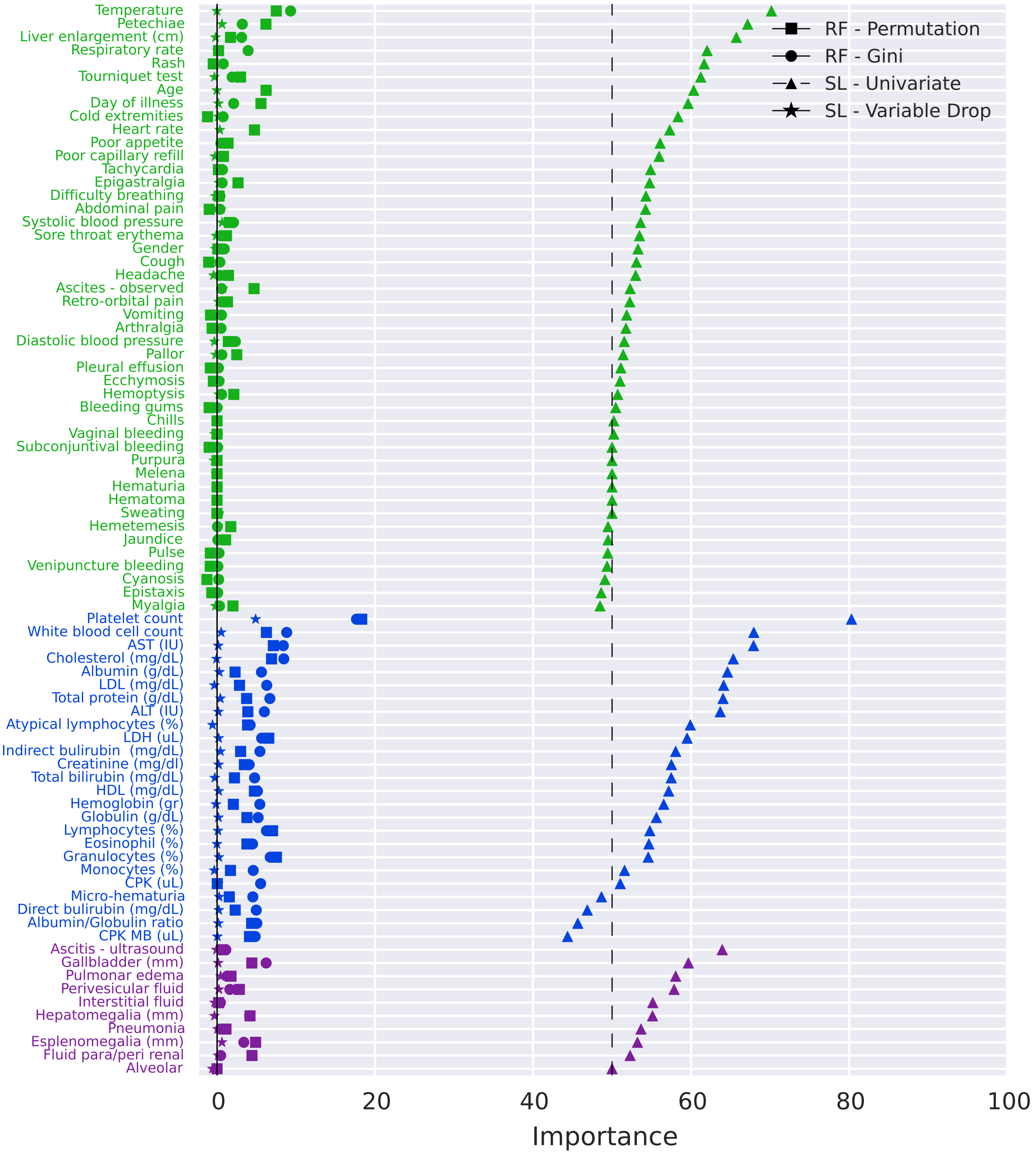

Figure 2.15: Variable importance measures for OFI/DF vs. DHF/DSS analysis. Basic variables appear in green, lab variables in blue, and costly variables in purple. Cross-validated AUC values (for super learner methods) have been scaled by 100, with 50 representing “no importance” for the univariate analysis (dotted line) and 0 representing “no importance” for other methods (solid line).

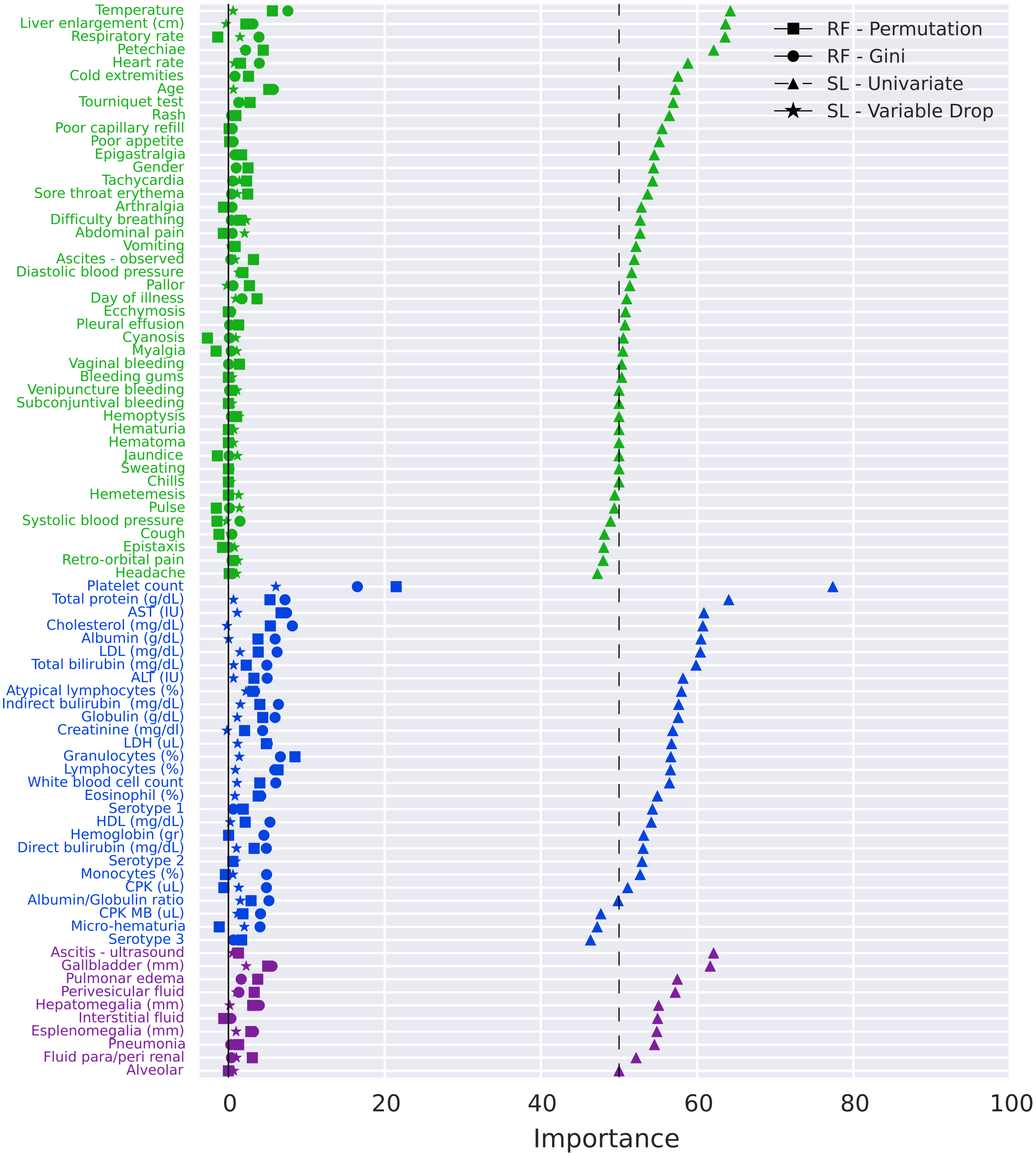

Figure 2.16: Variable importance measures for DF vs. DHF/DSS analysis. Basic variables appear in green, lab variables in blue, and costly variables in purple. Cross-validated AUC values (for super learner methods) have been scaled by 100, with 50 representing "no importance" for the univariate analysis (dotted line) and 0 representing "no importance" for other methods (solid line).

2.5 Conclusions

We have tested our ability to diagnose dengue and to predict dengue's severe manifestations across an array of model specifications. We have found that clinical information is strongly predictive of dengue diagnosis and of severe dengue prognosis with complex algorithms yielding significantly better results than simpler algorithms (e.g., Super Learner versus CART). Expensive clinical information (e.g., ultrasound and X-ray measurements) does not add much information beyond that provided by less expensive clinical data. In contrast, moderately expensive lab results do significantly improve predictions (particularly white blood cell count for diagnosing dengue and platelet count for predicting severe dengue).

Chapter 3

Feature Extraction and Prediction using LC-MS Data

This chapter explores methods for using Liquid chromatography – mass spectrometry (LC-MS) data to diagnose suspected dengue patients. The assumption here is that we have unrestricted access to LC-MS data and wish to explore its full potential. (In Chapter 4 we will restrict ourselves to using a small subset of LC-MS features.)

3.1 Background

Liquid chromatography – mass spectrometry is an analytical chemistry technique that combines the physical separation capabilities of liquid chromatography (LC) with mass analysis capabilities of mass spectrometry (MS). It is a powerful technique for the identification of a wide range of molecules, including proteins, lipids, salts, and metabolites [45, 35]. Metabolomics and lipidomics have already proven useful for diagnostic purposes in a variety of settings [14, 18, 69], but there is still much to be learned through the application of this field of research to dengue fever.

LC-MS laboratory procedure

There are six parts to the typical LC-MS laboratory process: quenching, extraction, reconstitution, chromatography, ionization, and assessment of mass. Here we describe each of these steps. Afterwards, we will describe methods for analyzing and interpreting the resulting data.

1. Liquid chromatography:

(a) Quenching: The biological sample is quenched to stop the metabolic pathways of the cells from continuing to function in order to capture the desired cell state for analysis. Quenching is generally done using a mixture of water and methanol.

(b) Extraction: The sample is mixed with a solvent and then run through a centrifuge to separate out insoluble material, which is thrown away. The resulting liquid is kept for analysis. Note that the particular solvent used influences the type and quantity of compounds present in the resulting sample. Sometimes a two-phase extraction process, involving two different solvents, is used.

(c) Reconstitution: A solvent is added to the liquid resulting from the extraction process. This solvent, known as the *mobile phase*, is chosen to facilitate smooth traveling through the LC elution column.

(d) Chromatography: The reconstituted sample is sent through the LC elution column, which itself contains a solvent known as the *stationary phase*. The purpose of this step is to separate the molecules, thus reducing the total number of analytes entering the mass spectrometer at a given time, which in turn reduces the competition for electric charges during the ionization stage and results in higher mass-spectrometry resolution. The *retention time* (RT), also known as *elution time*, refers to the time it takes for an analyte to pass through the LC system, starting from the column inlet and ending with the mass spectrometer. An analyte's retention time is influenced by its physiochemical characteristics and is one of the inputs for compound identification. Specifically, retention time is determined by a combination of molecular size, charge, hydrophobicity, and specific binding interactions. We can categorize modern LC into two sub-classes based on the stationary phase and corresponding required polarity of the mobile phase:

    i. *Reverse phase liquid chromatography* (RP-LC) is best suited for detecting mildly polar compounds. It uses octadecylsilyl and related organic-modified particles as stationary phase with pure or pH-adjusted water-organic mixtures such as water-acetonitrile and water-methanol for the mobile phase.

    ii. *Normal phase liquid chromatography* (NP-LC) is best suited for detecting highly polar compounds. It uses materials such as silica gel as stationary phase with neat or mixed organic mixtures for the mobile phase. This type of chromatography is also known as *hydrophilic interaction chromatography* (HILIC).

2. Mass spectrometry:

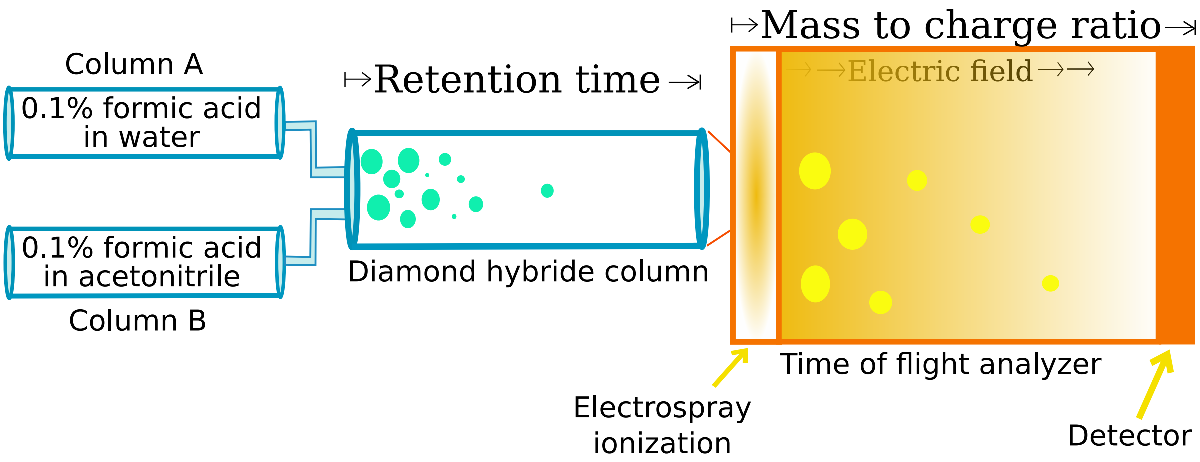

Figure 3.1: Summary of LC-MS laboratory procedure. In the liquid chromatography step (illustrated in blue), the time an analyte takes to travel through the diamond hybrid column (or another type of LC column) is recorded as the retention time. In the mass spectrometry step (illustrated in orange), the time taken for the ionized analyte to travel through the electric field maps to the analyte's mass-to-charge ratio.

(a) Ionization: Mass spectrometers work by converting the analytes coming out of the LC elution column to a charged (*ionized*) state. There are many techniques for ionization, with *dual electrospray ionization* (ESI) being the most widely used ion source for biological molecules, though neutral and low polarity molecules such as lipids may not be efficiently ionized by this method. When operated in positive ion mode, ESI will charge molecules by adding protons; while operated in negative ion mode, molecules are charged by removing protons.

(b) Assessment of mass: Ions and fragment ions produced during the ionization step are analyzed using one of several techniques. A time-of-flight (TOF) analyzer works by accelerating ions through a high voltage. The velocity of the ion, and hence the time taken to travel down a flight tube to reach the detector, indicates the *mass-to-charge ratio* (also known as the *m/z value*) of the ion.

A simplification of the above process is illustrated in Figure 3.1.

LC-MS Data interpretation

The above steps produce a large vector of tuples consisting of elution time, mass-to-charge ratio, and corresponding abundance. From here, we would ideally like to estimate the abundance levels of small molecule biomarkers (SMBs). Our task would be relatively simple if each SMB in our biological samples were to consistently map to a unique m/z - RT combination in a one-to-one fashion. But this is not the case. Instead, different molecules can map to the same m/z and RT values (within the window of instrument measurement precision), resulting in multiple possible identities for a given observed (m/z, RT) value [78]. Additionally, isotopes, adducts, charge state dispersion, competitive ionization, and degradation products, which we describe below, further complicate the mapping from observed molecular features (MFs) to the identities and quantities of SMBs in the original samples [28].

1. **Isotopes**: Atoms of a given chemical element can have different numbers of neutrons, thus resulting in slightly different mass-to-charge ratios.

2. **Adducts and clusters**: Two or more distinct molecules can combine to form adducts and clusters. For example, adduction of cations (e.g. M+NH4+, M+Na+, M+K+) and anions (e.g. M+formate-, M+acetate-) can occur when salts are present.

3. **Charge state dispersion**: Larger molecules and molecules with several charge-carrying functional groups such as proteins and peptides can exhibit multiple charging, resulting in ions such as M+2H2+, M+3H3+ etc. Due to charge state dispersion, one peptide species may register peaks at a series of m/z locations, with the distribution across m/z locations not well-characterized.

4. **Competitive ionization**: The abundance of some SMBs will appear suppressed when in the presence of SMBs that compete for ions. This could be a serious problem when it comes to translating the results of an LC-MS analysis to a point of care diagnostic (POC) test; we may think a particular SMB is suppressed in dengue patients when it in is fact present in equal amounts across patients.

5. **Degradation products**: When exposed to stress brought by conditions such as as changes in temperature or pH, compounds can degrade, resulting in degradation products.

In Chapter 3.3.1 we discuss approaches for translating raw LC-MS data into a profile of SMBs. Due to the above complications, this process will not result in a definitive SMB list, but rather a list of possible SMBs.

Literature on using LC-MS for dengue diagnostics

Prior studies have shown lipids to facilitate *Flaviviridae* viral entry, replication, and release [22, 51, 77], but few metabolomic studies exist that describe the human host response to

DENV [4, 12], and none to the author's knowledge have used LC-MS data with machine-learning algorithms to develop predictive models with validated performance measures for dengue diagnostics.

3.2 Data description

Sample collection

Biological samples (urine, saliva, and serum) were collected from a subset of clinic and hospital patients, with a preference for patients who arrived to the health facilities early in their disease progression and for patients who developed severe dengue. In all, we ran normal phase LC-MS (procedure described below) on 88 serum samples, 85 saliva samples, and 80 urine samples. In addition, reverse phase LC-MS (also described below) was run on 90 serum samples. These sample sizes are large enough to be suggestive of LC-MS's power to distinguish patient disease states but too small to achieve definitive results.

About a third of the serum LC-MS samples came from patients who were eventually determined to have an other febrile illness (OFI), about a third came from patients were were diagnosed with non-severe dengue fever (DF), and about a third came from patients eventually diagnosed with dengue hemorrhagic fever or dengue shock syndrome (DHF or DSS). Just over half of the serum samples that came from severe dengue patients (those with DHF or DSS) were from patients who were already displaying symptoms of severe dengue at the time of sample collection (Tables 3.1 and 3.2).

Initial Diagnosis	Final Diagnosis				Overall
	OFI	DF	DHF	DSS	
Suspected DF	29	29	8	5	71
Suspected DHF	0	0	7	5	12
Suspected DSS	0	0	0	5	5
Overall	29	29	15	15	88

Table 3.1: Serum samples used in normal phase LC-MS analysis for which final diagnosis is known.

The saliva and urine samples came primarily from DF patients. Of the 85 saliva samples, 48 are from DF patients, 30 are from OFI patients, and 7 are from DHF/DSS patients. Similarly, of the 80 urine samples, 44 are from DF patients, 29 are from OFI patients, and 7 are from DHF/DSS patients. With so few samples from severe dengue patients, we will not be able to run a meaningful analysis on identifying severe dengue using saliva and urine

Initial Diagnosis	Final Diagnosis				Overall
	OFI	DF	DHF	DSS	
Suspected DF	32	30	9	3	74
Suspected DHF	0	0	5	3	8
Suspected DSS	1	0	0	7	8
Overall	33	30	14	13	90

Table 3.2: Serum samples used in reverse phase LC-MS analysis for which final diagnosis is known.

samples, though we will proceed with an analysis to distinguish DENV (which we define to include DF, DHF, and DSS) from OFI using these samples.

Despite efforts to collect samples early in the disease progression, about two-thirds of our serum samples and about half of our saliva and urine samples were collected more than 72 hours after fever onset. It should also be noted that all of our DENV serum samples were of serotype 2 so we will need to exercise caution in extrapolating results to other serotypes. Meanwhile, most DENV saliva and urine samples are of serotype 1, with a handful of serotypes 2 and 3.

Laboratory details

Different LC-MS laboratory procedures can result in very different molecular discoveries (e.g., some procedures work best for lipid detection and others for sugar detection). The procedures used in this study are optimized for the detection of metabolites – the end products of cellular regulatory processes – while also allowing for the discovery of some lipids.

Normal phase LC-MS: Serum samples were added to cold 100% methanol in a 1:3 ratio, vortexed for 1 minute, incubated at -20C for 20 minutes, and then centrifuge at 14,000 rpm for 20 minutes. The resulting supernatant was transferred to a new vial, dried using a speed vacuum at room temperature for 45-60 minutes, and reconstituted in 100% acetonitrile. The resulting mixture was then incubated at room temperature for 10 minutes, vortexed for one minute, and centrifuged at 4C for 5 minutes at 14,000 rpm. Finally, 25μl of the supernatant was transferred to a glass vial for LC-MS analysis.

Saliva samples were thawed at room temperature. Once in liquid state, samples were centrifuged at 14,000 rpm for 20 minutes at 4C, and then 50μl of supernatant were collected and placed in a new vial. 100μl of acetonitrile was added, and the vial was vortexed for 1 minute and placed at -20C for 10 minutes in order to precipitate proteins present in the sample. It was centrifuged at 14,000 rpm at 4C for 5 minutes. Finally, 25μl of the supernatant

Time	Gradient
0.2 - 30 minutes	95% B to 50% B
30 - 35 minutes	50% B
35 - 40 minutes	50% B to 20% B
40 - 45 minutes	20% B to 95% B

Table 3.3: Liquid-chromatography gradient used for normal-phase analysis. Solvent A consisted of water with 0.1% formic acid; solvent B consisted of acetonitrile with 0.1% formic acid.

Scan rate	1.4 spectrum per second
Capillary voltage	4000 V
Drying gas (N_2)	235C at 10 L/min
Nebulizer pressure	45 psi
Fragmentor	150 V
Skimmer	65 V
OctopoleRFPeak	750 V
Capillary pump	Flow 40μl/min, pressure 400 bar
Binary pump	Flow 0.4 ml/min, pressure 400 bar
Mass range	100-1700 Da
Calibrated	<2 ppm mass accuracy

Table 3.4: Description of time of flight mass spectrometry parameter values, normal phase analysis.

was transferred to a glass vial for LC-MS analysis.

Urine samples were normalized prior to analysis using a technique based on expected creatinine levels in order to control for variation in patient hydration levels. After normalization, 50 μl of the resulting mixtures were vortexed and centrifuged at 4C for 20 minutes at 14,000 rpm. Finally, 25 μl of the supernatant were transferred to a glass vial to be analyzed by LC-MS.

A Cogent hydrophilic high performance liquid chromatography (HPLC) column type-C silica diamond-hydride was used in conjunction with an Agilent 6520 Quadrupole time of flight mass spectrometer. The chromatography gradient specified in table 3.2 was used with 20 minute column re-equilibrations between runs. The column was tuned, calibrated, and cleaned after every 15 injections.

The mass spectrometer was coupled with dual electrospray ionization operated in positive ion mode with a mass range of 100-1700 m/z calibrated to less than 2ppm mass accuracy. The acquisition time spanned 45 minutes. Table 3.4 contains additional details regarding the mass spectrometer set-up.

Reverse phase LC-MS: Serum samples were extracted with 100μl of cold methanol and vortexed for 15 seconds. The samples were then incubated for 12 hours at -80C, followed by 30 minute centrifugation at 18,000xg. The supernatant was transferred to a new tube and dried in a vacuum concentrator. Extracted samples were resuspended in 40μl of 50% methanol, let stand at room temperature for 15 minutes, vortexed for 20 seconds, and centrifuged for 30 minutes at 18,000xg to remove insoluble debris. The supernatant was transferred to HPLC vials and stored at -80C prior to LC/MS analysis.

Samples were injected in randomized order on an Agilent 1290 HPLC system coupled to an Agilent 6224 time-of-flight mass spectrometer [Santa Clara, CA, USA]. Samples were run in positive ion mode. An Atlantis T3 C18 column (3μm particle size, 2.1x150mm, Waters Millford, MA, USA) at a flow rate of 250 μl/min was used. The mobile phase was composed of (A) water plus 0.1% formic acid and (B) 95% acetonitrile with 0.1% formic acid. A linear gradient elution from 0% to 98% of substance B (spanning 0-13 minutes) and 98% of substance B (spanning minutes 13-14) was applied. ESI source conditions were set as follows: gas temperature 325C, drying gas 5 L/min, nebulizer 20psi, fragmentor 120 V, skimmer 50 V, and capillary voltage 4000 V. Mass spectral data were collected in profile and centroid mode. The instrument ran in 2 GHz extended dynamic range mode with a range of 85 to 1700 m/z at a scan rate of 2 spectra/second.

Raw LC-MS data

Our normal phase laboratory procedure resulted in about 400 million intensity values per patient sample, which can be visualized on a grid with axes specifying retention time and mass-to-charge ratios (m/z values), as shown in Figure 3.2. Figure 3.3 summarizes raw LC-MS data obtained from six different serum samples by showing the maximum intensity value across the entire range of masses detected at each retention time.

3.3 Methods for extracting features from LC-MS data

We converted the raw HILIC data from machine binary format into mzML format using `msConvert` [10] and read the resulting mzML files using the `pymzML` [1]. To analyze this data, we employ biologically motivated methods that are specific to the LC-MS technology, as well as more general methods that could be applied to a variety of datasets. The advantage of the former methods is that they lend themselves to compound identification. The latter methods could be advantageous if prediction accuracy with LC-MS data is our goal, rather than the identification of compounds that do a good job discriminating between samples.

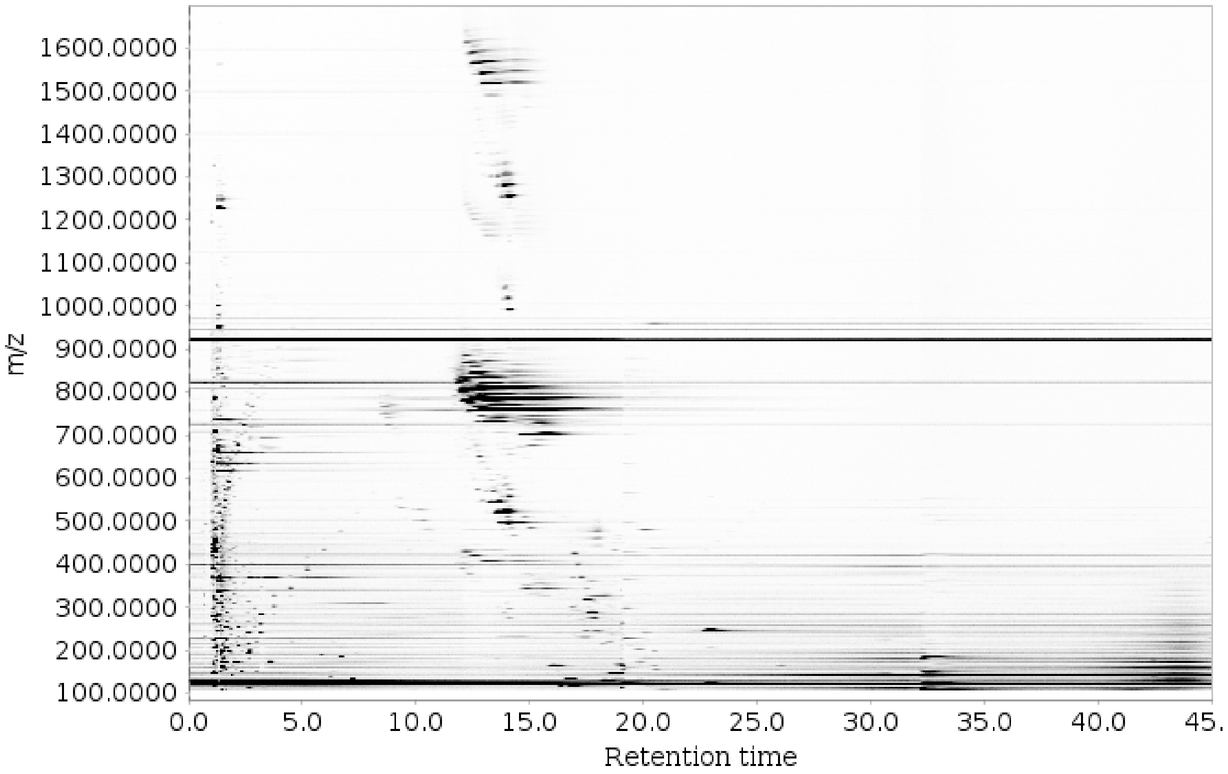

Figure 3.2: Visual representation of HILIC data from one serum sample. Color intensity indicates quantity of small molecules with the corresponding m/z and retention time values.

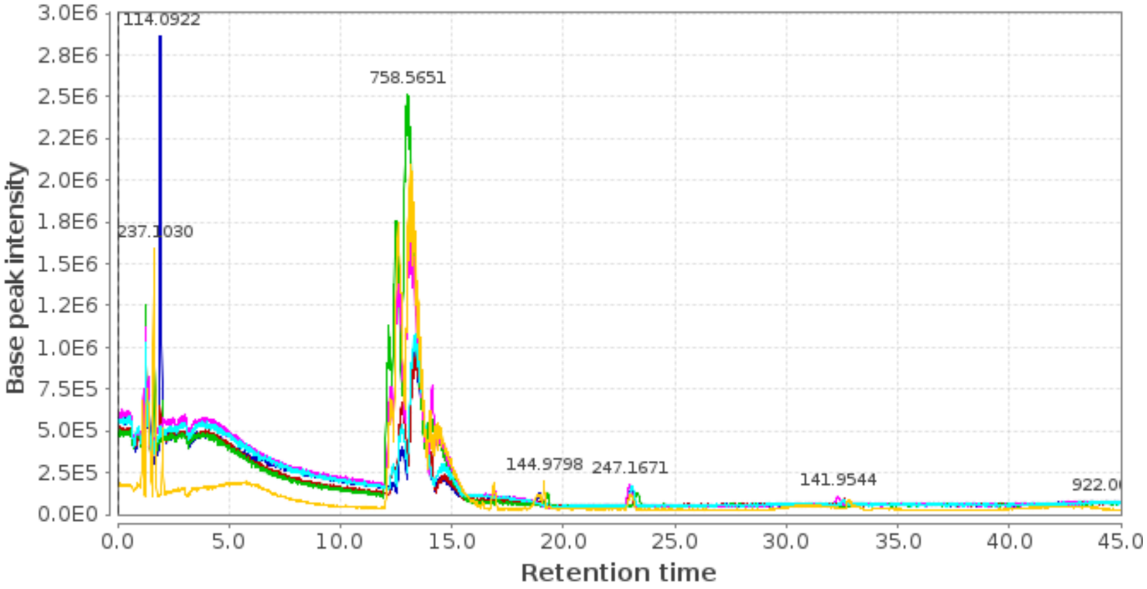

Figure 3.3: Chromatogram representation of raw LC-MS data. Each color represents a different serum sample.

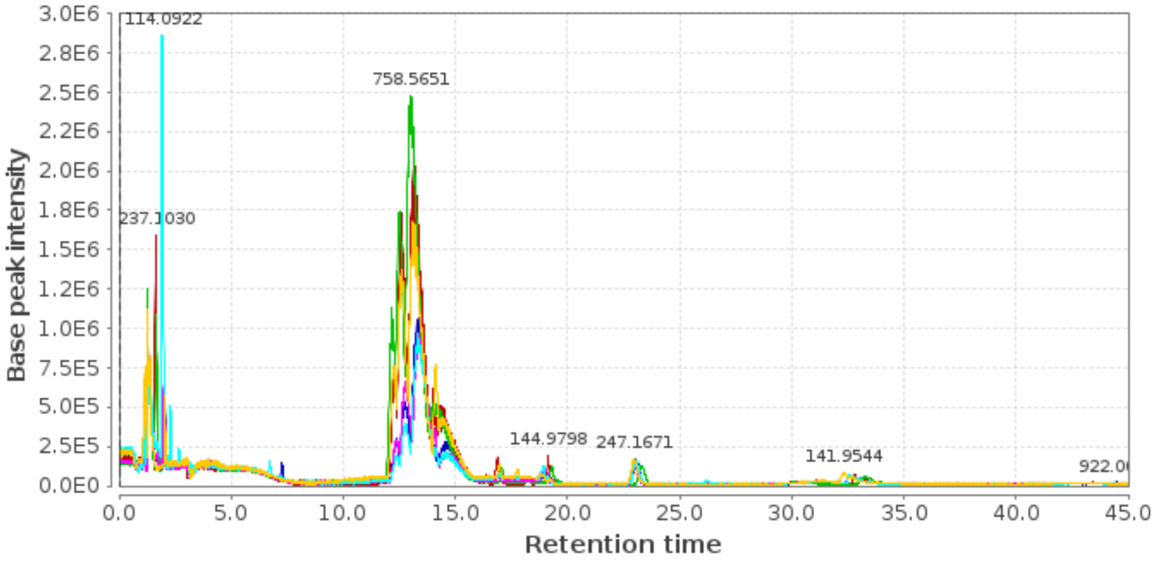

Figure 3.4: Chromatogram representation of the LC-MS data displayed in Figure 3.3 after baseline correction.

3.3.1 Industry standards: MZmine, Mass Hunter, and XCMS

`Mass Hunter` (provided by Agilent Technologies), `MZmine` [46, 26], and `XCMS` [55] are software tools that use LC-MS data to surmise which small molecule biomarkers (SMBs) are present in a given sample, and at what relative concentration. Each of these tools implements the following basic procedure, the specifics of which are determined according to user specifications, which we specify below. Note that "scan" and "spectrum" refer to the m/z and intensity values corresponding to one single retention time value.

Step 1: Baseline correction

The baseline correction is meant to correct intensity values by compensating for gradual shifts (also known as "warping") in the chromatographic baseline over time [6]. MZmine implements asymmetric least squares to estimate the baseline for each scan using the R package "ptw". This involves finding an optimal polynomial to describe the warping by assigning different weights to the data points that are above and below an iteratively estimated trendline. A smoothing parameter must be specified by the user. Using the total ion count chromatogram, the corrected intensity values are the maximum of 0 and $x_{original}(1 - \frac{x_{base}}{x_{max}})$. Figure 3.4 illustrates the same data shown in Figure 3.3 after implementing a baseline correction.

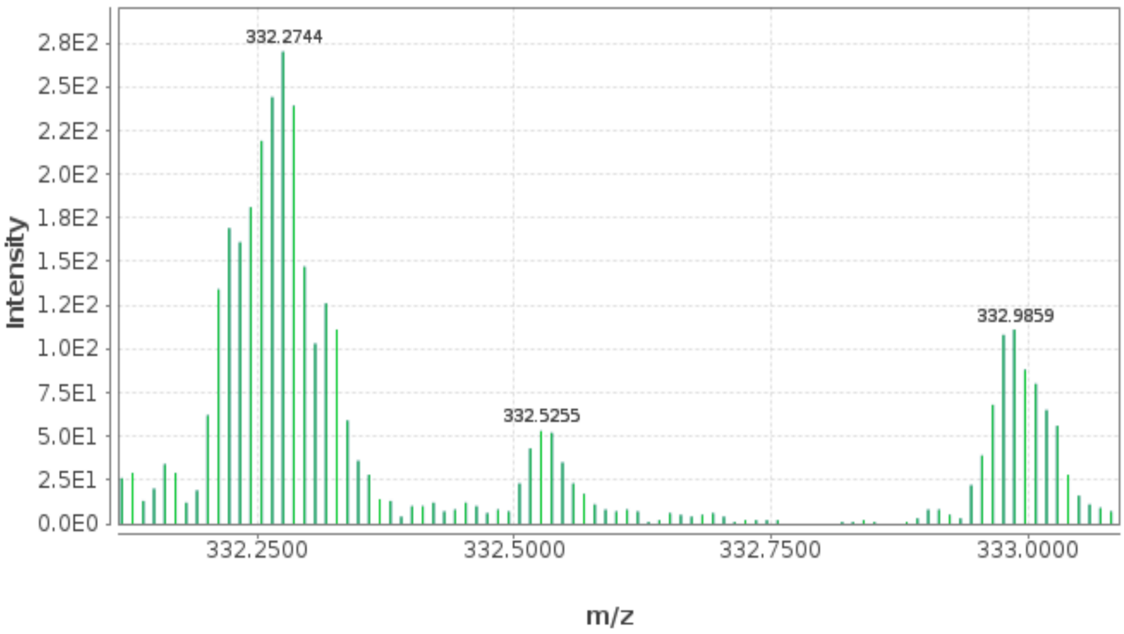

Figure 3.5: Close-up of data (for one serum sample) after combining intensities of nearby m/z values within each scan using a moving average filter.

Step 2: Filtering

The purpose of the filtering step is to remove noise by combining intensities of nearby m/z values within each scan. Filtering options include: (i) moving average filter, (ii) Savitzky-Golay filter, which involves performs a local polynomial regression to determine the smoothed value for each point, and (iii) m/z resample filter. The m/z resample filter involves segmenting each scan into m/z bins with widths defined by the user. The mass of the new data point will be in the middle of each m/z bin's space and its intensity is the average of the intensity of all the data points inside the bin.

Step 3: Peak detection

Peak detection involves further organizing our data into peak lists. To do this, we may conduct three sub-steps: mass detection, chromatogram building, and chromatogram deconvolution:

a. Mass detection The mass detection step results in a list of pairs of m/z and intensity values for each scan. Options for this step include:

- Centroid mass detection: each signal above user-specified noise level is considered a

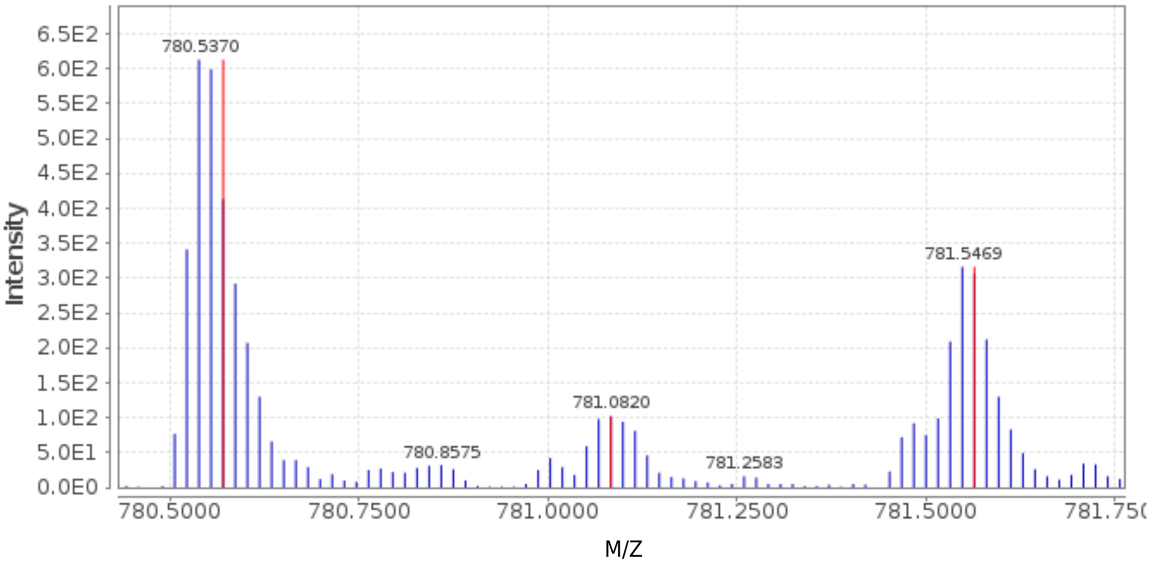

Figure 3.6: Close-up of data (for one serum sample) after running the recursive threshold peak detection algorithm. Identified peaks are colored red.

detected ion.

- Local maximum method: every local intensity maximum along the spectrum is considered a spectral peak provided it is above the user-specified noise level.

- Recursive threshold: user specifies the minimum and maximum allowable widths of m/z peaks, as well as the minimum intensity level for a data point to be considered part of the chromatogram (as opposed to noise). This algorithm initially looks at the whole range of data points. If the m/z width of this range in not within the given width limits, a minimum data point is found and used to split the range into two parts. The same algorithm is then applied recursively on each part. Recursion continues until all m/z ranges that fit into the given width limits are found. Final m/z values are determined as local maxima of the identified m/z ranges. Figure 3.6 illustrates the result of this recursive threshold method for mass detection.

- Wavelet transform mass detector: the Gaussian derivative wavelet, Mexican Hat wavelet, or an alternative wavelet is fit to the data. The final m/z value of the ion is generally calculated as an average of m/z values of surrounding data points weighted by their intensity.

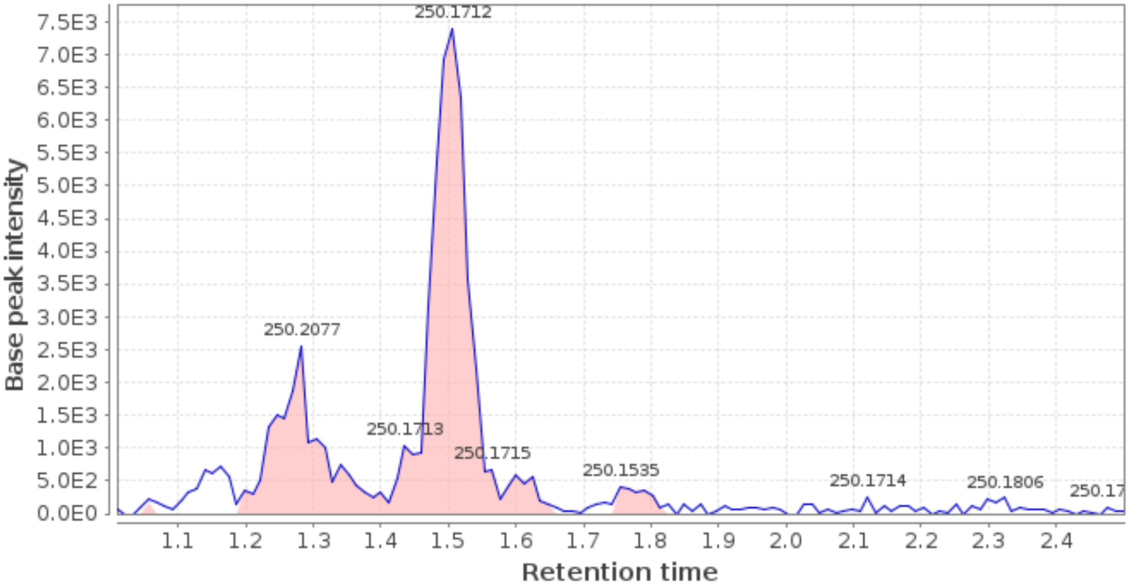

Figure 3.7: Example chromatogram built by combining each scan's peak list using the described chromatogram building method. Contiguous pink regions are each considered to be one peak (before deconvolution).

b. Chromatogram building Here, we process each scan's peak list (i.e., the peaks found in the "mass detection" step), beginning with the scan with the lowest RT value. Each scan's peaks are processed in the order of decreasing intensity. The current one-dimensional peak (which has an associated m/z, rt, and intensity value) is compared to the spectral peaks of the previous scan, and if it is within the m/z tolerance level of one of these peaks, then it is connected to it. If no matching m/z value is found, then a new chromatogram is created with the given m/z value. The connected peaks form two-dimensional strings of spectral peaks. When all scans have been processed, the chromatograms that do not meet the user-specified minimum time span and intensity requirements are eliminated from the final peak list. A typical result of the chromatogram building process is illustrated in Figure 3.7.

c. Chromatogram deconvolution Chromatogram building generally results in peak combinations that ought be split into multiple peaks in order to more closely correspond to distinct SMBs. We have the following method options for implementing this deconvolution:

- Baseline cut-off deconvolution: after removing the lowest part of the chromatogram according to a user-specified baseline level, remaining peaks are recognized if they meet the minimum acceptable height (intensity) for a chromatographic peak and have an RT width that is neither too narrow nor too wide.

- Noise-amplitude deconvolution: similar to the baseline cut-off method, but the baseline level is set individually for each chromatogram as follows: divide chromatogram into bins of user-specified size and set the baseline intensity to the average intensity of the bin with the highest number of data points.
- The “Savitzky-Golay” algorithm: the borders of individual peaks are detected using the smoothed second derivative of the chromatogram curve.
- Local minimum search: individual peaks are separated at local minimal points.
- Wavelets: briefly, a series of wavelets with different scales is fit to the chromatogram. Local maxima in the convolution results determine the locations of possible peaks. When these candidate peak locations co-occur at multiple scales then the scale with the strongest response indicates peak width. Given the candidate peak locations and scales, peaks can then be reconstructed from the original chromatogram.[1]

After mass detection, peak convolution and deconvolution, each LC-MS run i ($i = 1, ...N$) can be represented by a peak list:

$$P_i = \{P_{ijc}\};\ j = 1, ..., N_i,\ c = \{mz, \delta mz, rt, \delta rt, height, area\}$$

where N_i is the total number of peaks for run i and c is an index for values of each peak p_{ij}: mz is the mean m/z value for data points within the peak, δmz is standard deviation of m/z values within the peak, rt is retention time at the maximum intensity data point, δrt is the lengths of the peak in time, height is the height of the peak, and area is the area of the peak. Either a peak area or height can be used in further processing stages.

Step 4: Isotopic peak grouper

We can optionally run an isotopic peak grouper in order to filter out peaks that are likely isotopes of other peaks. When an isotope pattern is found, only the highest isotope is kept. More specifically, this procedure is implemented by processing the peak list in the order of decreasing height. For each peak, we try to find the most appropriate charge state by comparing the number of identified isotopes for each possible charge. For each charge state, peaks that fit the m/z and RT distance limits are considered as isotopes. The charge

[1] Tautenhahn’s centWave [61], Du’s continuous wavelet transform method [15], and Nguyen’s GDWavelet [42] are all wavelet-based methods for peak detection. Unlike some of the other algorithms, these methods provide some amount of shape-matching, which is appealing since ’true’ peaks are believed to have characteristic *shapes*, not just characteristic heights and widths [15].

state with the highest number of identified isotopes is selected, and the isotope pattern is generated. Recall that the difference between neighboring isotopes is a single neutron. The exact mass of 1 neutron is 1.008665 Da, but part of this mass is consumed as a binding energy to other nucleons so the actual mass difference between isotopes depends on the chemical formula of the molecule. Since MZmine does not know the formula at the time of deisotoping, it assumes the default distance of 1.0033 Da, with user-defined tolerance (the m/z tolerance parameter).

Step 5: Peak alignment and gap filling

Peaks are matched across LC-MS runs through a peak alignment procedure. This step will result in an aligned peak list with a column for each LC-MS run and a row for each peak that was matched in one or more of the original peak lists. Inside the cells will be intensity values corresponding to the max height and/or area under the intensity curve. There are several options for peak alignment:

- The "join aligner" begins by taking one peak from the peak list of one LC-MS run and finding the closest match in the master peak list, which starts as an empty list. If no good match is found (as determined by user-specified thresholds), then the peak is added as a new entry in the master list. The closeness of a match between a candidate peak and a peak from the master list (which may itself be composed of what had been multiple distinct peaks) is defined as: $k(mz_{candidate} - mz_{mean\,of\,peaks}) + (rt_{candidate} - rt_{mean\,of\,peaks})$, where k is a tuning parameter usually set to a large number to reflect the fact that peaks from the same compounds usually match closely in m/z values but often have substantial RT variation.

- The "RANSAC" (RANdom SAmple Consensus) algorithm, unlike the join aligner, is able to handle non-linear deviation of the retention times among samples. For details of MZmine's implementation of RANSAC, see Pluskal 2010 [46].

The master peak list will inevitably have some empty cells (occurring when an LC-MS run has no peak where other runs have a peak), which one may elect to fill. One of the simplest methods to fill these gaps consists of taking the maximum intensity value from the original data that falls within a user-specified range of the missing peak's rt and m/z values; if there is no registered intensity value within the specified range, then an intensity of zero is typically assumed.

Step 6: Normalization

The purpose of normalization is to eliminate intensity value variation between runs that is unrelated to the biological processes of interest, and to convert multiplicative noise into additive noise (variance-stabilizing transformation). Some options for normalization are as follows. Note that each relies on a set of assumptions regarding the nature of measurement and biological variability, and some of these methods can be used in combination with one another.

- Total intensity normalization: This method forces all samples to have equal total intensity by dividing each intensity value by a sample-specific normalization factor. The assumption that the total concentration of metabolites does not vary across biological samples is unlikely to be met, however, particularly for urine samples. This is thus not a method that we will employ.

- Linear normalization: All peak heights are divided by (a) average peak height, (b) average squared peak height, (c) max peak height, or (d) total raw signal. Each column (raw data file) of the peak list is normalized separately. In other words, the normalization factor is determined independently for each raw data file.

- Median fold change normalization: The median of the log fold changes of peak intensities between samples is set to be approximately zero. In other words, we can choose a target profile, $p_{i'}$, of intensity values $p_{i'1}...p_{i'J}$, and then scale the intensities of each sample $i* = 1, ..., N-1$ such that median$(log(\frac{p_{i*1}}{p_{i'1}}), ..., log(\frac{p_{i*J}}{p_{i'J}})) \approx 0$. This procedure is justified if we believe the reasonable assumption that peaks which differ in intensity level from sample to sample purely due to the effect of dilution will exhibit the same fold changes. The dilution factor has been shown to be well estimated through this method [68].

- Quantile normalization: Intensity values are adjusted such that all samples have identical post-adjustment peak intensity distributions, as measured by the chosen quantiles. Thus, this method assumes that unwanted sample-to-sample variation is peak-intensity-dependent and thus applies a scaling factor that varies across the range of peak intensity values.

- Standard compound normalizer: standard compounds are injected into the samples in known concentrations prior to the LC-MS analysis. Normalization can be done by dividing each peak by the height of the peak of the known standard that is closest to the peak needing to be normalized.

Minimum peak height	600 counts
Allowed ion species	+H and +Na
Isotope grouping	performed
Peak spacing tolerance	.0025 m/z; 7.0 ppm
Minimum height for compounds	3000 counts
Peak smoothing	0.2 times peak width
Compound ion threshold	two or more

Table 3.5: Parameter values used for feature extraction with Mass Hunter.

Step 7: Peak Identification

Identification of peaks can be performed either by searching a custom database of m/z values and retention times, or by connecting to an online resource such as PubChem [27], KEGG [25], METLIN [56], or HMDB [72, 73, 74]. “Neutral mass” is the primary term for a database search. Isotopic pattern similarity can be used as a second filter to select optimal candidates, by comparing the ratios of the detected isotopes and matching isotopes from the predicted isotopic pattern of the database compound. We will not bother with peak identification in this chapter, but will discuss it further in Chapter 4.

Summary of industry procedure used with normal phase data

For the normal phase analysis, molecular features were extracted using Mass Hunter’s Qualitative analysis operated with the parameter values specified in Table 3.5 [70]. Features extracted from Mass Hunter were further analyzed using Agilent’s Mass Profiler Pro program (MPP). Using MPP, data was further filtered (using a 5000 count threshold), aligned (using a retention time window width of .25 minutes and an m/z window of 15ppm + 2mDa), and normalized to the 75th percentile. These steps resulted in 15,930 molecular features. Lastly, features were eliminated if they were not found in at least 50% of samples for at least one of our three main diagnostic groups (OFI, DF, and DHF/DSS) [70], which reduced the number of molecular features to 744.

Summary of industry procedure used with reverse phase data

For the reverse phase analysis, the raw data files were processed for peak detection using ProteoWizard MS Convert version 3.0.6478. Retention-time correction, chromatogram alignment and metabolite feature annotation were performed using XCMS software version 1.46 in R version 3.2.2 [55]. Feature extraction parameters were optimized using the IPO software package [37]. The XCMS parameters were set as follows: centWave was used for metabolite feature detection [61] (m/z deviation=26.87 ppm, chromatographic peak width range = 9.92 – 45.5 seconds, signal to noise ratio threshold=5); retention-time correction was performed using the obiwarp algorithm (profStep=0.592); and parameters for chromatogram alignment were: mzwid=0.0614, minfrac=0.5, and bw=0.88. Related isotopic features were grouped using the CAMERA software verson 3.2 [30]. This process resulted in 15,657 molecular features among the serum samples.

3.3.2 Methods from outside of the LC-MS literature

In this section we discuss general methods (not just specific to LC-MS data) that exploit the natural ordering of our data's features (e.g., determined by retention times and mass-to-charge ratios) to reduce the dimension of our raw LC-MS data before proceeding with our various prediction algorithms.

Grid of summed intensities

One of the simplest methods for reducing the dimension of our raw data is to convert the original $H \times H'$ grid into a grid of dimension $G \times G'$ (where $G < H$ and $G' < H'$) by simply summing the intensity values in our raw data that belong to each square on our $G \times G'$ grid (illustrated in Figure 3.8). We can do this for a variety of grid specifications, treating G and G' as tuning parameters that can be fit data-adaptively via cross-validation. The natural extension to this grid of summed intensities method is to smooth the fit across regions.

The fused Lasso for Functional Data

We can conceptualize our raw data features $w_i(h, h')$ for sample i as being ordered according to the index values h and h'. The fused lasso modifies the lasso penalty to take into account the ordering of the features. In our 2-dimensional case it solves

$$\underset{\beta_0,\beta}{\text{argmax}} \sum_{i=1}^{N} \Big[y_i(\beta_0 + \beta^T w_i) - \log(1 + e^{\beta_0 + \beta^T w_i}) \Big]$$

$$- \quad \lambda_1 \sum_{h=1}^{H} \sum_{h'=1}^{H'} |\beta_{h,h'}| \quad - \quad \lambda_2 \sum_{h=1}^{H} \sum_{h'=2}^{H'} |\beta_{h,h'} \quad - \quad \beta_{h,h'-1}| \quad - \quad \lambda_3 \sum_{h=2}^{H} \sum_{h'=1}^{H'} |\beta_{h,h'} \quad - \quad \beta_{h-1,h'}|$$

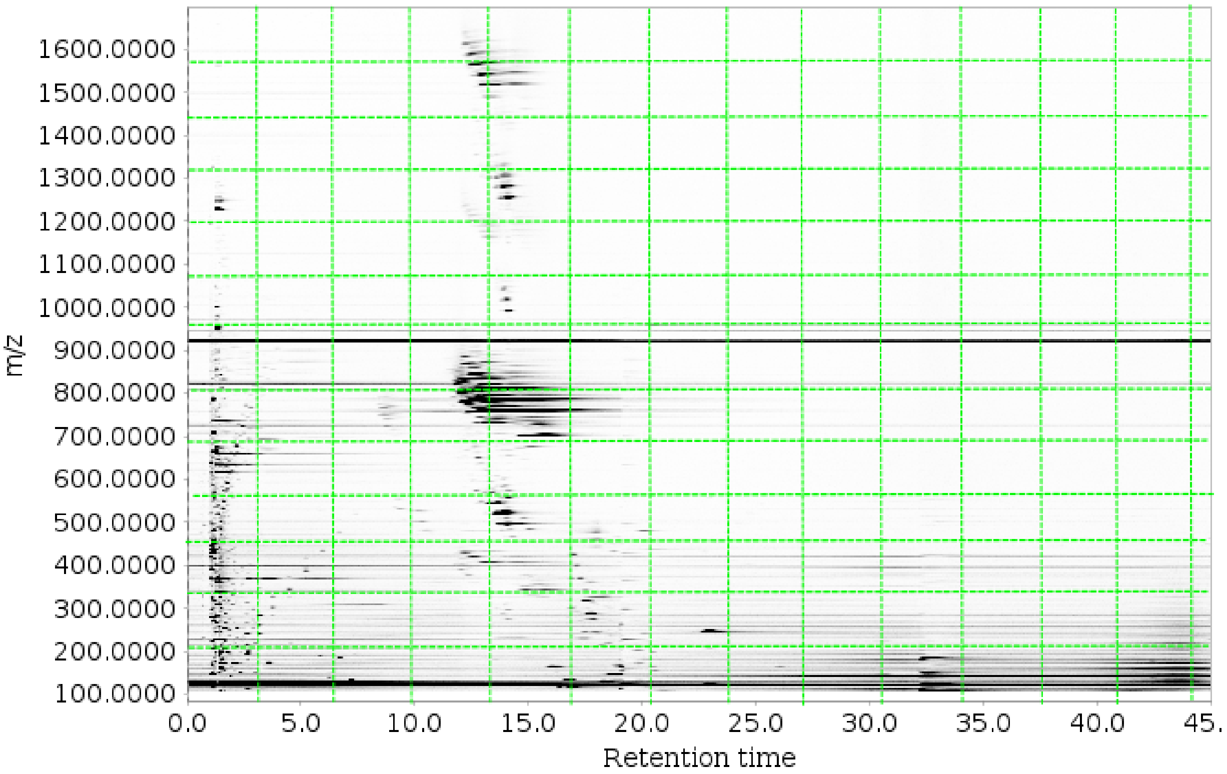

Figure 3.8: The "grid of summed intensities" approach for feature extraction consists of summing the intensities within each rectangle specified by the dotted green lines.

where the first penalty encourages the solution to be sparse while the second and third encourages it to be smooth over our covariate grid.

Deep learning

Deep learning is a class of machine learning algorithms that involve learning multiple levels of data representation and abstraction using model architectures. Deep learning is very flexible, with a huge number of options for model architectures. Here, we will first discuss the use of the most widely used types of deep network – deep convolutional networks – and will then describe some alternatives that could be better suited for particular LC-MS situations. This discussion builds upon the introduction to neural nets provided in Chapter 1, and borrows from the material presented in Michael Nielson's online book [43].

Convolutional neural networks use an architecture that exploits the spatial structure of data, making them a natural choice for classifying images or image-like data. Convolutional neural nets use three basic ideas: local receptive fields, shared weights, and pooling.

In Figure 1.1 (of Chapter 1), each neuron was connected to every neuron in the adjacent layers. To create a convolutional net, we will think of our data as a 2-dimensional grid of input neurons, and we will make connections in small, localized regions of the input image. For example, we can build a hidden neuron using just a 5 x 5 square of input neurons (known as a *local receptive field*), representing LC-MS intensities for adjacent retention time and m/z values. Each of these 25 input neurons will learn a weight, and the hidden neuron will learn an overall bias as well. We can then slide the local receptive field over by one retention time value (or by a larger *stride*) to connect to a second hidden neuron. This process can continue until we have swept through the entire input image. If the input image is 28 x 28 pixels, then there will be 24 x 24 units in the hidden layer (plus the bias term) since we can only move the local receptive field 23 units before reading the edge of the input image.

A key feature of convolutional neural nets is that each of the 24 x 24 neurons belonging to the hidden layer created by the method described above share the same set of 25 weights and the same bias. This means that all neurons in the hidden layer detect the same pattern of pixel values, just at different locations in the input image. As a result, convolutional networks are well-suited for classifying images in which the object of interest may be located in a variety of positions. The map from the input layer to the hidden layer is sometimes called a *feature map*, *kernel*, or *filter*. A complete convolutional layer consists of multiple feature maps, allowing for the detection of different kinds of localized features.

Directly following a convolutional layer, we can create a *pooling layer*, generated by condensing the information from the convolutional layer. There are many ways to construct a pooling layer. One simple method, known as *max-pooling*, involves creating pooling units that are each equal to the maximum activation value found in a defined region of the convolutional layer's feature maps. For example, we could take the maximum value found for each 2 x 2 square of each 24 x 24 feature map of the convolutional layer described above. If the convolutional layer contains three feature maps, then this pooling method would result in a pooling layer that is 3 x 12 x 12 neurons. Thus, pooling reduces the number of parameters needed in later layers.

The final layer is one which is fully connected, just like we described when introducing neural nets in Chapter 1 (i.e., there will be K x 3 x 12 x 12 connections between the pooling layer and the output layer). The network can be trained using stochastic gradient descent and back-propogation. Initializing all weights and biases to be 0 is an ad hoc procedure that generally works well enough in practice. In our example, taking the number of layers and units as fixed, there are (25 + 1) x 3 parameters to fit in going from the input layer to the convolutional layer, plus another (12 x 12 x 3 + 1) x K in going from the pooling layer to the output layer.

There are additional techniques that we can employ to potentially improve test set performance. One strategy, known as the *dropout* technique, involves removing individual neurons at random while training the network. This technique can reduce the tendency to over-fit by producing a model that is more robust to the loss of individual pieces of evidence, reducing reliance on what may be particular idiosyncrasies of the training data. An additional option to improve test performance is to use an ensemble of networks, whereby each network castes a “vote” to obtain a final predicted classification.

Two additional types of neural nets worth mentioning are recurrent neural nets and deep belief nets. Convolutional neural nets are *feed forward* nets in that the activations of neurons in later layers are completely determined by the activations of neurons in earlier layers. *Recurrent neural networks*, in contrast, base the behavior of hidden neurons not just on the activations in previous layers, but also at earlier times. In other words, the activation of a neuron may depend on earlier inputs. This time-varying behavior make them particularly well-suited for problems in natural language processing, and we can also imagine them being useful for accounting for the temporal ordering in which biological samples are collected and/or processed. The time-varying behavior unfortunately makes recurrent neural networks difficult to train, though incorporating an idea known as long short-term memory units can help. Deep belief nets are capable of both unsupervised and semi-supervised learning. That is, they can learn useful features even when the training images are unlabeled. In the LC-MS context, this feature could be useful if we had access to additional LC-MS data that lacked a diagnosis.

Deep learning has a huge amount of potential to maximize the diagnostic power of LC-MS data. It does, though, have two main drawbacks: (1) the results are not interpretable as SMBs, and (2) it generally requires a reasonably large amount of data.

3.4 Results using LC-MS data

Our results suggest that LC-MS data contains highly useful information for diagnostic purposes. We must provide the caveat, however, that the results in this chapter are based on very small samples so we cannot be sure of how well the observed results are indicative of what one would find in different samples. Note that while we display error bars in the figures throughout this chapter, these confidence intervals are determined according to the method described in Chapter 1.2, which is based in asymptotic theory; with such a small sample size, these error bars are unlikely to have the advertised 95% coverage. Thus, when we observe differences in the performance of methods, we do not know exactly what the probability is that in a new sample these same relative performances will be observed.

3.4.1 Distinguishing DENV from OFI

We are more successful at distinguishing dengue patients (DENV) from patients with other febrile illnesses (OFI) when using LC-MS data from serum samples than we are when using clinical information alone (Figure 3.9). Super learner achieved a cross-validated AUC of .90 when using only clinical data of the 88 patients for whom we had serum samples for normal phase LC-MS analysis. In contrast, LC-MS data yielded a cross-validated AUC of about .98. Prediction accuracy was similar when using LC-MS data alone, and when combining it with clinical data. Prediction accuracy was also similar when using the super learner algorithm with different LC-MS feature extraction methods, though the Mass Hunter feature extraction method out-performed the grid of summed intensities method when using other prediction methods such as Adaboost, Gradient Boost, and Random Forests (Figure 3.9). Note that all results presented in Figure 3.9 are based on the same set of 88 samples. Thus, when using clinical data alone, the Mass Hunter and intensity grid methods should yield identical results. However, due to the stochastic nature of the machine-learning methods, slight differences in results are observed.

Normal and reverse phase LC-MS results are very similar. Figure 3.10 suggests that reverse phase LC-MS might perform slightly better than normal phase LC-MS, given that

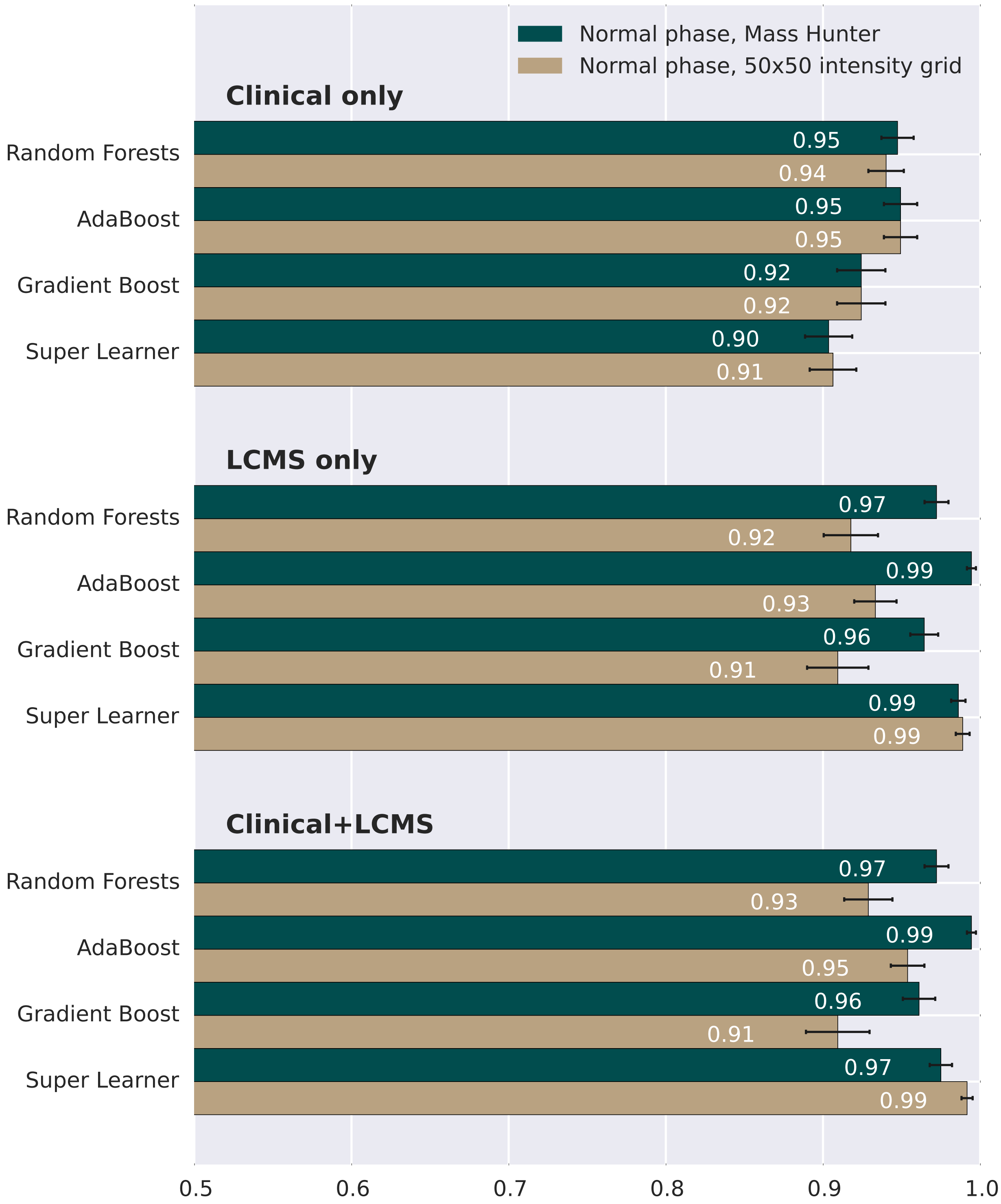

Figure 3.9: Cross-validated AUCs and corresponding 95% confidence intervals for various algorithms, predictor sets, and LC-MS data processing methods, OFI vs. DENV analysis. The same 88 serum samples were used for both analyses.

the data on which reverse phase LC-MS was conducted appears somewhat noisier than the data on which normal phase LC-MS was run (i.e., using only clinical information, prediction results were better for the patients from whom samples were extracted for the normal phase analysis), but differences appear insubstantial relative to the noise. Here we also observe super learner being out-performed (as measured by cvAUC) by some of the competing algorithms. Indeed, super learner cannot offer guarantees for such small sample sizes, as luck plays a larger role.

LC-MS analysis appears less useful when run using saliva and urine samples as compared to serum samples. In fact, the molecular features extracted using Mass Hunter from these non-invasive samples appear significantly less useful for predicting dengue than clinical information (Figure 3.11). Interestingly, LC-MS data combined with clinical data generates poorer prediction accuracy than clinical information alone. It is possible that the learners are over-fitting to irrelevant features found in the LC-MS data and are thus better off without access to such features. This problem may be mitigated by additional observations.

3.4.2 Identifying severe dengue cases

While we would ultimately like to predict which patients will experience severe dengue (DHF/DSS), restricting our LC-MS analysis to include only those who were not yet displaying DHF/DSS symptoms at the time of sample collection leaves us with too few observations to work with. Thus, we instead examine whether LC-MS is useful for distinguishing severe dengue patients from the rest, irregardless of symptoms at the time of sample collection. We find that the Mass Hunter feature extraction method does somewhat better than the binned intensity approach, but clinical data out-performs LC-MS data regardless of how it is processed. In fact, using LC-MS data in combination with clinical data does not improve prediction results beyond what is achieved using clinical data alone (Figure 3.12).

There is some evidence that reverse phase LC-MS is more helpful for diagnosing severe dengue than is normal phase LC-MS (Figure 3.13), with reverse phase results looking similar to the results generated using only clinical information.

3.4.3 Diagnosis mislabeling investigation

Altering the diagnoses of observations for which we have poor predictive power will clearly yield dishonest results if done without scientific justification. Our results presented throughout this study take the DENV labeling in the original data as fact. However, of the approximately 90 observations for which we have LC-MS data, one of them stands out as being

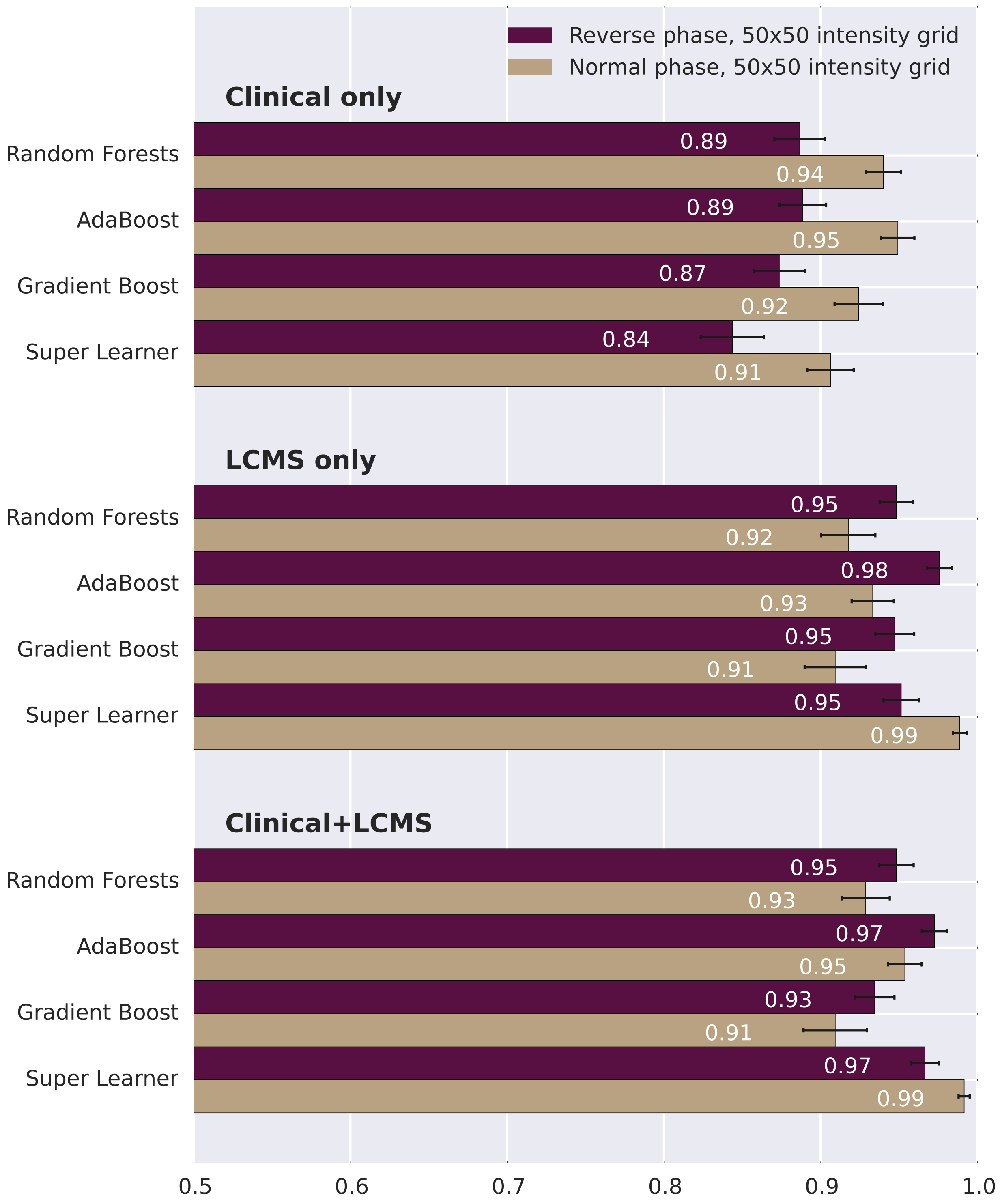

Figure 3.10: Cross-validated AUCs and corresponding 95% confidence intervals for various algorithms and predictor sets, OFI vs. DENV analysis. Serum samples used in the reverse phase LC-MS (n=91) were taken from different patients than those used in the normal phase LC-MS (n=88), resulting in substantially different performance results when using only clinical information.

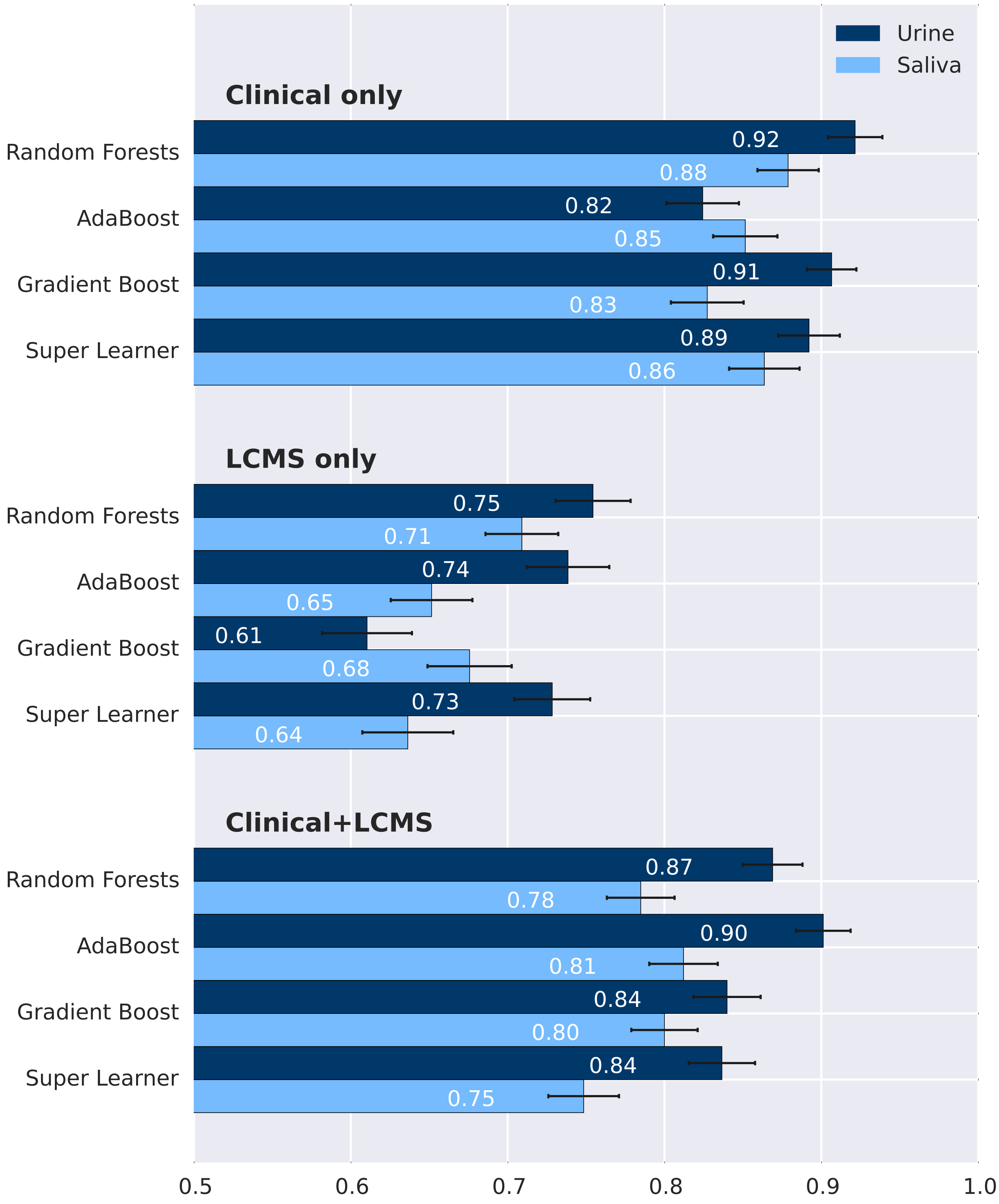

Figure 3.11: Cross-validated AUCs and corresponding 95% confidence intervals for various algorithms and predictor sets, OFI vs. DENV analysis. Samples used in the saliva analysis (n=85) were taken from different patients than those used in the urine analysis (n=80), resulting in different performance results when using only clinical information.

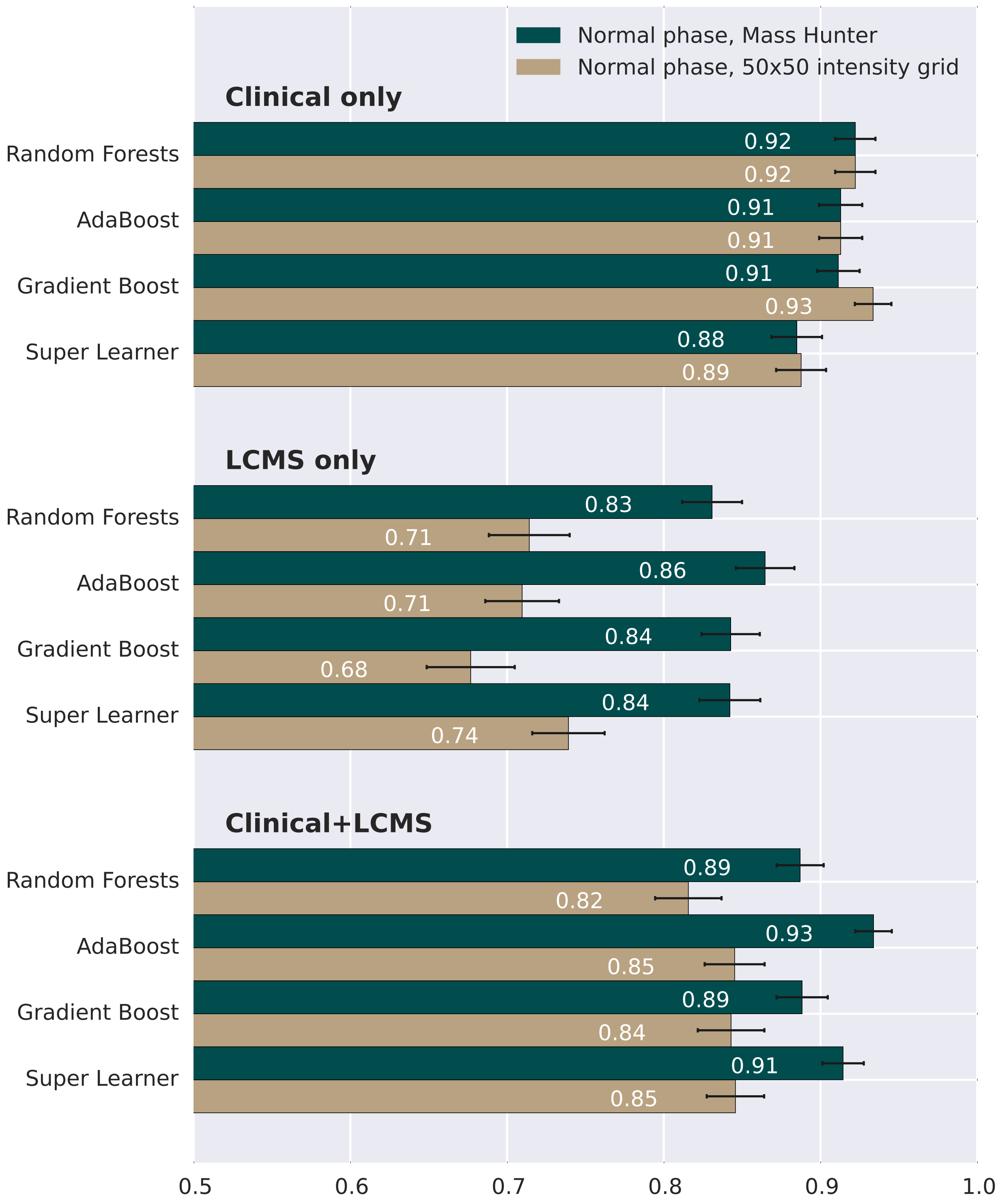

Figure 3.12: Cross-validated AUCs and corresponding 95% confidence intervals for various algorithms, predictor sets, and LC-MS data processing methods, OFI/DF vs. DHF/DSS analysis. Patients were included even if DHF/DSS symptoms were observed at the time of sample collection (n=88 serum samples).

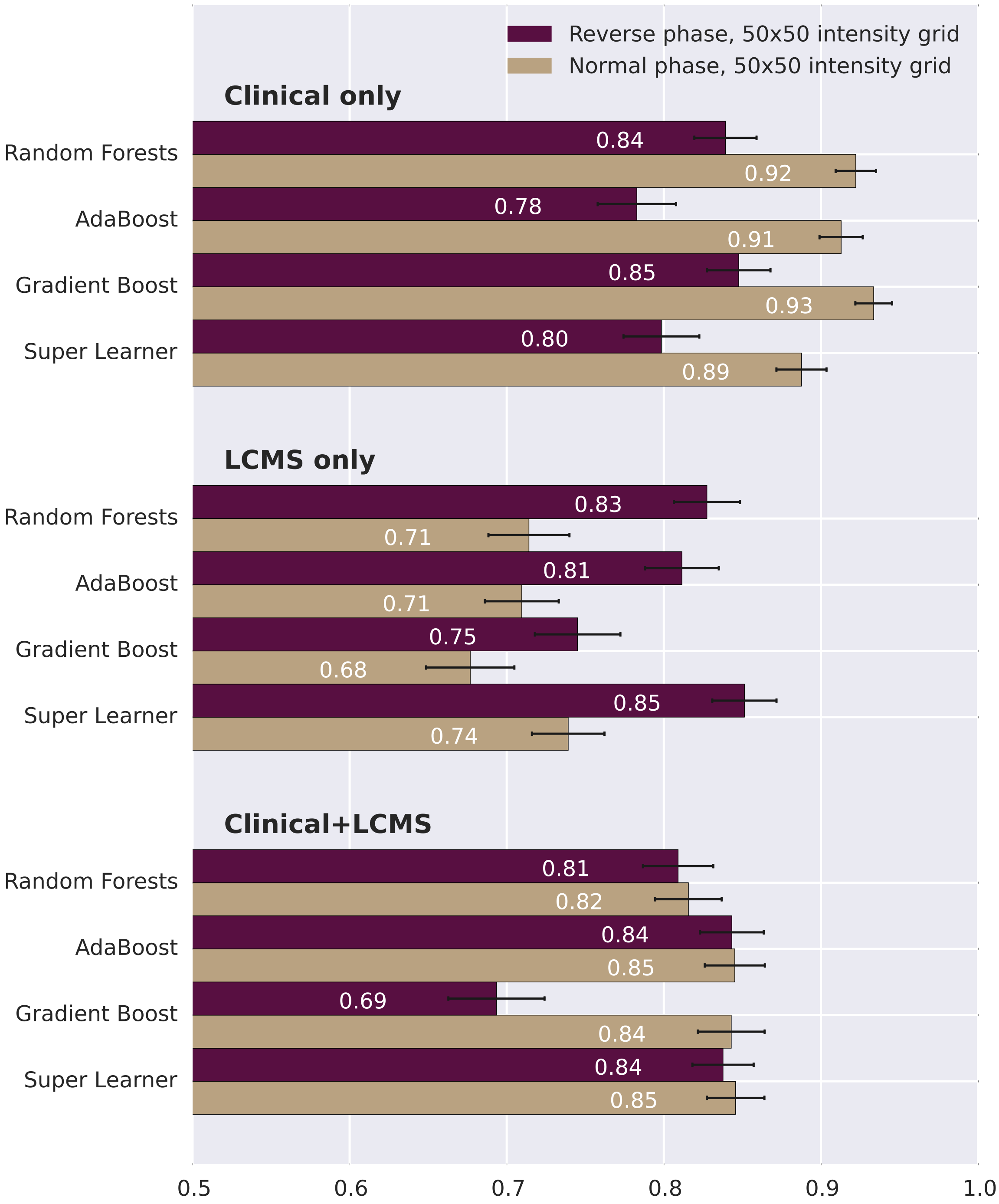

Figure 3.13: Cross-validated AUCs and corresponding 95% confidence intervals for various algorithms and predictor sets, OFI/DF vs. DHF/DSS analysis. Samples used in the reverse phase LC-MS (n=91) were taken from different patients than those used in the normal phase LC-MS (n=88), resulting in different performance results when using only clinical information. Patients were included even if DHF/DSS symptoms were observed at the time of sample collection.

consistently predicted to have a very low likelihood of being dengue-positive despite being labeled as DF in the data. Based on the time in which this sample was collected, as well as the fact that both clinical indicators and the metabolic features strongly predict this observation to be OFI, we believe this observation was incorrectly labeled in the data. To determine the impact of re-labeling this observation as OFI, we ran our three main analyses using serum normal phase LC-MS data after making just this one label change. While our previous results for predicting DENV were good, our new results are spectacular: using clinical and LC-MS data, we are now able to achieve near perfect prediction – an error rate of 1% corresponding to a 98% sensitivity and 100% specificity (Table 3.6).

Features	Modify?	NPV	PPV	Error	Sensitivity	Specificity	AUC 95% CI		AUC
Clinical only	No	0.79	0.95	0.11	0.88	0.90	0.88	0.93	0.90
LC-MS only	No	0.92	0.93	0.07	0.97	0.86	0.98	0.99	0.99
Clinical + LC-MS	No	0.93	0.93	0.07	0.97	0.86	0.96	0.99	0.98
Clinical only	Yes	0.83	0.91	0.11	0.91	0.83	0.90	0.94	0.92
LC-MS only	Yes	0.97	0.97	0.03	0.98	0.93	0.99	1.00	1.00
Clinical + LC-MS	Yes	0.97	1.00	0.01	0.98	1.00	0.99	1.00	1.00

Table 3.6: Cross-validated performance measures with and without modification of suspected false positive observation, OFI vs. DENV analysis. The positive predictive value (PPV), negative predictive value (NPV), error rate, sensitivity and specificity are based on the threshold value c (discussed in Chapter 1.4.1) which minimizes the error rate.

3.5 Discussion

Our analysis reveals that LC-MS is a powerful tool for distinguishing dengue disease states. Due to the small sample sizes, we were not able to conclude whether there is a significant difference between the predictive powers of different LC-MS laboratory methods (reverse phase versus normal phase), but our analysis is suggestive that serum samples contain more useful information than do urine and saliva samples for the purpose of prediction, and that LC-MS is more useful for diagnosing dengue (OFI versus DENV) than it is for distinguishing severe dengue patients from other patients (OFI/DF versus DHF/DSS).

It is worth noting that the dengue-positive serum samples used in this analysis were all of serotype 2. Without a more thorough investigation of which molecular features our analysis is deeming important, we might be uncovering predictors that are specific to serotype 2 and not more generally useful to other serotypes. Indeed the metabolomics study conducted by Birungi et al. [4] found each serotype to have a distinct metabolic signature.

It is also worth discussing the likely effect that the timing of our sample collection has on our results. We have already discussed the changes in clinical symptoms over the course of a dengue infection (Figure 1). It is no surprise, then, that the greatest metabolomic differential between OFI and DENV patients occurs within the first 72 hours [12]. If our samples had been collected earlier in the disease progression, we could therefore anticipate additional power for distinguishing OFI from DENV patients, though possibly a reduction in power for identifying severe dengue patients, as the symptoms of severe dengue grow more severe on days 4 and 5.

While this section revealed the broad usefulness of LC-MS data for diagnosing dengue, in the next chapter we will examine the number of metabolites that appear to be driving our predictive power.

Chapter 4

Selecting a Best Subset of LC-MS Features

We now wish to identify a small subset of LC-MS features which, when used in combination with clinical predictors, are most helpful for distinguishing patients in various disease states. Since we wish to identify the molecular composition of these features to help pave the way for a low cost diagnostic tool that will not require LC-MS, we use only the feature extraction methods that are biologically motivated (e.g., MZmine and Mass Hunter, discussed in Chapter 3). While our focus is still on predicting whether a given patient has dengue or severe dengue fever, we will also discuss the more basic problem of feature assessment in the context of multiple hypothesis testing.

4.1 Literature

Most studies in the dengue prediction literature use either a battery of univariate tests, regression trees (e.g., CART), or the coefficients of regression models to pick out both clinical variables of relevance (Tables 2.2 and 2.3) and metabolites of importance [38, 14]. Ju and Brasier [24] do use additional approaches, though only after first reducing their feature set to 25 variables based on univariate t-test results.[1] Furthermore, no paper to the author's knowledge apply dimension reduction strategies while controlling for the influence of other features as we do.

[1] This paper was not included in Chapter 2's literature review as it was based on only 51 observations (38 DF and 13 DHF patient samples) and so did not meet our sample size criteria. Additionally, the cross-validation procedure in this paper did not incorporate the initial dimension reduction procedure, thus producing optimistic point estimates of the test error.

4.2 Data description

All methods in this chapter are applied to the 88 serum samples described in Chapter 3. This data contains 29 other febrile illness (OFI) samples, 29 non-severe dengue (DF) samples, and 30 severe dengue (DHF/DSS) samples (Table 3.1). We consider all 744 molecular features extracted through the normal phase LC-MS procedure.

Clinical data comes from the hospital database described in Chapter 2, with missing clinical features imputed as described in Section 2.3.2. We thus have clinical information for 1,658 illness episodes, 673 of which are OFI, 745 of which are DF, and 240 of which are DHF/DSS (Table 2.5).

4.3 Methods

4.3.1 Feature assessment and the multiple-testing problem

Scientists often wish to identify the genes, proteins, or metabolites whose abundance differs between normal and diseased samples in order to learn about the disease and to develop targets for drug therapeutics. Thus, for completeness, we will discuss the traditional statistical topic of multiple hypothesis testing and will examine the number of molecular features that appear statistically significant for diagnosing dengue fever.

Using the most commonly applied test statistic – the two-sample t-statistic – for each LC-MS feature w_j, we get:

$$t_j = \frac{\bar{w}_{2j} - \bar{w}_{1j}}{se_j} \tag{4.1}$$

where $\bar{w}_{1j}$ is the mean value of w_j across the N_1 samples belonging to diagnostic group 1, and $\bar{w}_{2j}$ is the mean value of w_j across the N_2 samples belonging to diagnostic group 2. The quantity se_j is the standard error of differences of means. Assuming a common variance across diagnostic groups, we can use the pooled within-group standard error:

$$se_j = \hat{\sigma}_j \sqrt{\frac{1}{N_1} + \frac{1}{N_2}};\ \hat{\sigma}_j^2 = \frac{1}{N_1 + N_2 - 2} \left(\sum_i^{N_1} (w_{ij} - \bar{w}_{1j})^2 + \sum_i^{N_2} (w_{ij} - \bar{w}_{2j})^2 \right) \tag{4.2}$$

A traditional approach for accessing the results of all 744 molecular features involves computing a p-value for each feature using the t-distribution, which assumes that either the features are normally distributed, or the sample size is large enough to invoke the central limit theorem. Since our sample size is small, it is particularly dangerous to assume normally distributed parameter estimates and it is not necessary to do so: an appealing alternative approach is to use the permutation distribution, which does not rely on such an assumption. This approach consists of computing a test statistic t_j^k for each k permutation of the diagnostic labels. So if we were to apply this method to the problem of distinguishing OFI (n=29) patients from DENV patients (n=59), then we would calculate the test statistic for $K = \binom{88}{29}$ permutations of the diagnostic labels. The p-value for molecular feature j could then be calculated as:

$$p_j = \frac{1}{K} \sum_{k=1}^{K} I(|t_j^k| > |t_j|) \tag{4.3}$$

In other words, the p-value for feature j is the fraction of test statistics generated by the permutation of the diagnostic labels that are more extreme than t_j. Of course, $\binom{88}{29}$ is a large number; rather than computing test statistics for all possible permutations, we could base the p-value calculation on a random sample of, say, ten-thousand permutations.

Alternatively, we can pool the results for all molecular features in order to take advantage of the fact that they are measured on the same scale and are similar in other ways. Our p-value calculation then becomes:

$$p_j = \frac{1}{JK} \sum_{j'=1}^{J} \sum_{k=1}^{K} I(|t_{j'}^k| > |t_j|) \tag{4.4}$$

Relative to the approach described by Equation 4.3, this provides more granular p-values without imposing an additional computational burden.

Using the set of p-values from Equation 4.4, we will test the hypotheses:

- H_{0j}: disease state has no effect on molecular feature j

  versus

- H_{1j}: disease state has an effect on molecular feature j

for all $j = 1, 2, ..., J$ features. We reject H_{0j} at level α if $p_j < \alpha$. By construction, this means that the probability of falsely rejecting H_{0j} is α. Put another way, this test has a type-I error equal to α.

But with so many hypothesis tests, we are also concerned about the overall error rate. Two sensible measures of the overall error rate are the family-wise error rate (FWER) and the false discovery rate (FDR). FWER is the probability of at least one false rejection while

FDR is the expected proportion of falsely significant features. Using R to denote the number of hypotheses that are rejected and V to denote the number of true null hypotheses that are (falsely) rejected, the false discovery rate is then equal to $E(\frac{V}{R})$ while the family-wise error rate is $Pr(V \geq 1)$. We discuss each of these measures in more detail below.

The family-wise error rate depends on the correlation between tests. If the tests are independent each with type-I error rate α, then the probability that at least one of these tests is falsely rejected is $(1-(1-\alpha)^J)$. If the tests have positive dependence (i.e., the fact that test j was falsely rejected increases the probability that test k is falsely rejected), then the FWER will be less than $(1-(1-\alpha)^J)$.

The Bonferroni method is a very simple way of controlling the family-wise error rate, though it is too conservative, especially for large J. The Bonferroni method consists of lowering the p-value threshold by which we reject each null from α to α/J. This results in an FWER that is equal to at most α. There are many alternative ways to adjust the individual p-values to achieve an FWER of at most α. In particular, Dudoit et. al. (2002) [16] presents methods that avoid the independence assumption.

The notion of a false discovery rate and a testing procedure to limit it was first introduced by Benjamini and Hochberge (1995) [2]. This procedure, known as the Benjamini-Hochberg (BH) procedure (Algorithm 4.1), bounds the FDR by a user-defined level α. The p-values on which this method is based can be obtained from an asymptotic approximation to the test statistic, or from a permutation distribution. If the hypotheses are independent, Benjamini and Hochberge (1995) show that regardless of how many of the null hypotheses are true, and regardless of the distribution of p-values when the null hypothesis is false, the BH procedure results in an FDR that is at most α.

Algorithm 4.1 The Benjamini-Hochberg (BH) Method

```
1. Fix the false discovery rate $\alpha$ and let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(M)}$ denote the ordered p-values.
2. Define
       $L = max\left\{j : p_{(j)} < \alpha \cdot \frac{j}{J}\right\}.$    (4.5)
3. Reject all hypotheses $H_{oj}$ for which $p_j \leq p_{(L)}$, the BH rejection threshold.
```

A more intuitive approach for limiting the FDR could involve a direct plug-in estimate of the FDR, as described in Algorithm 4.2 [21]. By choosing the cut-off, C, such that $F\hat{D}R$ is as desired, this approach is actually equivalent to using the permutation p-values in the

BH procedure. The plug-in estimate is based on the approximation

$$E(V/R) \approx \frac{E(V)}{E(R)} \tag{4.6}$$

and in general $F\hat{D}R$ is a consistent estimate of FDR [57, 58]. Note that the numerator $E\hat{(V)}$ in Algorithm 4.2 actually estimates $(J/J_0)E(V)$, since the permutation distribution uses J rather than J_0 null hypotheses (where $J_0 < J$ is the number of true null hypotheses among J hypotheses tested). Hence a better estimate of FDR can be obtained if an estimate of J_0 is available. Specifically, we could decrease our estimate $F\hat{D}R$ by multiplying it by $\frac{\hat{J_0}}{J}$. (The most conservative estimate of FDR uses $J_0 = J$.)

Algorithm 4.2 The Plug-in Estimate of the False Discovery Rate

1. Create K permutations of the data, producing t-statistics t_j^k for features $j = 1, 2, ..., J$ and permutations $k = 1, 2, ..., K$.

2. For a range of values of the cut-point C, let

$$R_{obs} = \sum_{j=1}^{J} I(|t_j| > C),\ E\hat{(V)} = \frac{1}{K}\sum_{j=1}^{J}\sum_{k=1}^{K} I(|t_j^k| > C). \tag{4.7}$$

3. Estimate FDR by $F\hat{D}R = \frac{E\hat{(V)}}{R_{obs}}$.

The methods described above are based on the absolute value of the test statistic and therefore apply the same cut-points to both positive and negative values of the test statistic. But there is no theoretical justification for this symmetry. In fact, in some experiments most or all of the differentially expressed molecular features may be in the positive (or in the negative) direction. Thus, we will employ an approach that derives separate cut-points for the two directions of differential expression. This approach is known as the *significance analysis of microarrays* (SAM) procedure [63].

The SAM procedure is similar to the procedure described by Algorithm 4.2 insofar as we also calculate a test statistic t_j for each molecular feature and estimate attributes of the distribution of t_j under the null by generating permuted samples. Unlike previously, however, we will choose a threshold *band* by which to evaluate our hypotheses. This process is described in Algorithm 4.3.

The methods described above will give us a sense for the number of LC-MS features that distinguish disease states when evaluated individually. But certainly it is possible for LC-MS features to appear significant when examined in isolation, but insignificant when we control

Algorithm 4.3 The SAM procedure

1. Calculate and rank the J test statistics $t_1, t_2, ..., t_J$ to obtain the order test statistics $t_{(1)} \leq t_{(2)} \leq ... \leq t_{(M)}$.

2. Create K permutations of the data, producing t-statistics t_j^k for features $j = 1, 2, ..., J$ and permutations $k = 1, 2, ..., K$.

3. Calculate the expected order statistics from the permutations of the data:

   $$\tilde{t}_{(j)} = \frac{1}{K} \sum_{k=1}^{K} t_{(j)}^k$$

   where $t_{(1)}^k \leq t_{(2)}^k \leq ... \leq t_{(J)}^k$ are the ordered test statistics from permutation k.

4. Define a threshold Δ and find the smallest positive $t_{(j)}$ such that $|t_{(j)} - \tilde{t}_{(j)}| \geq \Delta$, and the largest negative $t_{(j)}$ such that $|t_{(j)} - \tilde{t}_{(j)}| \geq \Delta$. Call these values t_1 and t_2, respectively.

5. Each molecular feature that has a t_j value greater than t_1 or less than t_2 is considered significant (for given Δ).

6. Estimate $F\hat{D}R$ by estimating the number of (falsely) significant genes under the null and dividing by the number of genes called significant.

7. Repeat steps 4 - 6 until desired $F\hat{D}R$ is achieved.

for other factors. In the next section, we describe methods that will inform the extent to which LC-MS features provide unique information, both in combination with clinical data and in combination with one another.

4.3.2 Subset selection

In previous chapters, we discussed some classification methods that build a model involving only a subset of available features (e.g., elastic net logistic regression, nearest shrunken centroids, fused lasso). We used these classification methods in order to reduce over-fitting and to thus strike a proper variance-bias balance with the goal of reducing expected test error.

In contrast, we now wish to select features because we are up against practical constraints: we cannot expect a low-cost point-of-care diagnostic test for dengue fever to involve the collection of a full profile of metabolite information such as LC-MS provides. Therefore, we wish to find a small subset of LC-MS features that, when used in combination with the available clinical data, provide strong predictive power. If we had no computational constraints, we could do this by simply trying all possible combinations of J' LC-MS features and taking the combination that, when used with the clinical indicators, minimizes our test error. But if we want the best combination of, say, 5 LC-MS features, then with a total of 744 predictors to choose from, we have $\binom{744}{5} \approx 1.87 \times 10^{15}$ combinations to try. Thus, we will instead implement less computationally demanding methods. It should be noted that while these methods may work well in practice, none are guaranteed to find the subset of features that minimizes the test error.

We will employ a couple of different methods, specified by Algorithm 4.4, to select a subset of LC-MS features while controlling for clinical features. For the first of these methods ("Method 1"), we use data from our 88 patient-samples for which we have normal phase LC-MS serum data. While controlling for patient clinical information, we choose a subset of LC-MS features using one of the procedures described below. We then run prediction models using clinical data plus the selected LC-MS features. To achieve an honest estimate of prediction error, the entire described procedure will need to be embedded in a cross-validation step, as indicated by Algorithm 4.4.

One weakness of the method described above is that it does not make use of the 1,570 observations for which we have clinical information but lack LC-MS information. For the second method ("Method 2"), we begin by fitting the super learner model using these 1,570 observations that have clinical information but not LC-MS information. We then use this fitted model (which only uses clinical predictors) to obtain predicted probabilities for each of the 88 patient-samples that has associated LC-MS data. These predicted probabilities will then be treated as a variable in a new prediction problem that uses LC-MS information to obtain a subset of J' LC-MS variables. While this method will allow us to identify the molecular features (if any) that improve our prediction beyond what can be achieved with clinical information alone, we note that, unlike Method 1, it does not allow us to incorporate interactions between individual clinical indicators and LC-MS molecular features when judging the importance of LC-MS features. A hybrid of these two methods ("Method 3") consists of using the predicted probabilities from all clinical data in combination with the clinical data for the LC-MS sample subset (n=88) when choosing a subset of LC-MS features.

Algorithm 4.4 Framework for choosing "best" subset of LC-MS features.

1. Fit super learner using clinical data for all observations that lack LC-MS information.
2. Use fit from Step 1 to obtain predicted probabilities for each observation that *has* LC-MS information.
3. Split data from Step 2 into V folds.
4. For $v = 1$ to V:
    (a) Choose a subset of J' LC-MS variables using all observations from Step 2 except for those in fold v. There are many options for doing this, and for each option we could either use the LC-MS variables in conjunction with the data's clinical indicators ("Method 1"), the LC-MS variables with just the predicted probabilities obtained in Step 2 ("Method 2"), or both ("Method 3").
    (b) Fit super learner using all observations from Step 2 except those in fold v. For predictors, use the LC-MS variables selected in step 4a along with the clinical indicators used in step 4a.
    (c) Obtain predicted probabilities for fold v using fit from step 4b.
5. Access performance using predicted probabilities obtained in step 4c.
6. Compare to performance achieved by running step 4 but selecting zero LC-MS variables.

The methods described above all involve a variable selection procedure (Step 4a of Algorithm 4.4). Here we describe some options for this procedure.

Forward-stepwise variable selection

Forward stepwise variable selection, also known as the "greedy" approach, normally begins with an empty model (e.g., the predicted outcome is an average of the outcomes in the training set without regard to covariate values). In our case, though, we would begin with a model that includes clinical predictors while excluding LC-MS features. One by one, each LC-MS feature is considered for inclusion in the model, and the reduction in risk accompanying that variable's inclusion is calculated. The variable that brings the greatest risk reduction is added to the model and the process is repeated until some stopping criteria is satisfied. While there is no guarantee that this process will result in a subset of predictors that minimizes our loss function, it does have the appealing property of working with any number of chosen prediction routines, including super learner, such that we do not have to make assumptions on the functional form of the relationship between our predictors and outcome.

Multivariable Adaptive Regression Splines (MARS)

MARS is an adaptive procedure for regression that uses expansions in piecewise linear basis functions. The building strategy proceeds much like it does for forward stepwise logistic regression. However, rather than just considering each predictor in its original form, we consider the collection of basis functions with knots at each observed covariate value. When evaluated for model inclusion, we further consider the product of each candidate basis function with each of the terms of the current model (i.e., all interactions are considered). At the end of the model-building process, we generally have a large model that over-fits the data. Thus, we employ a backward deletion process that, at each stage, consists of removing the term which results in the smallest increase in risk. The final model size can theoretically be determined using cross-validation, but the MARS procedure instead uses the generalized cross-validation criterion, which strikes a balance between training risk and model complexity without cross-validation's computational burden.

L_1-penalized logistic regression

We can run an L_1-penalized logistic regression (as described in Chapter 1.5) with a complexity parameter chosen such that our desired number of LC-MS variables are selected. One weakness of this method is that it does not consider interaction terms unless the interaction terms are explicitly entered in the feature list. While theoretically we could add all interactions to our design matrix, a more targeted approach could consist of including just the interactions identified by a separate procedure as being important (e.g., multiple additive regression trees is one option for building this list of interactions).

Heuristic clustering-based methods

Another approach to selecting the best subset of LC-MS features could involve clustering the LC-MS indicators into C groups based on some correlation measure (e.g., k-means using Euclidean distances). We could then take the feature within each of the C clusters that best predicts our dengue outcome. Finally, by making C small enough, we can employ an exhaustive search over all combinations of features to find the best subsets (e.g., try all $\binom{C}{2}$ combinations to find best subset of 2 features, all $\binom{C}{3}$ combinations to find best subset of 3 features etc.).

Tree-based relative importance measures

As described in Chapter 1.5, random forests builds trees using bootstrap samples. When the *bth* tree is grown, we can pass the observations left out of the *bth* bootstrap sample (the so-called out-of-bag samples) through the tree to obtain a prediction accuracy estimate, much as we do in the cross-validation procedure. We can then permute the values of the *jth* variable in the out-of-bag samples and again pass these observations through the *bth* tree to obtain another prediction accuracy measurement. The difference between these accuracy measurements is averaged across all trees to give a relative variable importance measurement for each of the J predictors. Note that this importance measure is different from that which would be obtained if we were to compare the model's prediction accuracy *with* the *jth* predictor versus *without* the *jth* predictor. (This is because other variables could be used as surrogates if the model were refitted without predictor j; while useful for some purposes, the danger with the refitting strategy is that predictors with high predictive power but also high correlation with other variables would rank low in the variable importance measure.)

Alternatively, we can build a variable importance measure using the improvement in the split-criterion (e.g., gini coefficient) attributable to the splitting variable at each split in each tree. These improvements can be accumulated over all trees in the forest separately for each variable. This importance measure easily applies to other tree-based methods, not just to random forests. As with the out-of-bag permutation measure, the split-criterion-based measure gives *relative* variable importance; thus, it is common to assign the largest a value of 100 and then scale the others accordingly.

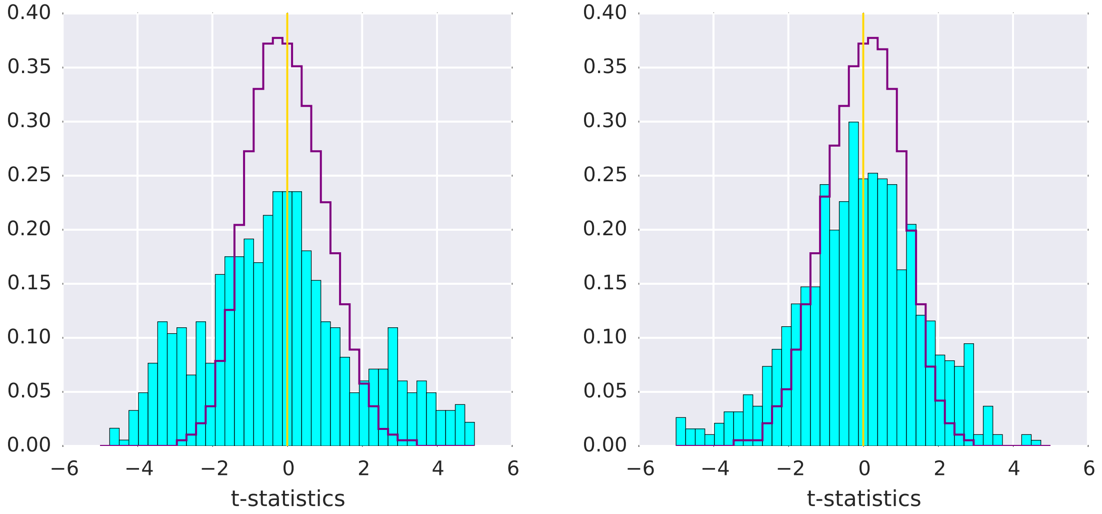

Figure 4.1: Histograms, in blue, of the t-statistic comparing OFI versus DENV samples (left graphic) and OFI/DF versus DHF/DSS samples (right graphic). Outlined in purple are the histograms of the t-statistics from 10,000 permutations of the respective diagnostic labels.

4.4 Results

4.4.1 LC-MS feature significance

Our analysis supports the notion that a large number of the 744 LC-MS features are statistically significant, both for distinguishing OFI patients from DENV patients, and for distinguishing severe dengue patients (DHF/DSS) from others (OFI/DF).

First, from Figure 4.1 it is clear that the test statistics (using Equation 4.1) are, on average, much larger in absolute value when calculated using our data than when calculated under the null distribution (generated using the permutation method).

Correspondingly, we find significantly lower p-values in our data than predicted by the null distributions (Figure 4.2). Mirroring what we saw in Figure 4.1, the LC-MS data appears to contain more features that are useful for predicting DENV than features that are useful for predicting severe dengue. In fact, if we choose a false discovery rate of .20, then the Benjamini-Hochberg method classifies about 320 features as "significant" for OFI versus DENV prediction, but only about half as many for predicting severe dengue.

The significance analysis of microarrays (SAM) procedure with an FDR of .20 also classifies just over 300 features as significant for distinguishing OFI from DENV and about 150 features as significant for distinguishing OFI/DF patients from DHF/DSS patients (Figure 4.3). Given that we are only assessing the importance of 744 molecular features, both

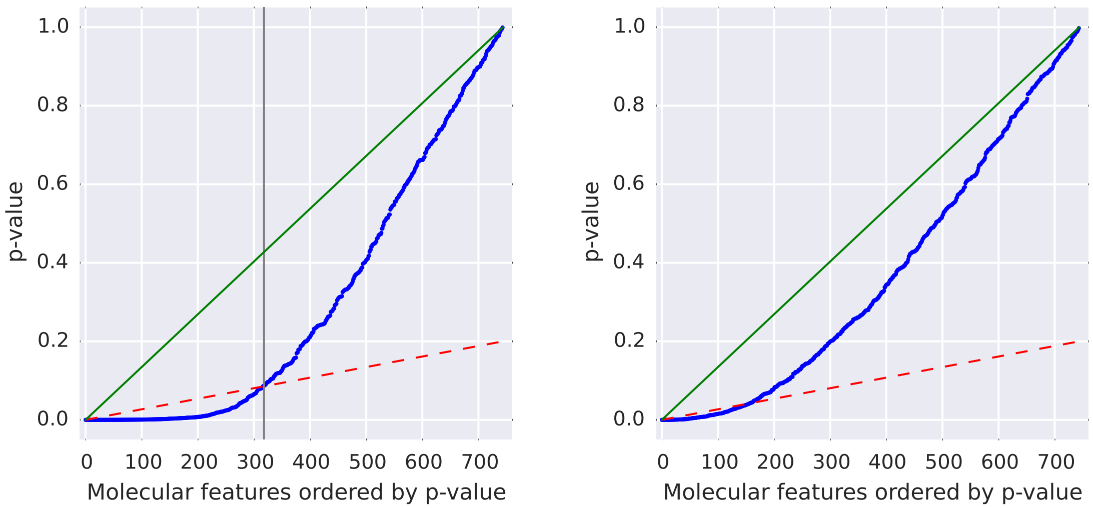

Figure 4.2: Plots of ordered p-values (blue dots) for OFI versus DENV analysis (left graphic) and OFI/DF versus DHF/DSS analysis (right graphic). Molecular features with p-values below the red dotted lines are considered significant by the Benjamini-Hochberg method with an overall false discovery rate of .20. If no features were actually associated with disease state, then p-values would be uniformly distributed between 0 and 1, as indicated by green lines.

numbers are impressively large.

4.4.2 Prediction performance of subset methods

It is clear from the section above that the LC-MS data contains useful features for both diagnosing dengue disease and for determining disease severity. In this section, we examine the benefits of using LC-MS data in combination with clinical data, and examine whether gains are achieved when a greater number of LC-MS features are used. To provide a basis of comparison, we contrast the methods presented in Algorithm 4.4 with the method of taking the LC-MS features that perform best according to a battery of univariate tests.

As expected, we find that choosing a "best subset" of LC-MS features using a method that selects each feature in isolation from other data features (the "ttest" method) performs worse than the method of choosing features in a manner that takes into account correlations with other predictors (the "topRF"[2] and "greedyRF"[3] methods). Our OFI versus DENV

[2] Specifically, we took the top J' features using random forests' gini importance measure.

[3] As the name implies, this procedure consists of choosing features in a forward step-wise fashion with random forests whereby feature selection is based on cross-validated AUC.

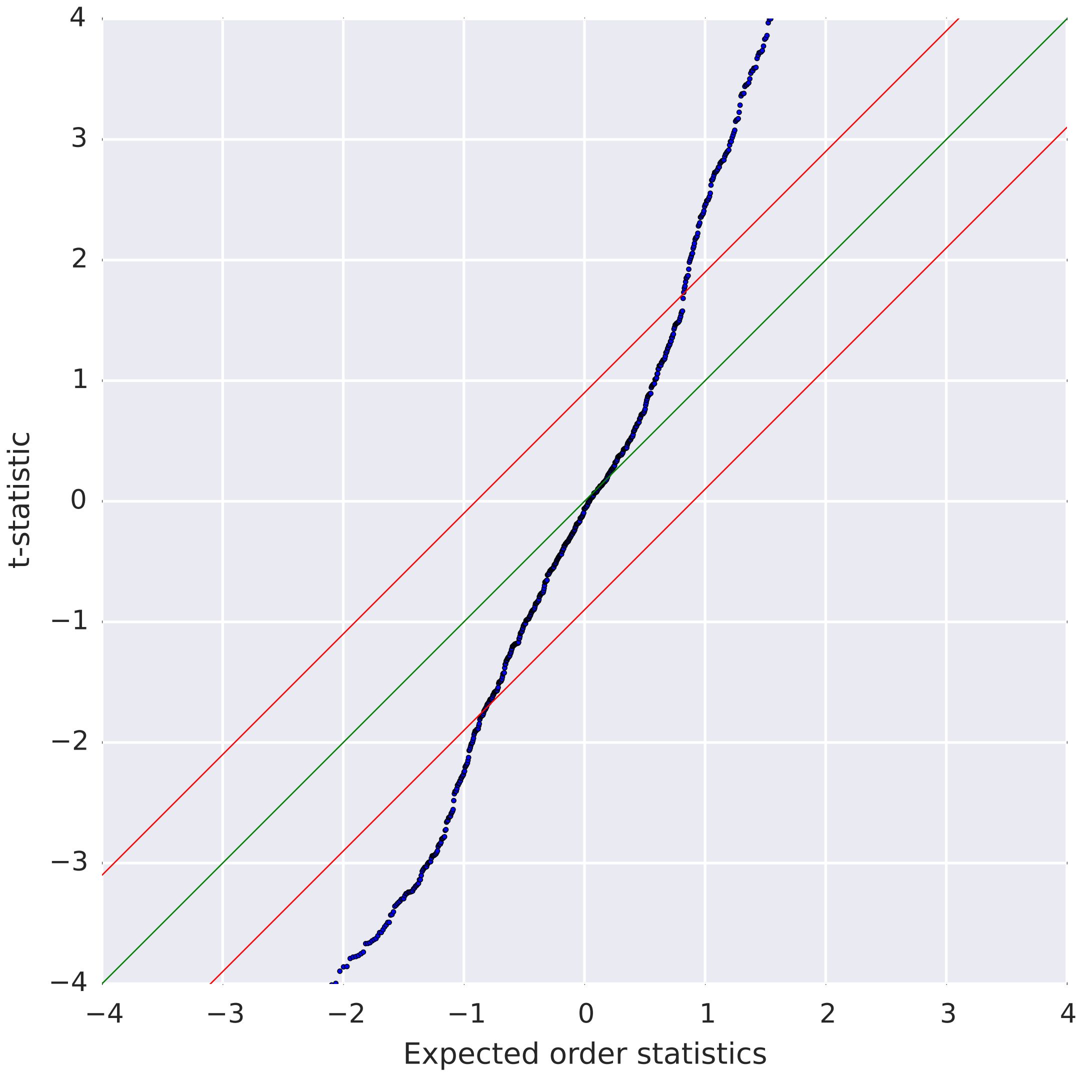

Figure 4.3: A plot of the ordered test statistics (vertical axis) versus the expected order statistics (horizontal axis) from permutations of the data, OFI versus DENV analysis. The two dotted lines are Δ units away from the 45 degree line, illustrating the significance criteria of the SAM procedure for an FDR of .20.

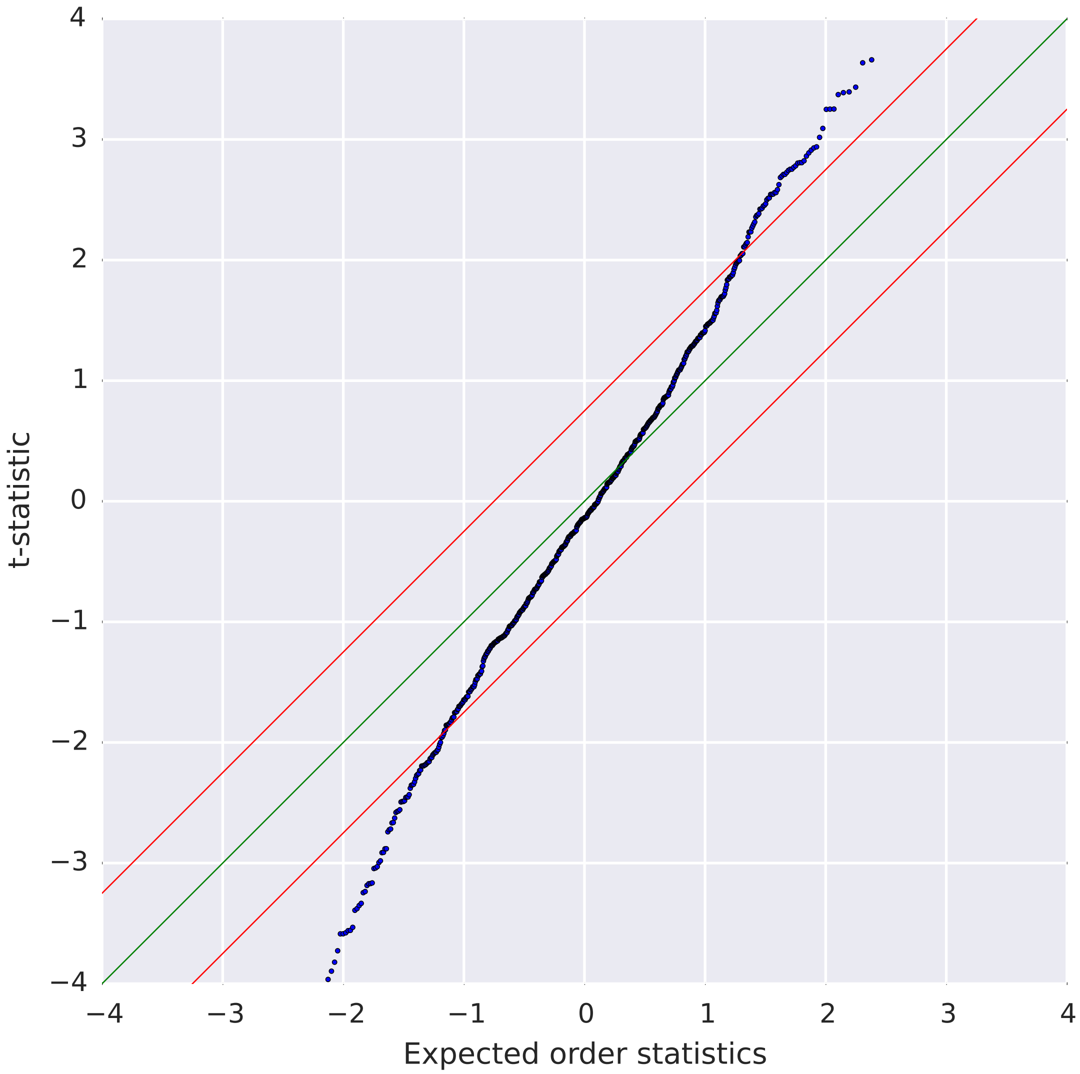

Figure 4.4: A plot of the ordered test statistics (vertical axis) versus the expected order statistics (horizontal axis) from permutations of the data, OFI/DF versus DHF/DSS analysis. The two dotted lines are Δ units away from the 45 degree line, illustrating the significance criteria of the SAM procedure for an FDR of .20.

analysis (Figure 4.5) also supports the notion that little is to be gained by adding LC-MS features beyond the most predictive two or three; while our data is very noisy, prediction accuracy does not appear to change to any measurable degree when we increase the number of LC-MS features beyond three. Methods 1 and 3 tend to do somewhat better than Method 2, suggesting that interactions between LC-MS features and clinical features may be useful for predicting dengue diagnosis, though we do not have enough data to obtain statistical significance for this difference.

Results from the OFI/DS versus DHF/DSS analysis (Figure 4.6) tell a different story. Here, the “topRF” and “greedyRF” procedures do not consistently outperform the “ttest” procedure for variable selection, and Method 1 is often out-performed by the others. These results are not actually surprising given that when it comes to severe dengue prediction, our LC-MS features do not improve prediction accuracy beyond what is achieved using clinical information alone (as seen previously in Figure 3.12). Thus, if none of the molecular features seem to add detectably useful information, then the observed differences among the AUC point estimates for our three procedures (“ttest”, ”topRF”, and “greedyRF”) is effectively noise.

4.4.3 Compound identification

Peak identification methods using custom and online databases were discussed briefly in Chapter 3.3. As mentioned, the mass measurement and retention time information coming from an LC-MS experiment is typically not sufficient for conclusively determining the formula of an unknown compound [52], though these methods can give us candidate small molecule biomarkers. Tandem MS (also known as MS/MS) can then be used to help resolve some of the ambiguity. This procedure consists of two or more stages of mass analysis applied independently, giving a more precise estimate of mass. Typically, for a confirmatory analysis, LC-MS/MS will be run using purchased “standards” containing a candidate small molecule biomarker, and the resulting LC-MS/MS profile will be compared to the LC-MS/MS profile achieved using the data sample(s) in question.

4.5 Discussion

Additional work is needed to both confirm the predictive power of the molecular features that we have found to be useful in our data sample, and to identify which small molecule biomarkers these molecular features represent. Beyond that, we will also need to think about

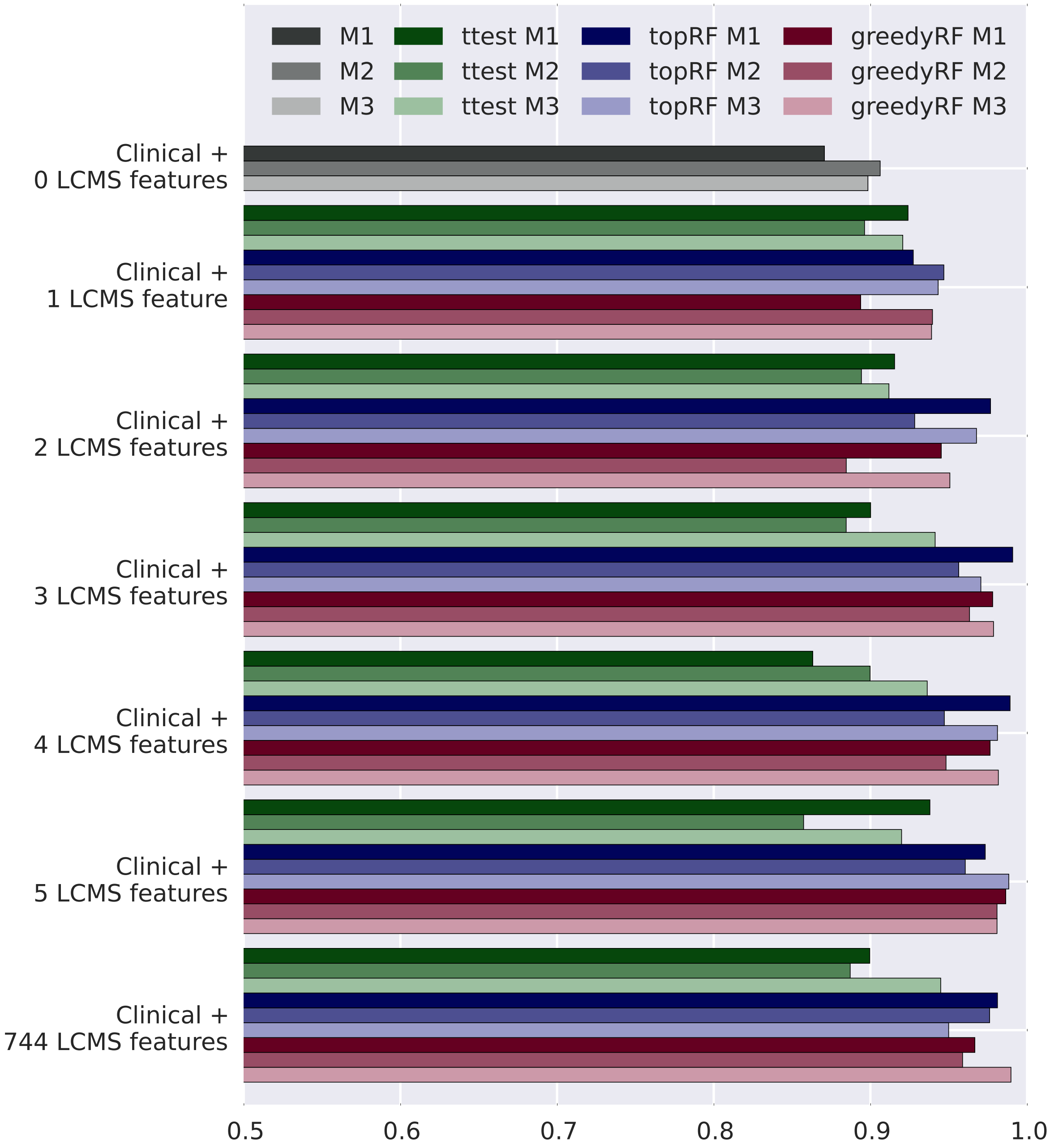

Figure 4.5: Cross-validated AUCs for super learner run with clinical data and various numbers of LC-MS features using various feature selection methods, as described by Algorithm 4.4 for the OFI vs. DENV analysis. "M1" stands for "Method 1", "M2" for "Method 2" and "M3" for "Method 3".

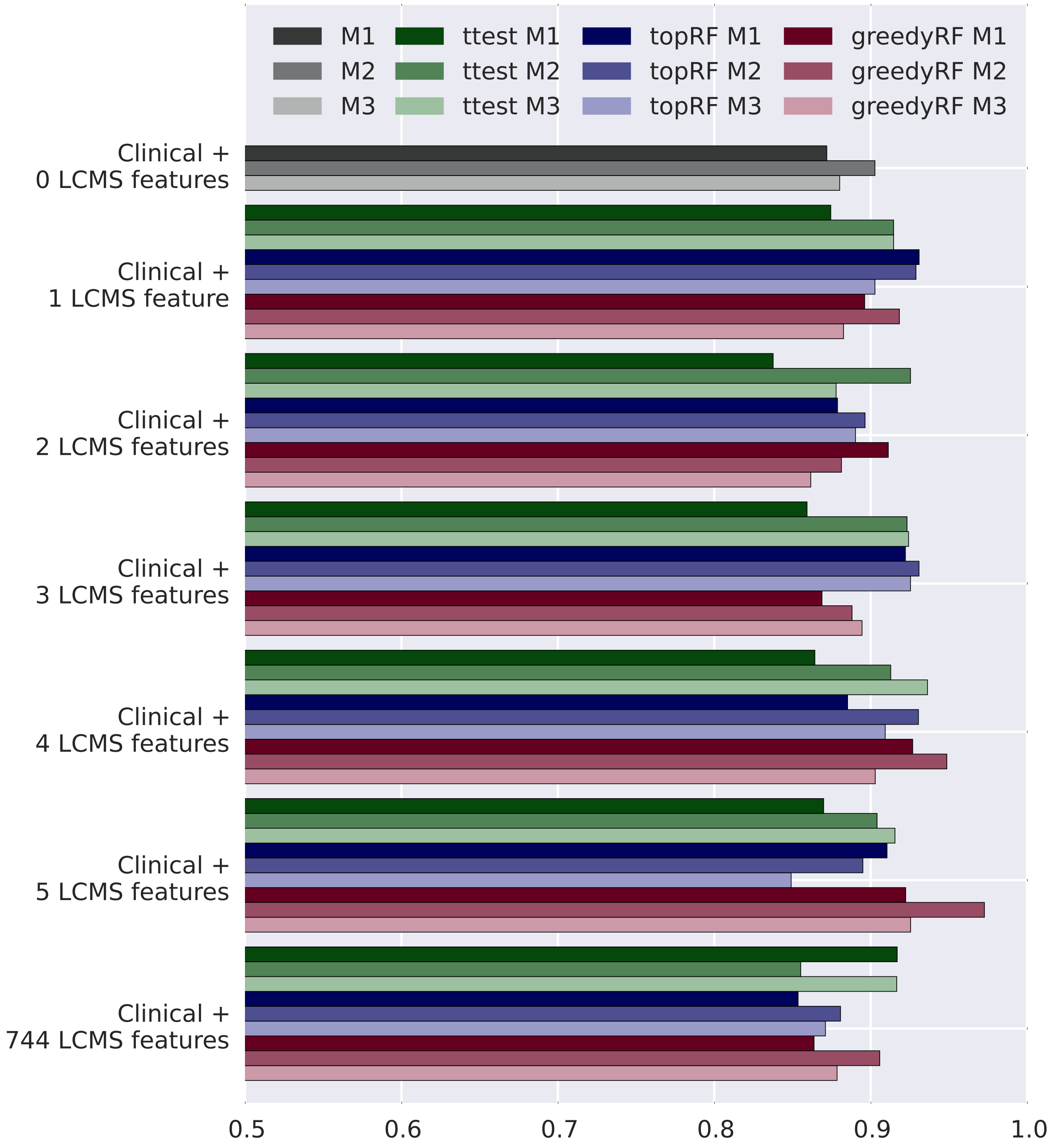

Figure 4.6: Cross-validated AUCs for super learner run with clinical data and various numbers of LC-MS features using various feature selection methods, as described by Algorithm 4.4 for the OFI/DF vs. DHF/DSS analysis. “M1” stands for “Method 1”, “M2” for “Method 2” and “M3” for “Method 3”.

the precision of a hypothetical point-of-care diagnostic test to detect a molecule of interest; if it is only able to give a binary response as to whether a particular molecule is present in one's biological sample, without information on quantity, then we might want to run our LC-MS analysis with dichotomized intensity values to reflect the concentration level that a point-of-care test is likely to be able to detect.

Finally, it should also be noted that the signs and symptoms of severe dengue, such as hemorrhaging, are not exclusively dengue symptoms. In particular, the Chikungunya virus can cause similar symptoms and we in fact do observe patients who display the signs and symptoms of severe dengue while not actually testing as dengue positive. Indeed if we instead predict the signs and symptoms of severe dengue, rather than a DHF/DSS diagnosis, our prediction accuracy may improve. The implication is that we are not necessarily detecting the metabolites specific to the dengue virus and if this is one's goal, then other methods (such as viral isolation) are safer. On the other hand, predicting whether a patient is likely to hemorrhage, regardless of diagnosis, may be of greater relevance in terms of patient care and could serve as an alternative prognostic goal.

Bibliography

[1] Till Bald, Johannes Barth, Anna Niehues, Michael Specht, Michael Hippler, and Christian Fufezan. PymzML-Python module for high-throughput bioinformatics on mass spectrometry data. *Bioinformatics*, 28(7):1052–1053, 2012.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing, 1995.

[3] Samir Bhatt, Peter W Gething, Oliver J Brady, Jane P Messina, Andrew W Farlow, Catherine L Moyes, John M Drake, John S Brownstein, Anne G Hoen, Osman Sankoh, and Others. The global distribution and burden of dengue. *Nature*, 2013.

[4] Grace Birungi, Sheryl Meijie Chen, Boon Pheng Loy, Mah Lee Ng, Sam Fong, and Yau Li. Metabolomics Approach for Investigation of Effects of Dengue Virus Infection Using the EA . hy926 Cell Line research articles. *Journal of proteome research*, 9:6523–6534, 2010.

[5] Hope H Biswas, Oscar Ortega, Aubree Gordon, Katherine Standish, Angel Balmaseda, Guillermina Kuan, and Eva Harris. Early clinical features of dengue virus infection in nicaraguan children: a longitudinal analysis. *PLoS neglected tropical diseases*, 6(3):e1562, jan 2012.

[6] Hans F M Boelens, Paul H C Eilers, and Thomas Hankemeier. Sign constraints improve the detection of differences between complex spectral data sets: LC-IR as an example. *Analytical Chemistry*, 77(24):7998–8007, 2005.

[7] Leo Breiman. Random Forests. *Journal of Machine Learning*, 45(1):5–32, 2001.

[8] Maria Rosario Capeding, Ngoc Huu Tran, Sri Rezeki S Hadinegoro, Hussain Imam Hj Muhammad Ismail, Tawee Chotpitayasunondh, Mary Noreen Chua, Chan Quang Luong, Kusnandi Rusmil, Dewa Nyoman Wirawan, Revathy Nallusamy, Punnee Pitisuttithum, Usa Thisyakorn, In Kyu Yoon, Diane van der Vliet, Edith Langevin, Thelma

Laot, Yanee Hutagalung, Carina Frago, Mark Boaz, T. Anh Wartel, Nadia G. Tornieporth, Melanie Saville, and Alain Bouckenooghe. Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial. *The Lancet*, 384(9951):1358–1365, 2014.

[9] David Chadwick, Barbara Arch, Annelies Wilder-Smith, and Nicholas Paton. Distinguishing dengue fever from other infections on the basis of simple clinical and laboratory features: Application of logistic regression analysis. *Journal of Clinical Virology*, 35:147–153, 2006.

[10] Matthew C Chambers, Brendan Maclean, Robert Burke, Dario Amodei, L Daniel, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Katherine Hoff, Darren Kessner, Natalie Tasman, Nicholas Shulman, Barbara Frewen, Tahmina a Baker, Mi-youn Brusniak, Christopher Paulse, Brent Lefebvre, Frank Kuhlmann, Joe Roark, Paape Rainer, Trey Chadick, Krisztina Holly, Josh Eckels, Eric W Deutsch, and Robert L Moritz. NIH Public Access. 30(10):918–920, 2013.

[11] Nancy R. Cook. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115:928–935, 2007.

[12] Liang Cui, Yie Hou Lee, Yadunanda Kumar, Fengguo Xu, Kun Lu, Eng Eong Ooi, Steven R Tannenbaum, and Choon Nam Ong. Serum metabolome and lipidome changes in adult patients with primary dengue infection. *PLoS neglected tropical diseases*, 7(8):e2373, jan 2013.

[13] G. Cybenko. Degree of approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 9(3):303–314, 1989.

[14] Judith R Denery, Ashlee a K Nunes, Mark S Hixon, Tobin J Dickerson, and Kim D Janda. Metabolomics-based discovery of diagnostic biomarkers for onchocerciasis. *PLoS neglected tropical diseases*, 4(10), jan 2010.

[15] Pan Du, Warren a Kibbe, and Simon M Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics (Oxford, England)*, 22(17):2059–65, sep 2006.

[16] S Dudoit, Y H Yang, M J Callow, and T P Speed. Statistical methods for identifying differentially expressed genes in replicated c{DNA} microarray experiments. *Stat. Sinica*, 12(1):111–139, 2002.

[17] Sandrine Dudoit and Mark J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, jul 2005.

[18] David I Ellis, Warwick B Dunn, Julian L Griffin, J William Allwood, and Royston Goodacre. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*, 8(9):1243–1266, 2007.

[19] Maria G Guzman, Scott B Halstead, Harvey Artsob, Philippe Buchy, Jeremy Farrar, Michael B Nathan, Jose Luis Pelegrino, Cameron Simmons, and Sutee Yoksan. Europe PMC Funders Group Dengue : a continuing global threat Europe PMC Funders Author Manuscripts. 8(12 0):1–26, 2015.

[20] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class AdaBoost. *Statistics and Its Interface*, 2(3):349–360, 2009.

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009.

[22] Nicholas S Heaton, Rushika Perera, Kristi L Berger, Sudip Khadka, Douglas J Lacount, Richard J Kuhn, and Glenn Randall. Dengue virus nonstructural protein 3 redistributes fatty acid synthase to sites of viral replication and increases cellular fatty acid synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(40):17345–50, 2010.

[23] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[24] Hyunsu Ju and Allan R Brasier. Variable selection methods for developing a biomarker panel for prediction of dengue hemorrhagic fever. *BMC research notes*, 6:365, 2013.

[25] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(D1):D199–D205, 2014.

[26] Mikko Katajamaa and Matej Orešič. Processing methods for differential analysis of LC/MS profile data. *BMC bioinformatics*, 6:179, 2005.

[27] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian

Zhang, and Stephen H. Bryant. PubChem Substance and Compound databases. *Nucleic Acids Research*, page gkv951, 2015.

[28] Tobias Kind and Oliver Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, 8:105, jan 2007.

[29] Guillermina Kuan, Aubree Gordon, William Avilés, Oscar Ortega, Samantha N Hammond, Douglas Elizondo, Andrea Nuñez, Josefina Coloma, Angel Balmaseda, and Eva Harris. The Nicaraguan pediatric dengue cohort study: study design, methods, use of information technology, and extension to other infectious diseases. *American journal of epidemiology*, 170(1):120–9, jul 2009.

[30] Carsten Kuhl, Ralf Tautenhahn, Christoph Böttcher, Tony R Larson, and Steffen Neumann. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical chemistry*, 84(1):283–9, 2012.

[31] Mark J. Van Der Laan, Sandrine Dudoit, and Aad W. Van Der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics & Decisions*, 24:373–395, 2006.

[32] Erin Ledell, Maya Petersen, and Mark J. van der Laan. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic journal of statistics*, 9(2011042):1–20, 2015.

[33] Vernon J. Lee, D C Lye, Y Sun, and Y S Leo. Decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in Singapore. *Tropical medicine & international health*, 14(9):1154–9, sep 2009.

[34] Vernon J. Lee, David C B Lye, Yan Sun, Gina Fernandez, Adrian Ong, and Yee Sin Leo. Predictive value of simple clinical and laboratory variables for dengue hemorrhagic fever in adults. *Journal of Clinical Virology*, 42:34–39, 2008.

[35] Qunhua Li, Jimmy K. Eng, and Matthew Stephens. A likelihood-based scoring method for peptide identification using mass spectrometry. *The Annals of Applied Statistics*, 6(4):1775–1794, dec 2012.

[36] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[37] Gunnar Libiseller, Michaela Dvorzak, Ulrike Kleb, Edgar Gander, Tobias Eisenberg, Frank Madeo, Steffen Neumann, Gert Trausinger, Frank Sinner, Thomas Pieber, and Christoph Magnes. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics*, 16(1):118, 2015.

[38] Sebabrata Mahapatra, Ann M Hess, John L Johnson, Kathleen D Eisenach, Mary a DeGroote, Phineas Gitta, Moses L Joloba, Gilla Kaplan, Gerhard Walzl, W Henry Boom, and John T Belisle. A metabolic biosignature of early response to anti-tuberculosis treatment. *BMC infectious diseases*, 14(1):53, jan 2014.

[39] G. N. Malavige, S. Fernando, D. J. Fernando, and Suranjith L Seneviratne. Dengue viral infections. *Postgraduate Medical Journal*, (1):68–78, 2004.

[40] Natasha Evelyn Anne Murray, Mikkel B. Quam, and Annelies Wilder-Smith. Epidemiology of dengue: Past, present and future prospects. *Clinical Epidemiology*, 5(1):299–309, 2013.

[41] Federico Narvaez, Gamaliel Gutierrez, Maria Angeles Pérez, Douglas Elizondo, Andrea Nuñez, Angel Balmaseda, and Eva Harris. Evaluation of the traditional and revised WHO classifications of Dengue disease severity. *PLoS neglected tropical diseases*, 5(11):e1397, nov 2011.

[42] Nha Nguyen, Heng Huang, Soontorn Oraintara, and An Vo. Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet. *Bioinformatics*, 26(18):i659–65, 2010.

[43] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[45] James J Pitt. Principles and Applications of Liquid Chromatography- Mass Spectrometry in Clinical Biochemistry. *Clinical Biochemistry Review*, 30(February):19–34, 2009.

[46] Tomás Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Oresic. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, 11:395, jan 2010.

[47] James A. Potts, Robert V Gibbons, Alan L Rothman, Anon Srikiatkhachorn, Stephen J Thomas, Pra-On Supradish, Stephenie C Lemon, Daniel H Libraty, Sharone Green, and Siripen Kalayanarooj. Prediction of dengue disease severity among pediatric Thai patients using early clinical laboratory indicators. *PLoS neglected tropical diseases*, 4(8):e769, jan 2010.

[48] James a Potts and Alan L Rothman. Other Febrile Illnesses in Endemic Populations. 13(11):1328–1340, 2009.

[49] James A. Potts, Stephen J. Thomas, Anon Srikiatkhachorn, Pra On Supradish, Wenjun Li, Ananda Nisalak, Suchitra Nimmannitya, Timothy P. Endy, Daniel H. Libraty, Robert V. Gibbons, Sharone Green, Alan L. Rothman, and Siripen Kalayanarooj. Classification of dengue illness based on readily available laboratory data. *American Journal of Tropical Medicine and Hygiene*, 83(4):781–788, oct 2010.

[50] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015.

[51] Christopher Rothwell, Aude LeBreton, Chuan Young Ng, Joanne Y H Lim, Wei Liu, Subhash Vasudevan, Mark Labow, Feng Gu, and L. Alex Gaither. Cholesterol biosynthesis modulation regulates dengue viral replication. *Virology*, 389(1-2):8–19, 2009.

[52] Theodore R. Sana, Joseph C. Roark, Xiangdong Li, Keith Waddell, and Steven M. Fischer. Molecular formula and METLIN personal metabolite database matching applied to the identification of compounds generated by LC/TOF-MS. *Journal of Biomolecular Techniques*, 19(4):258–266, 2008.

[53] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4:Article32, 2005.

[54] Cameron P Simmons, Jeremy J. Farrar, Nguyen van Vinh Chau, and Bridget Wills. Dengue. *The New England Journal of Medicine*, 366:1423–1432, 2012.

[55] CA Smith, J Elizabeth, G O'Maille, Ruben Abagyan, and Gray Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using Nonlinear Peak Alignment, Matching, and Identification. *ACS Publications*, 78(3):779–787, 2006.

[56] Colin a Smith, Grace O'Maille, Elizabeth J Want, Chuan Qin, Sunia a Trauger, Theodore R Brandon, Darlene E Custodio, Ruben Abagyan, and Gary Siuzdak.

METLIN: a metabolite mass spectral database. *Proceedings of the 9th international congress of therapeutic drug monitoring & clinical toxicology*, 27(6):747–751, 2005.

[57] John D Storey. A direct approachto false discovery rates. *J. R. Statist.Soc. B*, 64(3):479–498, 2002.

[58] John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 66(1):187–205, 2004.

[59] Susan L Stramer, F Blaine Hollinger, Louis M Katz, Steven Kleinman, Peyton S Metzel, Kay R Gregory, and Roger Y Dodd. Emerging infectious disease agents and their potential threat to transfusion safety. *Transfusion*, 49:1S—-29S, 2009.

[60] Lukas Tanner, Mark Schreiber, Jenny G H Low, Adrian Ong, Thomas Tolfvenstam, Yee Ling Lai, Lee Ching Ng, Yee Sin Leo, Le Thi Puong, Subhash G Vasudevan, Cameron P Simmons, Martin L Hibberd, and Eng Eong Ooi. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS neglected tropical diseases*, 2(3):e196, jan 2008.

[61] Ralf Tautenhahn, Christoph Böttcher, and Steffen Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, 9:504, jan 2008.

[62] Nguyen Minh Tuan, Ho Thi Nhan, Nguyen Van Vinh Chau, Nguyen Thanh Hung, Ha Manh Tuan, Ta Van Tram, Nguyen Le Da Ha, Phan Loi, Han Khoi Quang, Duong Thi Hue Kien, Sonya Hubbard, Tran Nguyen Bich Chau, Bridget Wills, Marcel Wolbers, and Cameron P. Simmons. Sensitivity and Specificity of a Novel Classifier for the Early Diagnosis of Dengue. *PLOS Neglected Tropical Diseases*, 9(4):e0003638, 2015.

[63] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–21, 2001.

[64] Aad W. Van Der Vaart, Sandrine Dudoit, and Mark J. Van Der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24:351–371, 2006.

[65] Mark J. van der Laan, Sandrine Dudoit, and Sunduz Keles. Asymptotic Optimality of Likelihood-Based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 2004.

[66] Mark J. van der Laan, Eric C Polley, and Alan E Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[67] Mark J. van der Laan and J. M. Robbins. *Unified Methods for Censored Longitudinal Data and Causality.* Springer-Verlag, New York, 2003.

[68] Kirill a. Veselkov, Lisa K. Vingara, Perrine Masson, Steven L. Robinette, Elizabeth Want, Jia V. Li, Richard H. Barton, Claire Boursier-Neyret, Bernard Walther, Timothy M. Ebbels, István Pelczer, Elaine Holmes, John C. Lindon, and Jeremy K. Nicholson. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry*, 83(15):5864–5872, 2011.

[69] Nawaporn Vinayavekhin, Edwin a. Homan, and Alan Saghatelian. Exploring disease through metabolomics. *ACS Chemical Biology*, 5(1):91–103, 2010.

[70] Natalia Voge. *Metabolics-based diagnosis and prognosis of dengue virus infections and NS1 antigen detection for diagnosis and surveillance in humand and mosquitoes.* PhD thesis, Colorado State University, 2013.

[71] WHO. Dengue and severe dengue, 2012.

[72] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807, 2013.

[73] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, and I. Forsythe. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(Database):D603–D610, 2009.

[74] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie-Aude Coutouly, Ian Forsythe, Peter Tang, Savita Shrivastava, Kevin

Jeroncic, Paul Stothard, Godwin Amegbey, David Block, David D Hau, James Wagner, Jessica Miniaci, Melisa Clements, Mulu Gebremedhin, Natalie Guo, Ying Zhang, Gavin E Duggan, Glen D Macinnis, Alim M Weljie, Reza Dowlatabadi, Fiona Bamforth, Derrick Clive, Russ Greiner, Liang Li, Tom Marrie, Brian D Sykes, Hans J Vogel, and Lori Querengesser. HMDB: the Human Metabolome Database. *Nucleic acids research*, 35(Database issue):D521–6, 2007.

[75] World Health Organization. Dengue Guidelines for Diagnosis, Treatment, and Control. Technical report, 2009.

[76] Sophie Yacoub and Bridget Wills. Predicting outcome from dengue. *BMC medicine*, 12(1):147, sep 2014.

[77] Elena Zaitseva, Sung-Tae Yang, Kamran Melikov, Sergei Pourmal, and Leonid V. Chernomordik. Dengue Virus Ensures Its Fusion in Late Endosomes Using Compartment-Specific Lipids. *PLoS Pathogens*, 6(10):e1001131, 2010.

[78] Jianqiu Zhang, Elias Gonzalez, Travis Hestilow, William Haskins, and Yufei Huang. Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry. *Current Genomics*, 10:388–401, 2009.

[79] H Zou and T Hastie. Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society*, 67:301–320, 2005.

Appendix A: Case Report Forms

CONSULTATION CHART: CHILD INTEGRAL CARE STUDY CODE: __/__/__/__ DATE: __ /__ /__

FILE:__/__/__/__/__/__/__ NAME:____ TIME:_________am/pm

Wt_____Kg Ht_____Cm Age____(d / m / a) Sex M F B/P ______ Resp Rate _____ Heart Rate _____ Temp _____°C

Consultation:[ ] Initial [ ] Followup [ ] Conv. Shift:[ ] Reg. [ ] Night [ ] Wkend Time Consult______T Med._____°C

DOS_______ DOF_______Last Day Fever________ am/pm Last dose antipyretic___/___/___ Time:________ am/pm

General State	Y	N	U
Fever			🟩
Weakness/asthenia			🟩
Abnormally sleepy			
Poor general state			
Unconscious			
Restless and irritable			
Convulsions			
Hyptothermia			
Lethargy			

Head			
Headache			🟩
Stiff Neck			
Conjuctival infection			
Subconjuntival hemorrhage			
Retro orbital pain			🟩
Bulging fontanelle			
Conjuntival jaundice			

Throat			
Erhythema			
Sore throat			🟩
Cervical adenopathy			
Exudate			
Petechiae in mucosa			

Respiratory Symtoms			
Cough			🟩
Rhinorrhea			🟩
Nasal congestion			🟩
Ear ache			🟩
Nasal flaring			
Apnea			
Rapid respiration			
Expiratory grunt			
Resting Stridor			
Chest indrawing			
Wheezing			
Rales			
Hoarseness			

Gastrointestinal	Y	N	U
Loss of appetite			🟩
Nausea			🟩
Difficulty eating			🟩
Vomiting (# in the last 12h ___)			🟩
Diarrhea			🟩
Diarrhea with Blood			🟩
Constipation			🟩
Intermitent abdominal pain			🟩
Continuous abdominal pain			🟩
Epigastric pain			🟩
Oral intolerance			🟩
Abdominal distension			🟩
Hepatomegaly (___ cm)			

Dehydration			
Dry tongue/mucosal surfaces			
Poor skin turgor			
Reduced urine output			🟩
Drinks avidly, thirsty			
Sunken eyes			
Sunken fontanelle			

Renal			
Urinary symptoms			🟩
Leucocyturia ≥10 x Field			🟩
Nitrites			🟩
Erythrocytes ≥6xField			🟩
Bilirubinuria			🟩

Referred to pediatrician		
Hospital referral		
Hospital referral for Dengue		
Hospital referral for SARI		

Osteomuscular	Y	N	U
Arthalgia			🟩
Myalgia			🟩
Lumbalgia			🟩
Neck pain			🟩

Cutaneous		
Localized rash		
Generalized Rash		
Ehtrythmatous rash		
Macular ash		
Papular rash		
Mottled skin		
Flushed face		
Echymosis		
Central cyanosis		
Jaundice		

Nutricional state			
Obese			🟩
Overweight			🟩
Suspected problem			🟩
Normal			🟩
Underweight			🟩
Severely underweight			🟩

IMC_______

Breastfeeding			🟩
Vaccinations up-to-date			🟩
Influenza vaccination			🟩
Vaccination date _________			

ILI		
SARI		
Other DOF		
New DOF____________		

Category: [A] [B] [C] [D] [NA] **Change in category: [Yes] [No]**

Complete if Category A or B

	Y	N	U
Hemorrhagic manifestations			
Positive tourniquet test			
Petechiae ≥10 in PT			
Petechiae ≥20 in PT			
Cold skin and extremities			
Paleness in extremities			

	Y	N	U
Epistaxis/Nosebleed			
Gingival bleeding			
Spontaneous Petechiae			
Capillary refill > 2 seg.			
Cyanosis			

Atypical lymphocytes ____ % Date _____

	S	N	D
Hypermenorrhea			🟩
Hematemesis			
Melena			
Hemoconcentration (___ %)			🟩

Have you been hospitalized in the past year [Yes] [No], if Yes specify date and cause

Did you receive a blood transfusion in the past year [Yes] [No], if Yes specify date

Are you currently under medication [Yes] [No], if Yes specify

Did you take any other medication in the past 6 months [Yes] [No], if Yes specify

Digitation Stamp Physician and Supervisor Stamps **Revised 29 June 2011 Version 10**

Exams	Y	N
CBC		
Dengue Serology		
Thick blood smear		
Differential/extended smear		
Urinalysis		
EGH		
Fecal Cytology		
Rheumatoid factor		
Albumin		
AST/ALT		
Bilirubin		
CPK		
Cholesterol		
Influenza		
Others		

Treatment	Y	N
Acetaminophen		
Aspirin		
Ibuprofen		
Antibiotics		
Penicillin		
Amoxicillin		
Dicloxacillin		
Other:___________________		
Furazolidone		
Metronidazole/Tinidazole		
Albendazole/Mebendazole		
Iron Sulfate		
ORS		
Zinc Sulfate		
IV fluids		
Prednisone		
Hydrocortisone IV		
Salbutamol		
Oseltamivir		

History and Physical Exam

Diagnosis

1. ____

2. ____

3. ____

4. ____

Emer. Tele.____________________ Next Appt.:_________________

School: ____

Schedule of Classes: [AM] [PM] [NA]

Position	Code	Date	Time	Signature
Physician				
Nurse				
Supervisor				

Physician and Supervisor Stamps

Plan

Digitation Stamp

Revised 29 June 2011 Version 10

Case Report Form

Prospective Hospital-based Study of Dengue Classification, Case Management and Diagnosis in Nicaragua

File: ________________________

Study Code: ________________________

	YES	NO
CONSENT FOR STUDY PARTICIPATION	□	
SIGNED ASSENT FORM (For children >12 years)	□	
VERBAL ASSENT (For children >5 years)	□	
CONSENT FOR THE LONGITUDINAL STUDY	□	□
CONSENT TO STORE SAMPLES AFTER THE STUDY IS COMPLETED	□	□
DNA CONSENT	□	□
TYPE OF STUDY :		
CLINICAL STUDY	□	
PDVI (CSSFV COHORT STUDY)	□	
WITHDRAWN	□	

Physician Code, Signature, and Date:_______________________________________

Medical Supervisor Code, Signature, and Date:_______________________________

Admission and General Information

Name:

□ Female

Reason for Consultation

File No.: Study Code:

□ Male

□ Fever

Date of Birth(DD/MM/YY) Age years/months Date of Admission into Study

Time:

□ Malaise

Weight Height:

□ Headache

Date Admitted into Hospital:

Relation to Parent/Guardian:

□ Ocular Pain

Name of Parent/Guardian:

□ Body Pain/Aches

Address: Neighborhood: Telephone:

□ Mother

□ Rash

□ Father

□ Bleeding

Department: City: Cell:

□ Grandparent

□ Abdominal Pain

Nutricional Status

Eutrophic □ Malnourished Level 1 □ Malnourished Level 2 □ Malnourished Level 3 □ Overweight □

□ Other Family Member

□ Diarrhea

□ Not a Family Member

□ Vomiting

Pacient Originates from:

□ Primary public □ Secondary Public

□ Primary private □ Secondary Private

□ Primary provisional □ Secondary provisional

□ Home □ Other______

Corresponding Health Unit

Criteria of Hospitalization: (see reverse side of sheet and choose a number)

Other Criteria:

If referred from another hospital, date of admission:

Days Hospitalized:

Illnesses at time of Referral:

Referral:	Yes	No
Pneumonia	□	□
Nosocomial	□	□
Sepsis	□	□

□ Abdominal Discomfort

□ Loss of Consciousness

□ Sleepy/Drowsy

□ Lethargy

□ Irritability

Transfer Motive: Health Unit that Referred:

□ Unconscious

Samples take upon admission □ Red □ Purple □ Leucosep □ Blue

□ Cough

Physician Code, Signature, and Date

Medical Supervisor Code, Signature, and Date

Discharge Summary

Date of Discharge:

Care: □ Hospitalized (In-patient) □ Out-patient

Weight at Discharge

Number of Shock Episodes:

Evolution: □ Died in the Hospital, date and time:___________________ □ Transferred to other hospital, details:_________________

□ Total Recuperation □ Died at Home, date and time:_________________________

□ Incomplete Recuperation □ Abandoned Reason Abandoned:___________________

If follow-up appointment is planned, date of appointment :

Final Clinical Diagnosis

Other Diagnosis:

Type of other diagnosis: □ Concomitant □ Chronic □ Nosocomial

Laboatory Diagnosis:

Diagnosis according to the Revised WHO Classification:

Samples taken upon Discharge: □ Red □ Purple □ Leucosep □ Blue

Physician Code, Signature, and Date

Medical Supervisor Code, Signature, and Date

Follow-up Visit Summary

Date of Follow-up Visit:

Weight at time of visit

Health Status □ Healthy □ Other:_____________________

□ Child did not attend visit

Samples taken: □ Red □ Purple □ Leucosep □ Blue

Physician Code, Signature, and Date

Medical Supervisor Code, Signature, and Date

Note: Print the reverse side of page

Criteria for Hospitalization

1.Children younger than 1 year

2. Pacients with DHF

3. Obesity

4. Oral intolerance

5.Associated with chronic pathologies

6.Dehydration (whichever level)

7.Signs of Shock

- A) Temperature decrease below normal levels
- B) Cold and clammy skin
- C) Extreme paleness
- D) Oliguria
- E) Tachycardia
- F) Shortened pulse pressure (BP differential <20 mm Hg)
- G) Hypotension

8.Restlessness or Weakening

9.Evidence of capillary leakage

- A) Hemoconcentration
- B) Pleural effusion
- C) Ascites
- D) Marked decrease in frequency and amount of urine

10.Presence of Warning Signs

- A) Sustained and Intense Abdominal Pain
- B) Abdominal Distension
- C) Pain in Thorax
- D) Hepatomegaly
- E) Frequent Vomiting - Doesn't tolerate ORS
- F) Dificulty Breathing
- G) Bleeding from any part of body
- H) Trombocitopenia < 100,000 mm3

File No.________________ Study Code:________________

Clinical History

Date of Onset of Symptoms:_____/_____/_____ AM/PM Date of Onset of Fever: _______/________/________ AM / PM

Date of last menstruation ______________________________

		U	Y	N	How many days?
Fever		□	□	□	1 2 3 4 5 6 7
Has the patient been without fever at any moment during this episode of fever?		□	□	□	1 2 3 4 5 6 7
If affirmative, enter the date:	____/____/____				
Vomiting		□	□	□	1 2 3 4 5 6 7
If affirmative	□ only once				
How many times?	□ 2 or more				
	□ continuous				
Tendency to vomit while drinking?		□	□	□	1 2 3 4 5 6 7
While eating?		□	□	□	1 2 3 4 5 6 7
Cough		□	□	□	1 2 3 4 5 6 7
If affirmative, is it bloody?		□	□	□	1 2 3 4 5 6 7
Common cold?		□	□	□	1 2 3 4 5 6 7
Difficulty Breathing?		□	□	□	1 2 3 4 5 6 7
Loss of Appetite		□	□	□	1 2 3 4 5 6 7
If yes, still eating solids?		□	□	□	
If yes, still drinking liquids?		□	□	□	
Retroorbital pain		□	□	□	1 2 3 4 5 6 7
Abdominal pain		□	□	□	1 2 3 4 5 6 7
Osteomuscular pain		□	□	□	1 2 3 4 5 6 7
Articular pain		□	□	□	1 2 3 4 5 6 7
Irritability		□	□	□	1 2 3 4 5 6 7
Diarrhea		□	□	□	1 2 3 4 5 6 7
Mucus in the feces?		□	□	□	1 2 3 4 5 6 7
Blood in the feces?		□	□	□	1 2 3 4 5 6 7
Has the paciente drank more water than usual?		□	□	□	1 2 3 4 5 6 7
Mucosal bleeding					
Gums		□	□	□	1 2 3 4 5 6 7
Nose		□	□	□	1 2 3 4 5 6 7
Hematemesis		□	□	□	1 2 3 4 5 6 7
Melaena		□	□	□	1 2 3 4 5 6 7
Vaginal		□	□	□	1 2 3 4 5 6 7
Bleeding of the skin		□	□	□	1 2 3 4 5 6 7
Rash		□	□	□	1 2 3 4 5 6 7

Pathologic Background

	Yes	No
Peptic Ulcer	□	□
Asthma	□	□
Anemia	□	□
Diabetes	□	□
Allergy	□	□
Arterial Hypertension	□	□
Cardiac Disease	□	□
Kidney Disease	□	□
Tuberculosis	□	□
Chronic Hepatitis	□	□
Other chronic disease:		
Regular medications:		

Previos management

	Yes	No
Water	□	□
ORS	□	□
Tea	□	□
Milk	□	□
Juice	□	□
IV Rehydration	□	□
IV Blood	□	□
Others:		

Medications

	U	No	Yes
Aspirine	□	□	□
Ibuprofen	□	□	□
Acetaminophen	□	□	□
Diclofenac	□	□	□
Antibiotics	□	□	□
Vitamin C	□	□	□
Multivitamin	□	□	□
Furosemide	□	□	□
Other:			

Physician Code, Signature, and Date

Medical Supervisor Code, Signature, and Date

Case Report Form - Hospital Infantil Manuel de Jesús Rivera

Vital Signs Report

Date: ______/______/_______ **Date of Onset of Symptoms: ______/______/_____ Use of 12 hours (circle as appropriate) : 6 AM / PM to 6 PM / AM**

File No.: ________________________ **Study Code:__________________________** **Age_________ years/months***

First and Last Names:____ **Date of defervescence:____**

Male / Female **Height___________cm** **Weight:______________ kg** **BSA**

	To be Completed by Nurse/Physician													**Code &Signature**
Nurse & Phys	Code													
	Time													
Vital Signs	Temperature													
	Pulse[1]													
	Blood Pressure													
	Breathing Rate													
	Heart Rate													
Shock Information	Cap. Refill[2]													
	Extremities													
	Warm and Pink													
	Distal coldness													
	Cold and clammy													
	Dehydration													
	Oral Cyanosis													
	Perif Cyanosis													
	Has urinated													
	Oxygen Saturation													
	Level of Conscious.[3]													
Others	Number of Vomit													
	Hematocrit													
	Shock[4]													

Samples taken for Laboratory:

Time taken: _________________

	Yes	No		Yes	No		Yes	No
Red	❑	❑	Leucosep	❑	❑	Purple	❑	❑
Blue	❑	❑						

Principal Diagnosis					**Diagnosis according to the Revised Classification**
	Concomitant		❑		
Other Diagnosis	**Chronic**	❑	**Nosocomial**	❑	

[1] Specify (F) Strong and vigorous, (D) Weak and not vigorous, (A)Absent

[2] Specify "<2" seconds or ">2" seconds

[3] Specify 1- Clear and Lucid 2- Restless 3- Lethargic

[4] Specify (C) Compensated (D) Uncompensated

*** Si es menor de 1 año anotar la edad en meses**

Treating Physician Code, Signature, and Date ____

Medical Supervisor Code, Signature, and Date ____

6/24/2011

Case Report Form - Hospital Infantil Manuel de Jesús Rivera

File No.: Study Code:

Clinical Information

Date: ______/______/_______ Use of 12 hours (circle as appropriate) : 6 AM / PM to 6 PM / AM

To be Completed by Physician

Intensive Therapy □ **Infectology** □ **her Area:**

Breathing	Y	N
Difficulty	❑	❑
Type of Breathing		
❑ Reg ❑ Irreg ❑ Shallow		
❑ Deep ❑ Dissociated		
Cough	❑	❑
Pneumonia	❑	❑
Ronchi	❑	❑
Rales	❑	❑
Wheezing	❑	❑
Pleural Effusion	❑	❑
Right	❑	❑
Left	❑	❑
Intercostal Retractions	❑	❑
Thoracic Pain	❑	❑
Abdomen	Y	N
Abdominal Sensitivity	❑	❑
Abdominal Pain	❑	❑
Intermitant	❑	❑
Continuous	❑	❑
Epigastralgia	❑	❑
Abdominal Distension	❑	❑
Ascites	❑	❑
Icterus	❑	❑
Liver (cms)		
Spleen (cms)		
Drinks normally?	❑	❑
Eats normally?	❑	❑
Oral Intolerance	❑	❑
Liquid Stools	❑	❑
Number of Liquid Stools		

In the Previous 12 hours:	Y	N
Shock	❑	❑
Time of Shock:		
Signs of overhydration	❑	❑
Signs of dehydration	❑	❑
Cardiovascular	Y	N
Edema	❑	❑
Periorbital	❑	❑
Facial	❑	❑
Inferior Membranes	❑	❑
Hydrocele	❑	❑
Generalized	❑	❑
Pitting	❑	❑
Rash	Y	N
Cutaneous exanthem	❑	❑
Macular	❑	❑
Papular	❑	❑
Maculo papular	❑	❑
Erythema	❑	❑
Measles-like	❑	❑
Facial flushing	❑	❑
CNS	Y	N
Stiff neck	❑	❑
Meningismus	❑	❑
Lightheadedness [3]	❑	❑
Lethargy	❑	❑
Irritability	❑	❑
Convulsions	❑	❑
Glasgow (O,V,M) ___,___,___ (>5 years)		
Blantyre (O,V,M) ___,___,___ (<5years)		
Level of care(4)		

Bleeding	Y	N
Clinically significant?	❑	❑
Petechiae	❑	❑
Purpura	❑	❑
Ecchymosis/bruising	❑	❑
Hematoma	❑	❑
Hemoptysis	❑	❑
Epistaxis/Nosebleed	❑	❑
Gums	❑	❑
Melena	❑	❑
Hematemesis	❑	❑
Hematuria	❑	❑
Subconjuntival	❑	❑
Vaginal	❑	❑
Hipermenorrea	❑	❑
Venopuncture	❑	❑
Tourniquet Test[5]	10 20 N	
Lymphadenopathy[6]		
Procedures	Si	No
Ventilation	❑	❑
Dialysis	❑	❑
PVC	❑	❑
Ascitic fluid drainage	❑	❑
Use of parenteral fluids		
For Shock	❑	❑
Rehydration	❑	❑
Maintenance/Warning Signs	❑	❑
Use of Oxygen	❑	❑
Nebulized	❑	❑
Use of Inotropic drugs	❑	❑
Medication for liver failure	❑	❑
Diuretic	❑	❑

Treating Physician Code, Signature, and Date___________________________

Medical Supervisor Code, Signature, and Date ___________________________

Glasgow Scale (> 5 years)

Ocular: (1) Doesn't open eyes, (2) Opens in response to painful stimuli, (3) Opens on verbal command, (4) Opens spontaneously

Verbal: (1) Makes no sounds, (2) Incomprehensible sounds, (3) Inappropriate words, (4) Confused (5) Oriented, converses normally

Motor: (1) Makes no movementes, (2) Extension to painful stimulo, (3) Flexion to painful stimuli, (4) Withdrawal to painful stimuli, (5) Localizes painful stimuli, (6) Obeys commands

Blantyre Scale (< 5 years)

Ocular: (0) Fails to watch or follow, (1) Watches or follows

Verbal: (0) No response, (1) Abnormal cry with pain, (2) Cries appropriately with pain

Motora: (0) No response, (1) Withdraws from pain, (2) Localizes pain

[3] For pacientes older than 5 years

[4] Level of clinical care: 1 Basic hospital care. 2 Intermediate care. 3 Maximum care

[5] Number of petechiae per sq. inch

[6] Specify: (0) None (1) cervical (2) axilar (3) submandibular (4) inguinal (5) occipital (6) generalized

Revisado 22/07/09

Case Report Form - Hospital Infantil Manuel de Jesús Rivera

Laboratory and Office 2

Use of 12 hours (circle as appropriate) : 6 AM / PM to 6 PM / AM

File No.__________________________

Study Code: ____________________ **Date: ______/______/___**

Radiologic Signs		
Echocardiogram	**Yes**	**No**
Normal	□	□
Shortening fraction	□	□
%		
LV end-diastolic diameter	□	□
mL/m$_{2d}$		
LV end-systolic diameter	□	□
mL/m2d		
E Wave		
A Wave		
Ratio E/A		
Cardiac Index	□	□
L/min/m2		
Paracardial Effusion	□	□
Diameter:		
Ventricular disfuntion	□	□
Systolic:		
Diastolic:		
Time:		
No. Phys./Rad.		
Electrocardiogram	**Yes**	**No**
Normal	□	□
Rhythm		
Cardiac Frequency		
QRS Duration		
PR		
QTc		
T Wave		
ST-T		
Time:		
No. Phys./Rad.		

Shock	**Yes**	**No**
Refractory liquids	□	□
Crystalloids	□	□
Colloids	□	□
Refractory amines	□	□
Recurrent [1]	□	□
Prolonged [2]	□	□

Treating Physician Code, Signature, and Date________________________

Medical Supervisor Code, Signature, and Date _______________________

[1] Pacient persists with warning signs although adequate liquids are given

[2] Pacient persists with shock after 6 hours of treatment or doesn't improve after ≥ 60 ml/kg IV fluids are given.

Case Report Form - Hospital Infantil Manuel de Jesús Rivera

Laboratory and Office

Use of 12 hours (circle as appropriate) : 6 AM / PM to 6 PM / AM

File No.__________________________

Study Code: ___________________ **Date:** ______/______/___

Radiologic Signs		
Ultrasound	**Yes**	**No**
Normal	□	□
Vesicular Wall Thickening	□	□
mm:		
Perivesicular fluid	□	□
Hepatomegaly	□	□
mm:		
Splenomegaly	□	□
mm:		
Ascites	□	□
CC:		
Para/perirenal fluid	□	□
Pleural effusion	□	□
Right (Volume):		
Left (Volume):		
Pericardial effusion	□	□
Volume:		
Pulmonary Edema	□	□
Time:		
No. Phys./Rad.		
X-Rays[8] **PA □ AP □ PC □ No □**	**Yes**	**No**
Normal	□	□
Left pleural effusion	□	□
Right pleural effusion	□	□
Pleural Effusion Index [9]		
Pulmonary Edema	□	□
Interstitial	□	□
Alveolar	□	□
Pneumonia	□	□
Cardiomegaly	□	□
Cardiac Index		
Time:		
No. Phys./Rad.		
Interpreted by:		
Clinical Interpretation	□	□
Radiologic Interpretation	□	□

Clinical Laboratory	
Leukocytes (x1,000)	
Segmented (%)	
Lymphocytes (%)	
Monocytes (%)	
Eosinophils (%)	
Platelets(x1,000)	
Hematocrit (%)	
Hemoglobin(gr)	
Atypical Lymphocytes (%)	
ESR (mm/s)	
Troponin	
PT (s)	
PTT (s)	
AST (UI)	
ALT (UI)	
Total Bilirubin (mg/dL)	
Indirect Bilirubin(mg/dL)	
Direct Bilirubin (mg/dL)	
Cholesterol (mg/dL)	
HDL (mg/dL)	
LDL (mg/dL)	
LDH(uL)	
CPK (uL)	
Tot. Protein (g/dL)	
Albumin (g/dL)	
Globulin (g/dL)	
Ratio A/G	
Creatinine (mg/dl)	
K+ (mmol/dL)	
Ca+ (mg/dL)	
Na+ (mmol/dL)	
Cl- (mmol/dL)	

Uroanalysis	**Yes**	**No**
Urine test (7)		
Microscopic hematuria		
ABG	**Yes**	**No**
Resp. Acidosis	□	□
Resp. Alkalosis	□	□
Metab. Acidosis	□	□
Metab. Alkalosis	□	□

Treating Physician Code, Signature, and Date____________________________

Medical Supervisor Code, Signature, and Date ___________________________

[7] 0= no blood 1= 10 erythrocytes 2= Hematuria

[8] PA: Postero-Anterior, AP: Antero-Posterior, PC: Pancoast

[9] The largest diameter of pleural effusion, divided by the diameter of the right hemothorax, multiplied by 100.

External Consultation Form

Clinical Information

File No._______________________ **Study Code :** ________________

First and Last Names:____ **Date:**______________________

Age_____**years/months** **Height:**_______**cm** **Weight:**______**kg** **BSA:**_____**m2** **Diagnóstico probable en admision: DF / DHF / DSS**

Date of onset of symptoms (DD/MM/YY):_____/______/______**AM / PM** **Date of defervescence:**_____

Breathing	**Y**	**N**	**U**
Difficulty	❑	❑	❑
Respiratory Frequency			
Type of Respiration			
❑ Reg ❑ Irreg ❑ Shallow			
❑ Deep ❑ Dissociated			
Cough	❑	❑	❑
Pneumonia	❑	❑	❑
Ronchi	❑	❑	❑
Rales	❑	❑	❑
Wheezing	❑	❑	❑
Pleural Effusion	❑	❑	❑
Right	❑	❑	❑
Left	❑	❑	❑
Intercostal Retractions	❑	❑	❑
Thoracic Pain	❑	❑	❑
Abdomen	**Y**	**N**	**U**
Abdominal Sensitivity	❑	❑	❑
Abdominal Pain	❑	❑	❑
Intermitant	❑	❑	❑
Continuous	❑	❑	❑
Epigastralgia	❑	❑	❑
Abdominal Distension	❑	❑	❑
Ascites	❑	❑	❑
Icterus	❑	❑	❑
Liver (cms)			
Spleen (cms)			
Drinks normally?	❑	❑	❑
Eats normally?	❑	❑	❑
Oral Intolerance	❑	❑	❑
Liquid Stools	❑	❑	❑
Number of Liquid Stools			
Vomiting[1]			

Shock	**Y**	**N**	**U**
Shock	❑	❑	❑
Peripheral Cyanosis	❑	❑	❑
Oral Cyanosis	❑	❑	❑
Mottled skin	❑	❑	❑
Distal coldness	❑	❑	❑
Cardiovascular	**Y**	**N**	**U**
Capillary Refill (s)[2]			
Edema	❑	❑	❑
Periorbital	❑	❑	❑
Facial	❑	❑	❑
Inferior membranes	❑	❑	❑
Hydrocele	❑	❑	❑
Generalized	❑	❑	❑
Pitting	❑	❑	❑
Rash	**Y**	**N**	**U**
Cutaneous exanthem	❑	❑	❑
Macular	❑	❑	❑
Papular	❑	❑	❑
Maculo papular	❑	❑	❑
Erythma	❑	❑	❑
Measles-like	❑	❑	❑
Facial flushing	❑	❑	❑
CNS	**Y**	**N**	**U**
Stiff Neck	❑	❑	❑
Meningismus	❑	❑	❑
Lightheadedness [3]	❑	❑	❑
Lethargy	❑	❑	❑
Irritability	❑	❑	❑
Convulsions	❑	❑	❑
Glasgow (O,V,M) ___,___,___ (>5 years)			
Blantyre (O,V,M) ___,___,___ (<5years)			

Bleeding	**Y**	**N**	**U**
Bleeding	❑	❑	❑
Clinically significant?	❑	❑	❑
Petechiae	❑	❑	❑
Purpura	❑	❑	❑
Ecchymosis/bruising	❑	❑	❑
Hematoma	❑	❑	❑
Hemoptysis	❑	❑	❑
Epistaxis/Nosebleed	❑	❑	❑
Gums	❑	❑	❑
Melena	❑	❑	❑
Hematemesis	❑	❑	❑
Hematuria	❑	❑	❑
Subconjuntival	❑	❑	❑
Vaginal	❑	❑	❑
Hipermenorrea	❑	❑	❑
Venopuncture	❑	❑	❑
Tourniquet Test[5]	10 20 N		
Lymphadenopathy[6]			
General Information	**Y**	**N**	**U**
Fever	❑	❑	❑
Retroorbital Pain	❑	❑	❑
Headache	❑	❑	❑
Dehydration	❑	❑	❑
Hyporexia	❑	❑	❑
Urine (last 6 hours)	❑	❑	❑
Myalgia	❑	❑	❑
Arthalgia	❑	❑	❑
Vital Signs			
Temperature			
Cardiac Frequency			
Pulse[6]			
Presión Arterial			

Treating Physician Code, Signature, and Date___________________________

Medical Supervisor Code, Signature, and Date ___________________________

Glasgow Scale (> 5 years)

Ocular: (1) Doesn't open eyes, (2) Opens in response to painful stimuli, (3) Opens on verbal command, (4) Opens spontaneously

Verbal: (1) Makes no sounds, (2) Incomprehensible sounds, (3) Inappropriate words, (4) Confused, (5) Oriented, converses normally

Motor: (1) Makes no movementes, (2) Extension to painful stimulo, (3) Flexion to painful stimuli, (4) Withdrawal to painful stimuli, (5) Localizes painful stimuli, (6) Obeys commands

Blantyre Scale (< 5 years)

Ocular: (0) Fails to watch or follow, (1) Watches or follows

Verbal: (0) No response, (1) Abnormal cry with pain, (2) Cries appropriately with pain

Motora: (0) No response, (1) Withdraws from pain, (2) Localizes pain

3

[1] Specify (1) One occasion, (2) Two or more, (3) Continuous

[2] Specify"<2" seconds or ">2" seconds

[3] For pacients older than 5 years

[4] Number of petechiae per sq. inch

[5] Specify: (0)None (1) cervical (2) axilar (3) submandibular (4) inguinal (5) occipital (6) generalized

[6] Specify (F) Strong, (M) Moderate, (R) Rapid, (N) Impalpable

6/24/2011

Hospital Follow-up Form

Study Code: ___________

Use of 12 hours (circle as appropriate) : 6 AM / PM to 6 PM / AM

IV or PO Drug Administration

Time	Type	Amount in CC	Code and Signature

Treating Physician Code, Signature, and Date ___________________________________

Medical Supervisor Code, Signature and Date __________________________________

6/24/2011

Case Report Form - Hospital Infantil Manuel de Jesús Rivera

Laboratory and Office

Use of 12 hours (circle as appropriate) : 6 AM / PM to 6 PM / AM

File No.__________________________ **Study Code :** ___________________

Date: ______/______/___

Ultrasound	Y	N
Ultrasound performed?	□	□
Normal	□	□
Vesicular Wall Thickening	□	□
mm:		
Perivesicular fluid	□	□
Hepatomegaly	□	□
mm:		
Splenomegaly	□	□
mm:		
Ascites	□	□
CC:		
Para/perirenal fluid	□	□
Pleural effusion	□	□
Right (Volume):		
Left (Volume):		
Pericardial effusion	□	□
Volume:		
Pulmonary Edema	□	□
Time:		
No. Phys/Rad.		

X-Rays [8]	PC		PA		AP	
	Yes	No	Yes	No	Yes	No
Performed?	□	□	□	□	□	□
Normal	□	□	□	□	□	□
Left pleural effusion	□	□	□	□	□	□
Right pleural effusion	□	□	□	□	□	□
Pleural effusion index[9]						
Pulmonary edema	□	□	□	□	□	□
Interstitial	□	□	□	□	□	□
Alveolar	□	□	□	□	□	□
Pneumonia			□	□	□	□
Cardiomegaly			□	□	□	□
Cardiac Index						
Time:						
No. Phys./Rad.						
Interpreted by:						
Clinical Interpretation	□	□	□	□	□	□
Radiologic Interpretation	□	□	□	□	□	□

Clinical Laboratory			
Leukocytes (x1,000)		Direct Bilirubin	
Segmented (%)		Cholesterol (mg/dL)	
Lymphocytes (%)		HDL (mg/dL)	
Monocytes(%)		LDL (mg/dL)	
Eosinophils(%)		LDH (uL)	
Platelets(x1,000)		CPK (uL)	
Hematocrit (%)		Tot. Protein (g/dL)	
Hemoglobin (gr)		Albumin (g/dL)	
Atypical lymphocytes (%)		Globulin (g/dL)	
ESR (mm/s)		Ratio A/G	
Troponin		Creatinine (mg/dl)	
PT (s)		K+ (mmol/dL)	
PTT (s)		Ca+ (mg/dL)	
AST (ui)		Na+ (mmol/dL)	
ALT (ui)		Cl- (mmol/dL)	
Total Bilirubin			
Indirect Bilirubin			

Uroanalysis	Yes	No
Performed?	□	□
Urine test [7]		
Microscopic hematuria		
ABG	**Yes**	**No**
Performed?	□	□
Resp. Acidosis	□	□
Resp. Alkalosis	□	□
Metab. Acidosis	□	□
Metab. Alkalosis	□	□
Samples take for Lab.	**Yes**	**No**
Red	□	□
Leucosep	□	□
Purple	□	□
Blue	□	□
Time taken		

Treating Physician Code, Signature, and Date______________________________

Medical Supervisor Code, Signature, and Date ____________________________

[7] 0= no blood 1= 10 erythrocytes 2= Hematuria

8 PA: Postero-Anterior, AP: Antero-Posterior, PC: Pancoast

9 The largest diameter of pleural effusion, divided by the diameter of the right hemothorax, multiplied by 100.

MANAGEMENT OF PACIENT WITH DENGUE - INFECTOLOGY DEPT.

TRACKING SHEET FOR ORAL INTAKE AND EXCRETIONS (FECES, URINE)

FOR 12-HOUR USE

DATE

Beginning of rotation; 6AM or 6PM

Time	Alimentation				Feces		Urine	Vomiting	Personnel	
	Type of formula	Amount ordered (cc)	Oral gavage or	Amount taken (cc)	Normal (cc)	Abnormal (cc)	Amount (cc)	Amount (cc)	Signature	Code
Total										

Name of pacient:	File No.	Study Code

Treating physician code, signature and date
Medical resident code, signature, and date
Medical supervisor code, signature, and date

Case Report Form - Hospital Infantil Manuel de Jesús Rivera

MANAGEMENT OF PACIENT WITH DENGUE - INFECTOLOGY DEPT.
PERENETERAL FLUIDS
(HEMODERIVATIVES / CRYSTALLIODS) FOR 12-HR USE

DATE   Beginning of rotation; 6 AM or 6 PM

Inicial Hct	Time at inicial	Type of liquids	Calculated dose (cc/kg/h)	Completion time	Total Amount Absorbed (cc)	Final Hct	Personnel	
							Signature	Code
Total								

Name of pacient:	File No.	Study code
Treating physician code, signature, and date		
Medical resident code, signature, and date		
Medical supervisor code, signature, and date		

6/24/2011

Appendix B: Supplementary Tables and Figures

Name	Category	Type	Description	In cohort
Age	Demographic	continuous	Patient age on day of consult	1
Gender	Demographic	binary	Gender	1
Abdominal pain	Gen. symptom	binary	Abdominal pain	1
Arthralgia	Gen. symptom	binary	Joint pain	1
Ascites - observed	General sign	binary	Accumulation of fluid in peritoneal cavity, observed without ultrasound	0
Chills	Gen. symptom	binary	Chills	0
Cold extremities	General sign	binary	Cold extremities	1
Cough	Gen. symptom	binary	Cough	1
Cyanosis	General sign	binary	Blue or purple skin coloration due to tissues lacking oxygen	1
Day of illness	General	continuous	Days since fever onset (day 1 is first day of fever)	1
Diastolic blood pressure	General sign	continuous	Minimum pressure in arteries	1
Difficulty breathing	Gen. symptom	binary	Difficulty breathing	1
Headache	Gen. symptom	binary	Headache	1
Heart rate	General sign	continuous	Heart rate (beats per minute)	1
Jaundice	General sign	binary	Jaundice (indicator of hemoglobin breakdown)	1
Liver enlargement	General sign	continuous	Liver enlargement	1
Myalgia	Gen. symptom	binary	Muscle pain	1
Pallor	General sign	binary	Unhealthy paleness	1
Pleural effusion	General sign	binary	Excessive fluid around lung	0
Poor appetite	Gen. symptom	binary	Poor appetite	1
Poor capillary refill	General sign	binary	Takes ≥ 2 sec for color to return capillary bed after applied pressure	1
Pulse	General sign	categorical	Strong, moderate, rapid, or not palpable	0
Rash	General sign	binary	Rash	1
Respiratory rate	General sign	continuous	Respiratory rate (breaths per minute)	1
Retro-orbital pain	Gen. symptom	binary	Pain behind eyes	1
Sore throat erythema	Gen. symptom	binary	Sore throat	1
Sweating	General sign	binary	Sweating	0
Systolic blood pressure	General sign	continuous	Peak pressure in arteries (when ventricles contract)	1
Tachycardia	General sign	binary	Rapid heart beat	1
Temperature	General sign	continuous	Temperature (Celsius)	1
Vomiting	General sign	binary	Vomiting	1

Table 1: Descriptions of basic clinical variables - demographics & general signs and symptoms

Name	Category	Type	Description	In cohort
Bleeding gums	General - hemorrhagic	binary	Bleeding gums	1
Ecchymosis	General - hemorrhagic	binary	Hematoma that is 1 centimeter or larger	1
Epistaxis	General - hemorrhagic	binary	Nose bleeding	1
Hematoma	General - hemorrhagic	binary	Localized collection of blood outside the blood vessels (encompasses ecchymoses, petechiae, and purpura)	0
Hematuria	General - hemorrhagic	binary	Blood in urine	0
Hemetemesis	General - hemorrhagic	binary	Vomiting of blood	1
Hemoptysis	General - hemorrhagic	binary	Spitting up blood from respiratory tract	0
Melena	General - hemorrhagic	binary	Tarry feces caused by upper gastrointestinal bleeding	1
Petechiae	General - hemorrhagic	binary	Red or purple spots on skin that are less than 2 mm in diameter	1
Purpura	General - hemorrhagic	binary	Red or purple discoloration on skin 2 mm - 1 cm in diameter	0
Subconjuntival bleeding	General - hemorrhagic	binary	Bleeding of the eyes	1
Tourniquet test	General - hemorrhagic	categorical	Test of capillary fragility by counting number of petechiae resulting from specific application of blood pressure cuff	1
Vaginal bleeding	General - hemorrhagic	binary	Vaginal bleeding	0
Venipuncture bleeding	General - hemorrhagic	binary	Excessive bleeding from puncture with needle	0

Table 2: Descriptions of basic clinical variables - general signs of hemorrhaging

Name	Category	Type	Description	In cohort
Eosinophil (%)	Blood - count	continuous	Concentration of eosinophils (a type of white blood cell)	1
Granulocytes (%)	Blood - count	continuous	Concentration of granulocytes, a type of white blood cell. Also known as percent neutrophils.	1
Hemoconcentration	Blood - count	binary	Decreased fluid content of blood	1
Hemoglobin (gr)	Blood - count	continuous	Concentration of hemoglobin, the component of red blood cells that carries oxygen	1
Lymphocytes (%)	Blood - count	continuous	Combined concentration three subtypes of white blood cells	1
Monocytes (%)	Blood - count	continuous	Concentration of monocytes, a type of white blood cell	1
Platelet count	Blood - count	continuous	Blood platelet count (1,000 per mm^3)	1
White blood cell count	Blood - count	continuous	White blood cell count (1000 per mm^3)	1
Serotype	Blood - virology	categorical	Serotype of dengue virus	1
Micro-hematuria	Urine - count	continuous	Count of erythrocytes (red blood cells in urine)	1

Table 3: Description of lab clinical variables - blood and urine counts and virology

Name	Category	Type	Description	In cohort
Albumin (g/dL)	Blood - chemistry	continuous	The main protein in blood.	0
Albumin/Globulin ratio	Blood - chemistry	continuous	Relevant to the kidney	0
ALT (IU)	Blood - chemistry	continuous	Alanine aminotransferase (ALT) - could indicate liver inflammation	0
AST (IU)	Blood - chemistry	continuous	Aspartate aminotransferase (AST) - could indicate liver inflammation	0
Atypical lymphocytes (%)	Blood - chemistry	continuous	Atypical white blood cells	0
Cholesterol (mg/dL)	Blood - chemistry	continuous	HDL+LDL+LDH	0
CPK MB (uL)	Blood - chemistry	continuous	Level of creatine phosphokinase (CPK)	0
CPK (uL)	Blood - chemistry	continuous	Enzyme from heart and muscle – indicates breakdown of these tissues	0
Creatinine (mg/dl)	Blood - chemistry	continuous	Well-functioning kidneys should keep creatinine levels low	0
Direct bulirubin (mg/dL)	Blood - chemistry	continuous	Bi-product of metabolism of the liver	0
Globulin (g/dL)	Blood - chemistry	continuous	Concentration of globulin, a major blood protein	0
HDL (mg/dL)	Blood - chemistry	continuous	Concentration of high-density lipoprotein cholesterol	0
Indirect bilirubin (mg/dL)	Blood - chemistry	continuous	Bi-product of metabolism of the liver	0
LDH (uL)	Blood - chemistry	continuous	Lactate dehydrogenase, an enzyme which commonly marks injury and disease	0
LDL (mg/dL)	Blood - chemistry	continuous	Low-density lipoprotein cholesterol	0
Total bilirubin (mg/dL)	Blood - chemistry	continuous	Bi-product of metabolism of the liver	0
Total protein (g/dL)	Blood - chemistry	continuous	Combination of albumin and globulin (low means malnurishment or shock)	0

Table 4: Description of lab clinical variables - blood chemistry

Name	Category	Type	Description	In cohort
Ascitis - ultrasound	Ultrasound	binary	Accumulation of fluid in peritoneal cavity, indicated by ultrasound	0
Esplenomegalia (mm)	Ultrasound	continuous	Enlarged spleen	0
Fluid para/peri renal	Ultrasound	binary	Fluid in or around the kidney	0
Gallbladder (mm)	Ultrasound	continuous	Gallbladder wall thickening	0
Hepatomegalia (mm)	Ultrasound	continuous	Enlarged liver	0
Perivesicular fluid	Ultrasound	binary	Fluid around cavities	0
Pulmonar edema	Ultrasound	binary	Enlargement of the lungs due to fluid build-up	0
Alveolar	X-ray	binary	Pertaining to the lungs	0
Interstitial fluid	X-ray	binary	Fluid in space between cells	0
Pneumonia	X-ray	binary	Lung inflammation	0

Table 5: Description of costly variables

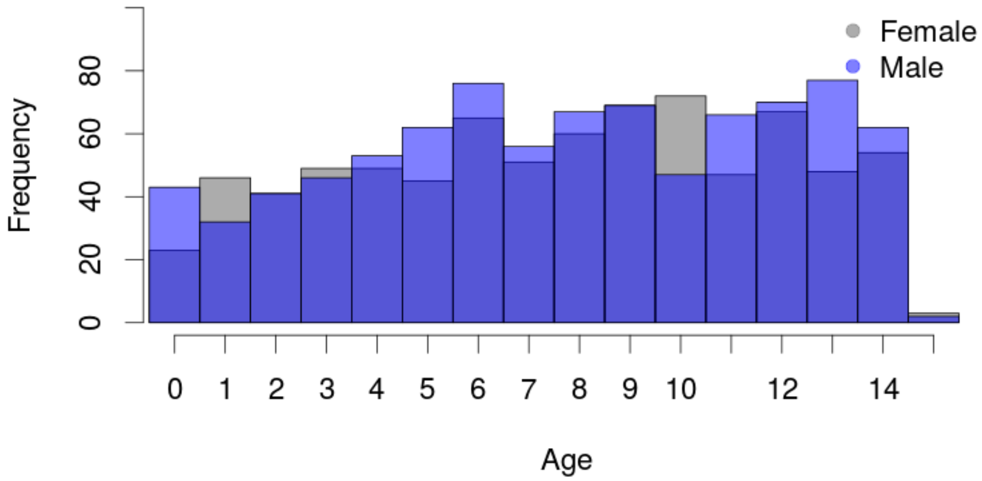

Figure 7: Age and gender distributions - hospital patients

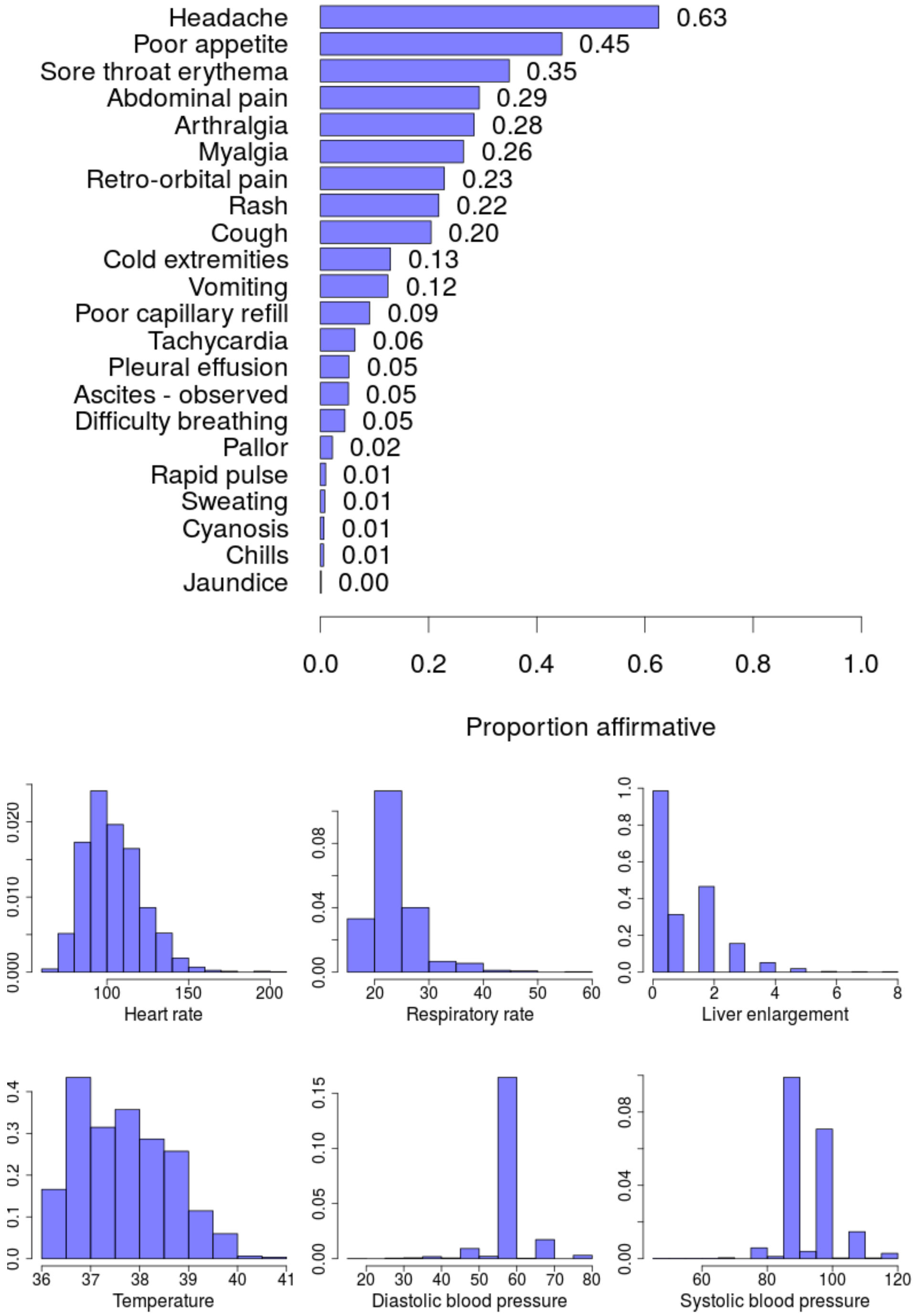

Figure 8: Basic clinical indicators - hospital patients

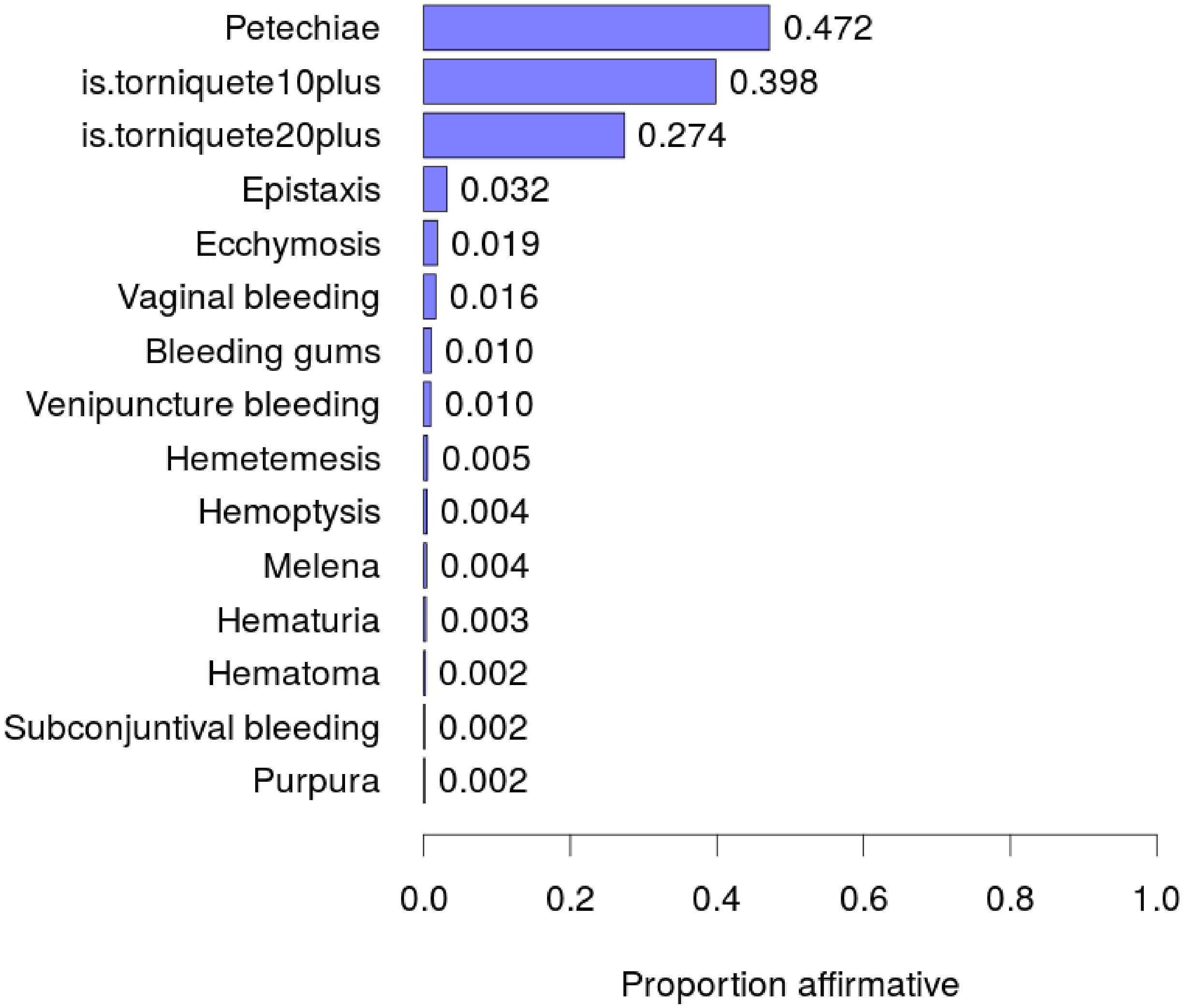

Figure 9: Clinical indicators of hemorrhaging - hospital patients. We find that each of the 15 indicators of hemorrhaging are found in fewer than 5% of hospital patients, with the exception of petechiae and the tourniquete test.

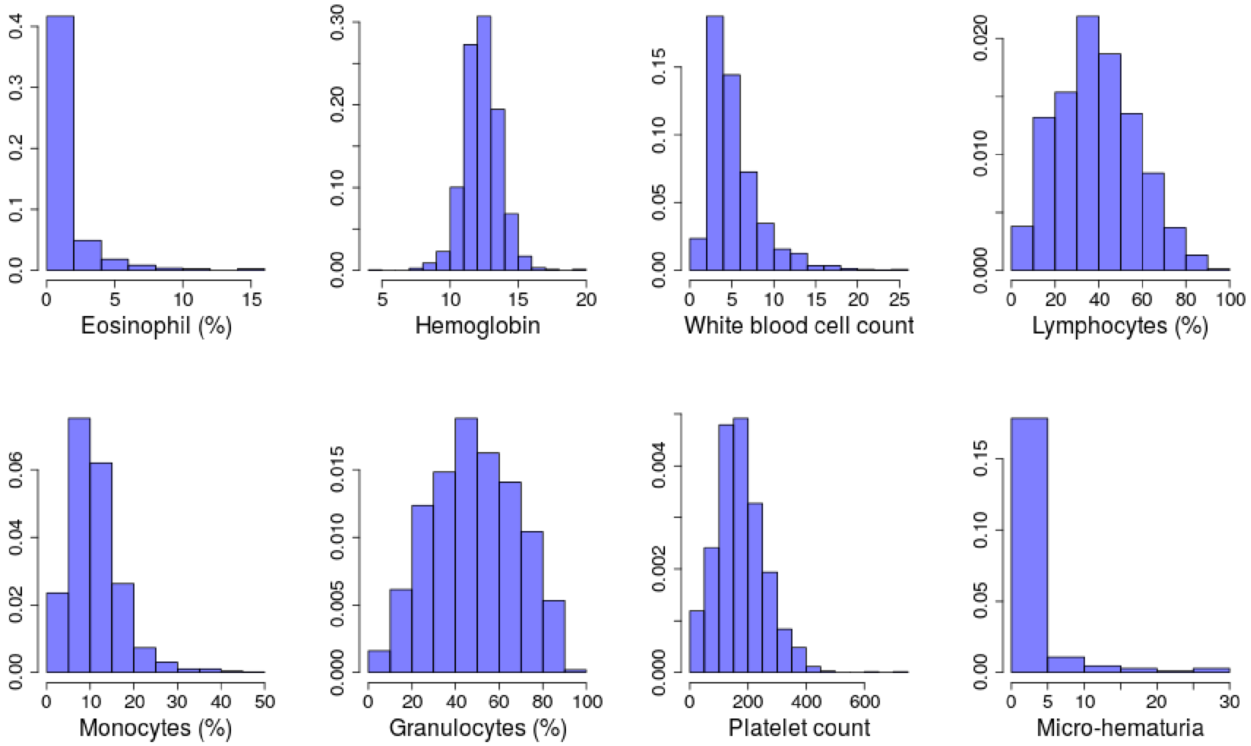

Figure 10: Blood and urine counts - hospital patients

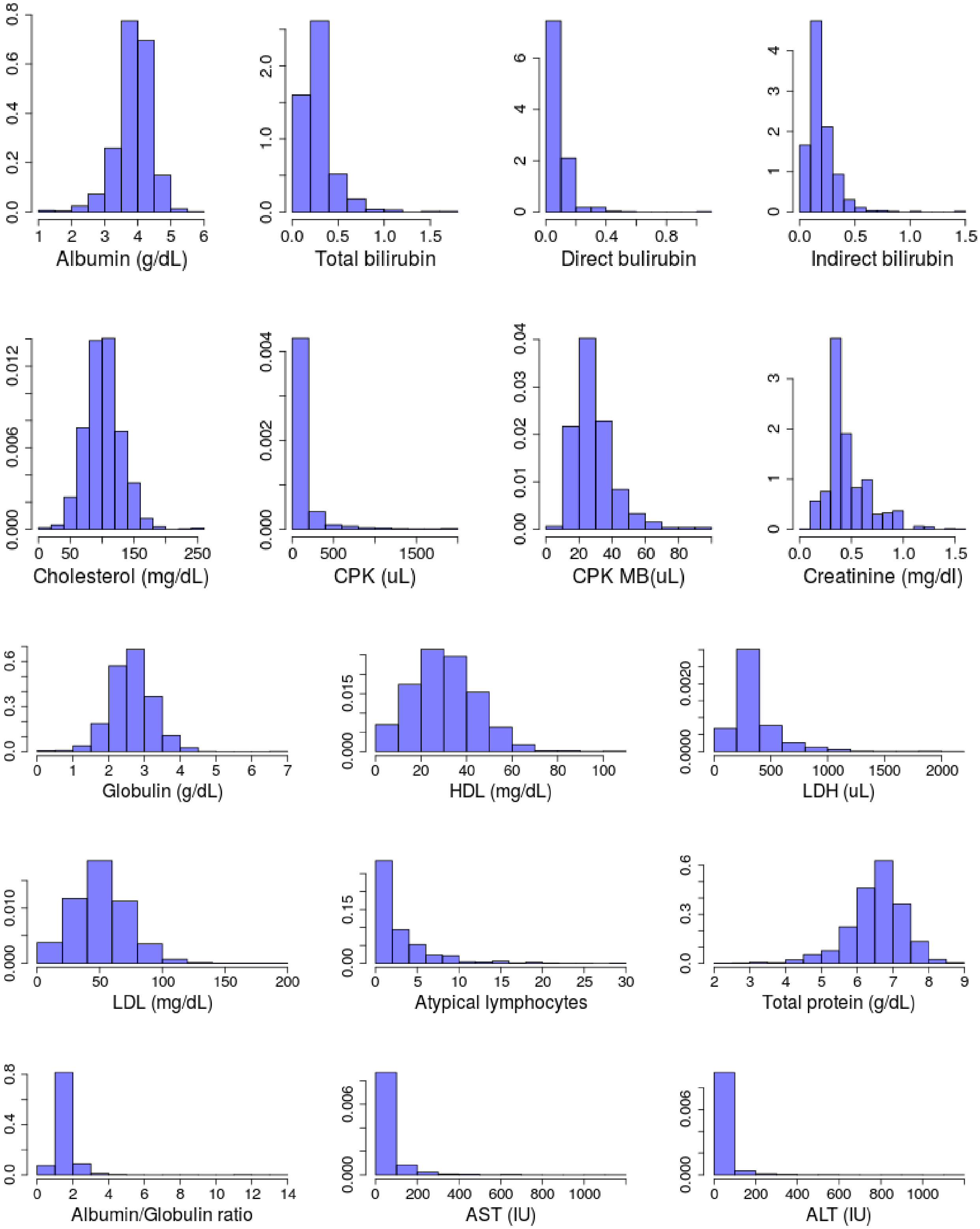

Figure 11: Blood chemistry - hospital patients

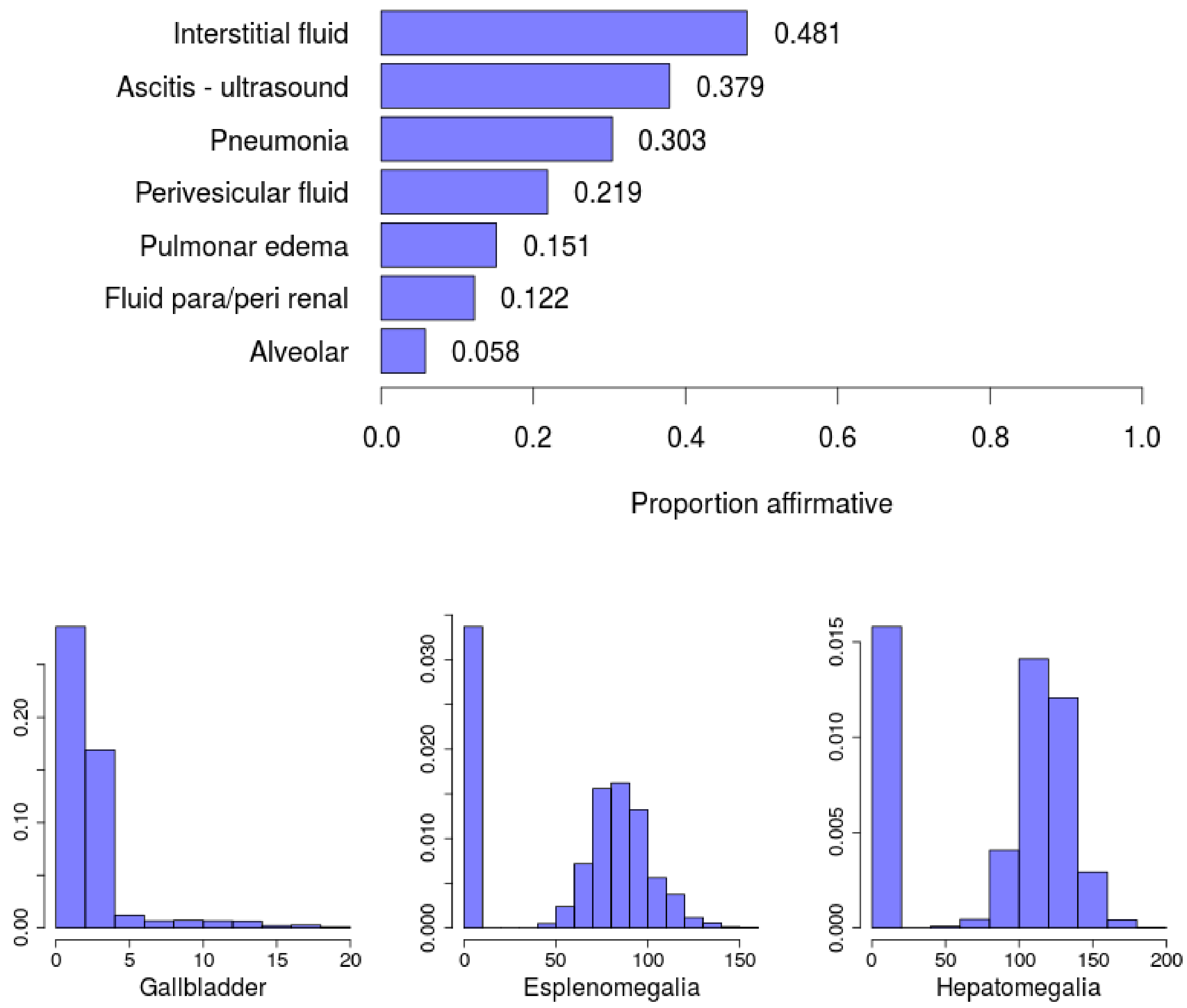

Figure 12: Ultrasound and X-ray indicators

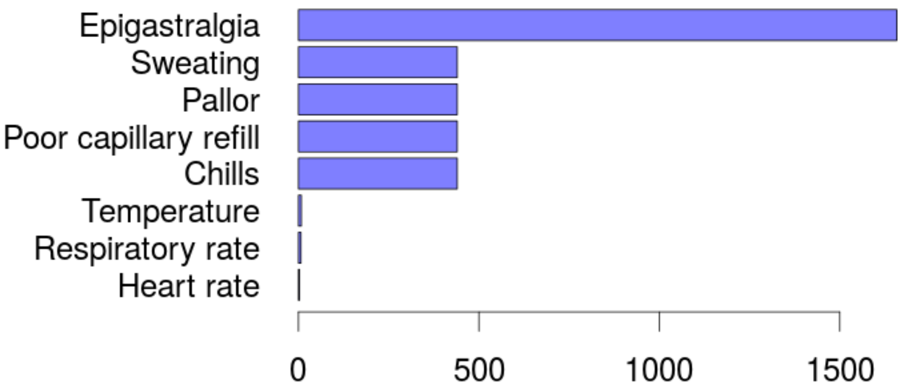

Figure 13: Counts of missing values among hospital patients - basic clinical variables

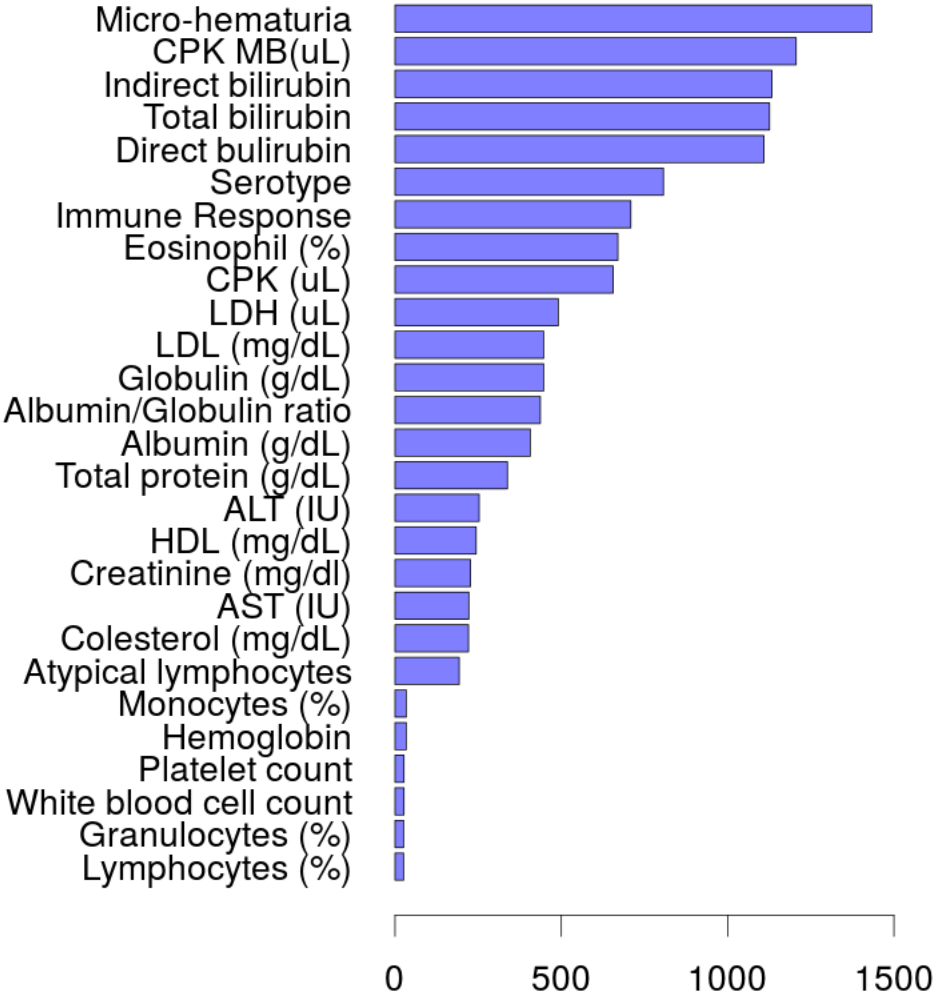

Figure 14: Counts of missing values among hospital patients - blood and urine lab variables

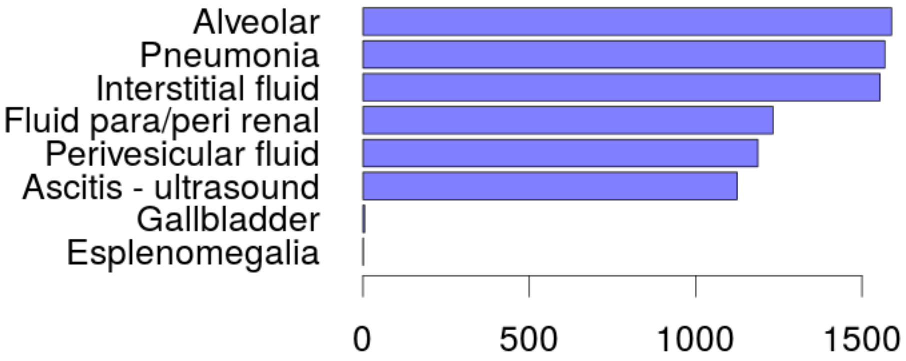

Figure 15: Counts of missing values among hospital patients - ultrasound and X-ray variables