Electrodiagnostic Nerve Tests: Understanding Healthy Peripheral Nerves

by

James Malcolm Bell

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Departments of Neuroscience and Computing Science

University of Alberta

Abstract

The Nerve Excitability Test (NET) is an electrodiagnostic test capable of non-invasive characterization of peripheral nerves in humans. It has utility in differentiating between healthy controls and subjects with peripheral nerve disorders. Full realization of the diagnostic potential of NET requires a substantial database of normative values. This thesis describes the process of combining NET data from multiple centers around the world (Canada, n=120, 57 male, ages 18-70; Japan n=85, 50 male, ages 19-86; Portugal n=42, 14 male, ages 22-84) to create the first international normative NET dataset for the human median nerve.

Since NET data often has missing values, we compared a number of approaches for filling missing values. An iterating cascading autoencoder performed the best. A simpler method, iterating linear regression, has similar performance while executing faster. The iterating cascading autoencoder was used to fill the missing values in the normative dataset.

Data collected from multiple locations can suffer from "batch effects": site-specific technical differences that reduce the homogeneity of the data. We developed a novel method for the detection of site-specific differences and found the homogeneity of the data from the three countries to be 95%, suggesting it is appropriate to combine the data into a single dataset. Comparison of the means of Canadian and Japanese NET data suggested that the remaining heterogeneity is due to technical differences in the stimulus-response curve, and that these differences have little or no impact on biological measures of

nerve health.

After establishing the normative dataset, we created a website which can be used as a clinical decision support system: NerveNorms.Bellstone.ca. The website also presents a nerve health score, interpretable as a p-value, to provide a quick and intuitive measure of the health of individual NET results relative to the normative dataset.

Acknowledgments

I would like to thank my supervisor, Kelvin Jones, for suggesting this project, mentoring me, providing additional funding, and making the process fun. His flexibility in tailoring the content and structure of my degree to my background and interests was invaluable. It was great to have the opportunity to return as an older student to cultivate my interest in neuroscience. Thank you for inviting me to be a part of this project.

I'm thankful to my second supervisor, Martha White, for teaching me so much about computing science. I couldn't possibly have done this project without her input.

I'm thankful for Clayton Dickson's willingness to be a member of my committee and that he ensured I had a solid understanding of the neuroscience relevant to my work.

Alona Fyshe provided valuable feedback as a member of my examining committee, but I especially appreciate that she encouraged me in my interest in the connections between machine learning and the brain (especially including an invitation to a CIFAR workshop about mind-machine links).

My daily status update calls with Matthias Stone were very helpful in keeping me motivated and focused in the times when it was hard to keep going. Also, his unnecessarily fast computer helped to speed up some long-running computations, and his suggestions for NerveNorms helped to speed up its development.

I'm grateful for the funding support provided by an Alexander Graham Bell Canada Graduate Scholarship (NSERC CGS-M), a QEII Graduate Scholarship, the University of Alberta (Walter H. Johns Graduate Fellowship), a Faculty of Medicine and Dentistry Graduate Student Recruitment Scholarship, and a Neuroscience and Mental Health Institute Stipend.

And of course I'm thankful to my family and friends for being supportive and making my life so good outside of school. You've had the biggest impact on shaping me into who I am today. You help me to remember to keep seeking excellence in everything I do, to find joy in my work, and—most importantly—to live a life that makes the world a better place.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Every day, millions of Canadians ask an artificial intelligence for help with everyday tasks like fetching the weather forecast or sending text messages. Every day, significantly fewer Canadians go to clinics to be tested for peripheral nerve disorders. The goal of this thesis is to begin the process of bringing those two things together, applying artificial intelligence techniques to improve peripheral nerve health. Eventually, artificial intelligence will aid in the automated diagnosis of peripheral nerve disorders.

In order to begin progress toward that goal, it was necessary to collect normative data in an adequate quantity to train machine learning algorithms, to fill missing values in that data, to test for the presence of site-specific differences in test results, and to develop a score for measuring the health of an individual's nerve. This thesis describes the progress made toward those goals. Further work will be necessary before differential diagnosis is possible, including collection of larger quantities of data from additional nerves, collection of data from participants with peripheral nerve disorders, and development of algorithms for differential diagnosis. Those aspects are not part of the present work; in this thesis, all data comes from healthy humans. After data collection via solicitation from international collaborators, the thesis project consisted of three phases: dealing with missing values in some nerve excitability measurements, determining whether it is statistically appropriate to combine international data, and development of the nerve health score.

Chapter 2 provides background information about the neuroscience and

computing science concepts used in this project. Later chapters describe the three phases of the project. The first phase, missing data, is discussed in Chapter 3. Phase two, site-specific differences, is the topic of Chapter 4. The third phase, creation of a clinical diagnostic support system, is the topic of Chapter 5. Chapter 5 considers the clinical implications of the normative dataset and introduces a website for browsing the data, which was designed as part of this project. The final chapter, Chapter 6, provides an overview of the work, its implications, and future directions.

Chapter 2

Background

This chapter provides background information about the neurophysiology of motor axons (Section 2.1) and the Nerve Excitability Test (Section 2.2) and basic concepts of machine learning (Section 2.3) and information theory (Section 2.4).

2.1 Motor Axon Physiology

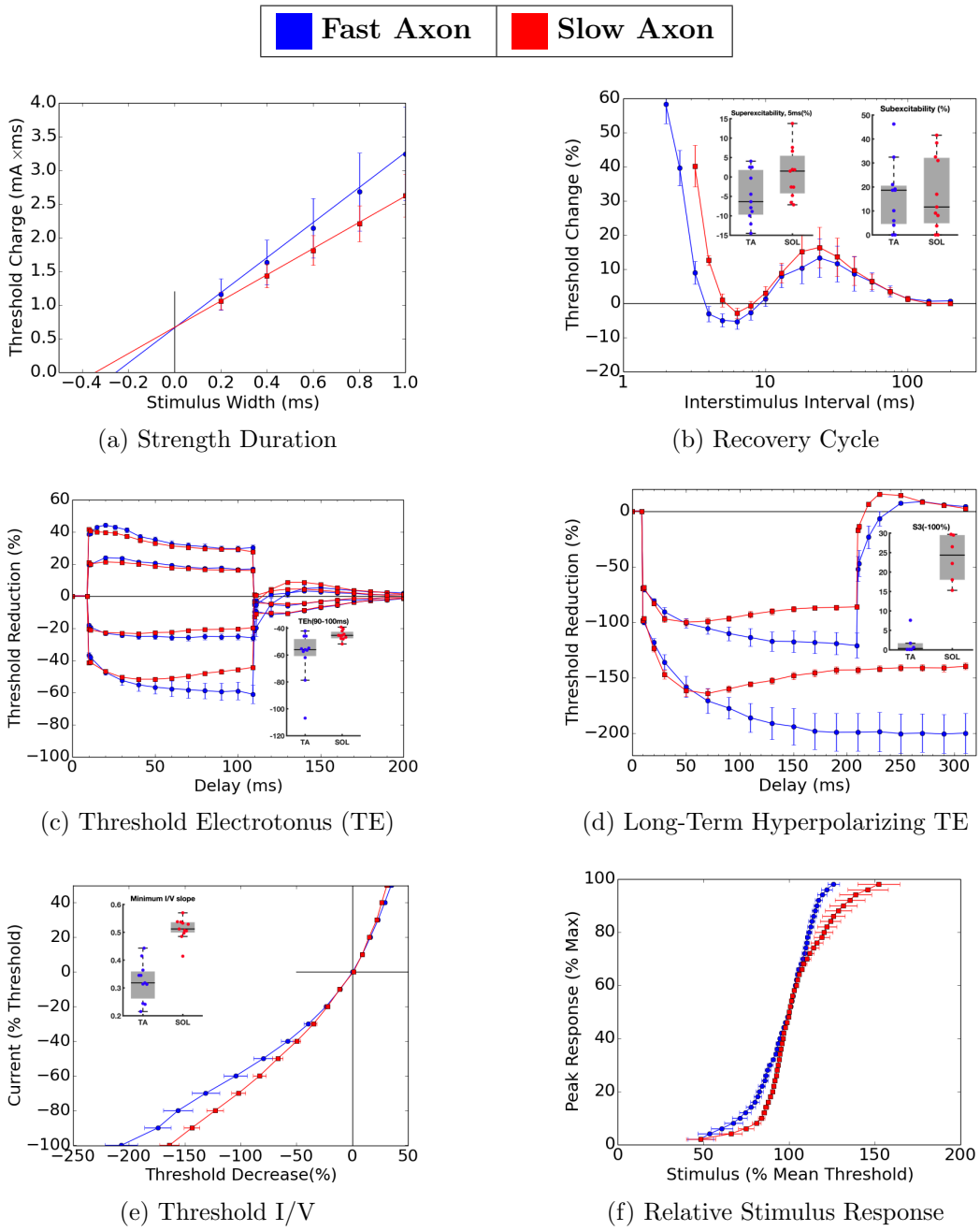
Axons propagate neuronal signals. These signals are in the form of action potentials: brief voltage spikes which propagate down the axon. Ion channels in the neuronal membrane open or close in response to the voltage across the membrane, selectively allowing specific ions, like Na^+ or K^+, to travel across the cell membrane. The mass transit of these ions changes the voltage, causing further changes in the ion channels. It is this movement of ions which causes voltage changes that propagate as action potentials. A neuron may fire repeatedly with some delay between action potentials; this firing encodes the signal being transmitted by the neuron. The firing of a neuron can be purposefully initiated by applying a current to the axon, either via needles or through the surface of the skin. When sufficient current charges the axonal membrane, an action potential propagates down the axon.

Motor axons propagate signals from motor neurons to muscle fibers. Specifically, action potentials in motor neurons effect the contraction of muscle fibers. Each motor neuron innervates hundreds of different muscle fibers, but each muscle fiber is innervated by a single motor neuron. When only a few mo-

tor neurons are excited, a weak contraction is produced. When many motor neurons are excited, a large contraction is produced. The fraction of the muscle activated by an applied stimulus to the axons can be measured using electromyography (EMG) and is called the compound muscle action potential (CMAP).

Understanding motor axons requires insight into their behavior in a variety of conditions. Traditional analysis of axonal behavior has been *in vitro* studies done on the nerve after dissection from the animal and placement in an artificial environment. These studies allow for unprecedented control of the solutions bathing the nerve, the introduction of drugs, and the use of small-scale electronic and optical devices for measurements. That is obviously problematic when studying living humans. Some methods allow for *in vivo* measurement, such as nerve conduction tests, which are conducted in clinics across Canada, but these noninvasive methods provide limited insight into the biophysics of nerves, *i.e.* the behavior of the underlying voltage-gated ion channels. The development of the Nerve Excitability Test (NET), described in the next section, has allowed for detailed, noninvasive study of *in vivo* axonal behavior.

One example of the utility of this test is our recent study of fast and slow axons in rats. Muscle fibers can be separated into two general categories: fast- and slow-twitch. Slow-twitch muscle fibers are involved in weaker, longer duration contractions, such as those used for postural control. They do not fatigue easily, but are unable to produce large force. Fast-twitch muscle fibers produce short, strong contractions for bursts of activity in movements like sprinting or jumping. Neurophysiologists had shown that the conduction velocity of motor axons to fast-twitch muscle fibers tends to be quicker than the velocity to slow-twitch, but this difference was explained by anatomical differences. There was little known about differences in the biophysical properties of fast and slow motor axons. We found that the axons leading to these muscles have functionally relevant differences in their biophysical properties. In particular, we found that hyperpolarization-activated, inwardly rectifying cation current (I_h) is stronger in slow axons, possibly due to their need to



(a) Strength Duration

(b) Recovery Cycle

(c) Threshold Electrotonus (TE)

(d) Long-Term Hyperpolarizing TE

(e) Threshold I/V

(f) Relative Stimulus Response

Figure 2.1: Differences between fast (tibialis anterior) and slow (soleus) motor axons in rat. The impact of I_h is especially notable in (d) Long-Term Hyperpolarizing Threshold Electrotonus. Mean values are plotted with standard error bars. Overlays show the distribution of individual rats for some notable excitability variables. This figure was reproduced from Bell et al. [4].

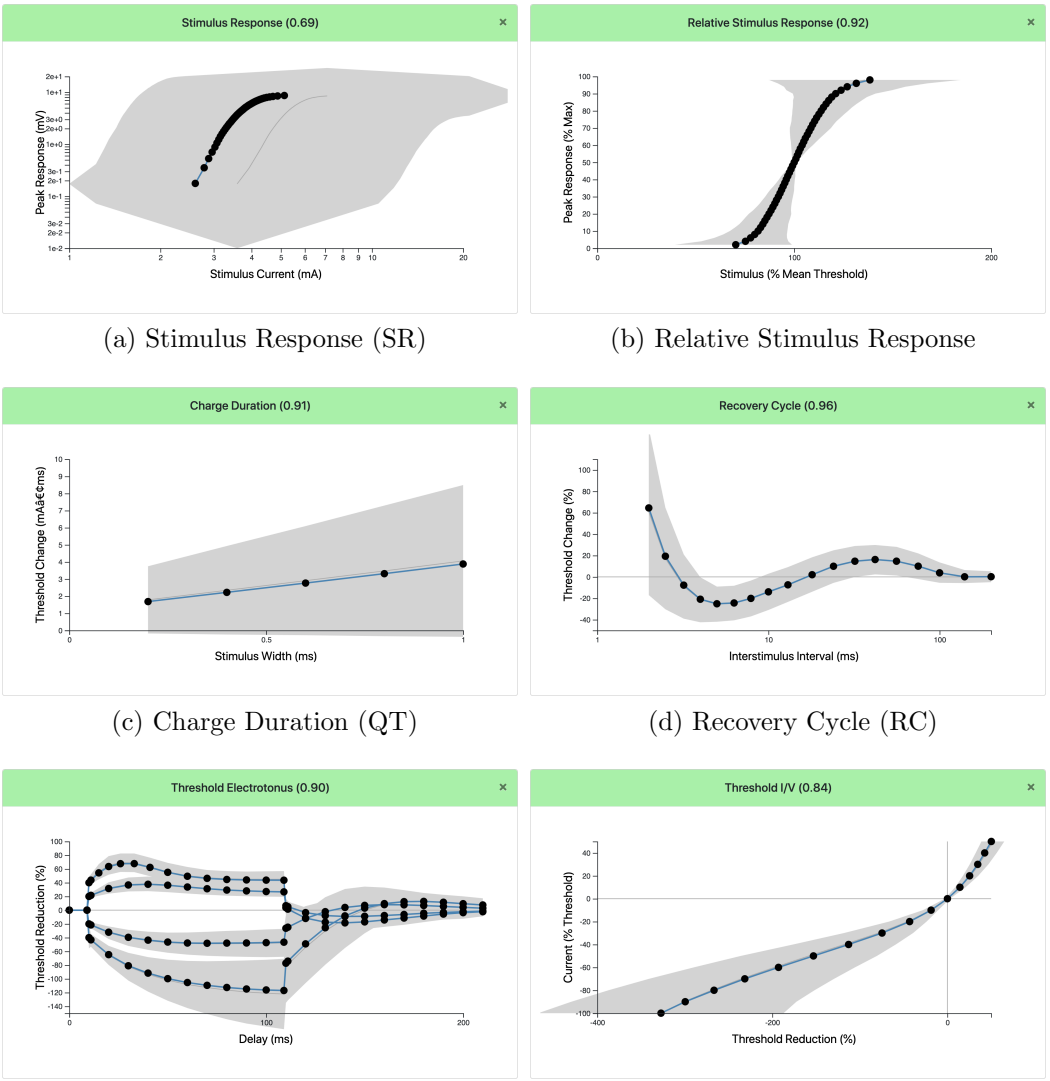
resist hyperpolarization over long periods of conducting impulses [4]. Figure 2.1 shows these differences between fast and slow axons, as measured by NET.

2.2 The Nerve Excitability Test (NET)

The Nerve Excitability Test was first established by Hugh Bostock and has been popularized with his QTRAC software. The protocol has gone through a number of revisions since the original TROND protocol, which was established in Trondheim, Norway in 1999 [31]. The physical setup of the test involves transcutaneous electrical stimulation using two electrodes placed over the nerve of interest. The normative data under consideration in this project was collected from the median nerve, so the stimulating cathode was placed over the wrist with the return anode on the forearm, intentionally placed away from the nerve. Two recording electrodes, which are placed over the muscle of interest, are used to collect electromyography (EMG) data. For the median nerve stimulation, this was the abductor pollicis brevis (APB, the largest thumb muscle). A fifth electrode is a ground. For the normative data, it was placed on the back of the hand. When the median nerve is stimulated, the motor axons innervating the APB muscle are activated causing a contraction. A larger stimulation activates more of the axons in the nerve, in turn recruiting more muscle fibers, which leads to a larger CMAP, which is observed as a large voltage at the recording electrodes. Before the test begins, the experimenter establishes the maximum CMAP by increasing the stimulating current until the CMAP stops increasing. With knowledge of the maximum CMAP, the QTRAC software begins an automated test consisting of five phases, or subtests. The data are then typically visualized in six characteristic plots.

Stimulus Response (SR)

Since most NET measurements consider the change in stimulus required to produce a target CMAP, it is necessary to first characterize the relationship between stimulus and CMAP. The amplitude of a 1-ms stimulating current is adjusted until it produces a CMAP with 2% of the magnitude of the maximum;



(a) Stimulus Response (SR)

(b) Relative Stimulus Response

(c) Charge Duration (QT)

(d) Recovery Cycle (RC)

(e) Threshold Electrotonus (TE)

(f) Threshold I/V (IV)

Figure 2.2: The six Nerve Excitability Test plots produced by QTRAC software. The data points are the mean of Canadian data. Shaded regions show 99% confidence interval for healthy controls (assuming a normal or log-normal distribution). Plots were generated by the NerveNorms website, which is discussed in more detail in Chapter 5.

this current and voltage are recorded. This process is repeated with amplitudes increasing to 100% of the current required to produce the maximum CMAP. This produces the *Stimulus Response* plot, as shown in Figure 2.2a. Unsurprisingly, this shows that a more activated nerve produces a stronger muscle contraction, but once the muscle is maximally stimulated, its response saturates.

Since SR has a lot of inter-individual variability, it is difficult to compare across individuals. To overcome this problem, the voltages are normalized to produce a *Relative Stimulus Response* plot. CMAP voltage is plotted relative to the maximum CMAP (i.e. 2% to 100%) while current is plotted relative to the current required to produce 50% of maximum CMAP. This plot, shaped like a tilted hourglass, is shown in Figure 2.2b. Since SR is less useful for comparison between individuals, it is often omitted (as it was in Figure 2.1).

The following tests are not as straightforward to measure as SR. Instead of merely sending a specific stimulus and measuring the response, these tests adjust a stimulus parameter (e.g. stimulus duration or amplitude) until a desired response is measured, and this stimulus parameter is the resulting dependent variable. In all cases, the target response is a CMAP equal to 40% of maximum. As can be observed in the relative SR plot (Figure 2.2b), the slope of the stimulus-response relationship is largest around 40%, so targeting 40% of maximum CMAP provides the most sensitivity to change. For example, if the stimulus results in a CMAP of 30% of maximum, the test will be repeated with a larger-amplitude stimulus current. If this results in 50% maximum CMAP, the stimulus will again be repeated, but with an amplitude between the previous attempts. This process is repeated until 40% maximum CMAP is achieved; the corresponding stimulating amplitude is recorded. This process of iteratively adjusting the stimulus amplitude in real-time to achieve a target response has been called **threshold tracking** in the NET literature. If the threshold of the motor axons increases, this is indicative of a decrease in their electrical excitability. It is the change in threshold from the control condition that is key dependent measure plotted in three of the NET graphs.

Charge-Duration (CD, or sometimes QT)

The charge amplitude and duration of a stimulus are inversely proportional; if the stimulus amplitude is decreased, a longer duration is required to produce the same CMAP. This relationship was first described by Weiss in 1901 as the "fundamental law of electrostimulation" [22]. This is measured by the *Charge-Duration* plot. The stimulus current required to produce 40% maximum CMAP is measured for five pulse widths ranging from 0.2ms to 1.0ms. These pulse widths are plotted against the threshold charge (i.e. current multiplied by pulse width). As shown in Figure 2.2c, this relationship is almost perfectly linear. The slope of this line is the rheobase in mA, which is the amount of current theoretically required for a pulse of infinite duration. The absolute value of the x-intercept of the line (not shown) is the strength-duration time constant (SDTC), which describes the behavior of the axonal cell membrane. Rheobase and SDTC are both excitability variables used to characterize the relative health of the nerve.

The remaining tests follow a condition-test paradigm, in which the nerve is conditioned in some way and then tested. The conditioning pulse puts the nerve into some state of interest, and the test pulse measures the excitability of the nerve in that state. As described above, the test stimulus is adjusted until a CMAP of 40% is observed, so multiple condition-test pairs are usually necessary. When measuring a single dependent variable value, the conditioning pulse is repeated with the same parameters, but the current of the test pulse is adjusted until it produces the required 40% maximum CMAP.

Recovery Cycle (RC)

For the *Recovery Cycle*, the conditioning pulse is a 1-ms stimulus with an amplitude greater than that needed for 100% maximum CMAP, *i.e.* a supramaximal stimulus amplitude. The 1-ms test stimulus is then sent at a variable delay, ranging from 2.5ms to 200ms after the conditioning pulse. This plot, shown in Figure 2.2d, demonstrates the response of the nerve to repeated stimuli. The y-axis shows the *relative* change in test stimulus amplitude re-

quired to produce a 40% maximum CMAP. The x-axis is time from 2.5ms to 200ms. Note that the data is not all observed in a single 200ms recording; each data point is measured as the result of a series of condition-test pairs which converge on the target CMAP. Each measurement takes time greater than the x-value of the point being observed because the conditioning and test pulses are separated by a time equal to the x-value, but an additional delay is required between each measurement to ensure the nerve has returned to its resting state.

This plot shows that most nerves are completely insensitive to repeated stimulation within the first couple of milliseconds; this is known as the absolute refractory period. During this time, the ion channels are still recovering from the first action potential; most notably, transient sodium channels are still inactivated and must be de-inactivated. There is then a period of increased threshold, the relative refractory period, which lasts for a few more milliseconds. In unmyelinated nerves (not shown), this relative refractory period is the only notable feature of the plot, which resembles a decaying exponential. However, in myelinated nerves, the complex interactions of the myelin capacitance and changing ion channel conductances gives rise to superexcitable and subexcitable periods. During the superexcitable period, the nerve is *more* responsive to stimulus; a smaller-amplitude stimulus results in an equal-magnitude response. This is followed by a late subexcitable period, during which, like the relative refractory period, the nerve is not excited as easily. Since these phenomena occur at exponentially increasing delays, the x-axis is plotted logarithmically.

In most tests, the electrical interactions between the stimulating and recording electrodes are unimportant. Though the current from the stimulating electrodes induces a large voltage at the recording electrodes, it is only observable in the first few milliseconds after the stimulus, and tests other than RC occur much later. However, in RC, this stimulus artifact can obscure recordings in the first 2–3ms after the stimulus, resulting in missing data at the beginning of some RC plots.

Threshold Electrotonus (TE)

In hyperpolarizing *Threshold Electrotonus* (see the lower half of Figure 2.2e), the conditioning pulse is a 100ms long hyperpolarizing stimulus. The amplitude of this hyperpolarizing pulse is equal to 40% of the amplitude required for a 1-ms stimulus to evoke the target CMAP, *i.e.*40% maximum CMAP. Like RC, the y-axis shows the relative change in the required stimulus amplitude and the x-axis is time. However, TE shows threshold *reduction*, so increases and decreases in excitability are in the opposite direction of RC, and TE time is linear, not logarithmic. The x-axis of TE indicates the delay between the onset of the conditioning pulse and the onset of the test pulse. The conditioning pulse always begins at the 10ms mark, so there is no change in threshold prior to 10ms. The presence of the hyperpolarizing pulse immediately causes a decrease in excitability; since the first measurement occurs as the 40% hyperpolarizing current is beginning, the increased current required to activate the nerve is 40%. A data point at an x-value of 20ms corresponds to the change in amplitude required to evoke 40% maximum CMAP after the nerve has been conditioned by a 10-ms, 40% hyperpolarizing current. As the duration of the hyperpolarizing current increases up to a duration of 100ms (at 110ms on the x-axis), the excitability of the nerve slowly returns to baseline. Data points after 110ms show the response of the nerve with the hyperpolarizing current turned off. For example, at 130ms, the nerve has been subject to a 100-ms conditioning pulse (from 10ms until 110ms), followed by 20ms of recovery. In this region, the nerve may exhibit a rebound increase in excitability which slowly decays toward zero.

The data collection process is repeated with a 40% depolarizing conditioning pulse (see the upper half of Figure 2.2e). It is usually also repeated with 20% hyperpolarizing and depolarizing pulse. These measurements provide an understanding of the dynamics of the nerve after hyperpolarization or depolarization, i.e., after long-term activity has changed the nerve's threshold. Most importantly, these tests reveal biophysical properties of ion channels that are located beneath the myelin sheath of the motor axons, *i.e.* the internode

region.

An optional extension to the test protocol uses 70% and 100% hyperpolarizing currents which last for up to 200ms and 300ms instead of 100ms (not shown in Figure 2.2, but see Figure 2.1d) [34]. This long-term TE is especially effective for observing I_h: the hyperpolarization-activated current. Since this current is activated by long-term hyperpolarization, it only begins to be evident near the end of the 100ms conditioning current, so long-term TE is more effective at showing its effect. It is an inwardly-rectifying current which counteracts the hyperpolarization, so the long-term TE in a nerve with large I_h will be attenuated.

Current-Voltage (IV)

The *Current-Voltage* relationship is similar to TE, but all of the measurements are made at the end of the hyperpolarizing/depolarizing conditioning pulse, and the conditioning pulse width is 200ms instead of 100ms. The independent variable is the strength of the conditioning pulse; instead of being fixed at 40% of the stimulus current required to produce 40% maximum CMAP, it ranges from -100% to +50% in increments of 10%. Contrary to standard plotting convention, the dependent variable is plotted on the x-axis, as shown in Figure 2.2f.

Excitability Variables

The excitability variables, also referred to as excitability indices, are generated by QTRAC to provide a summary of the results of the plots. These variables are measures taken from the plots, such as a particular y-value, an x-intercept, or the slope or maximum in a specified region. The participant's age, sex, and temperature are also included in the QTRAC list of derived excitability variables, though it is not usually appropriate to treat them in the same way. There are approximately 30 standard excitability variables, but the exact count depends upon which tests are included. Quantitative analysis in NET studies usually focuses on these variables. In keeping with that practice, this study

primarily considers those derived values extracted from the continuous data plots. Potential weaknesses of this approach will be discussed in Chapter 6.

For a more detailed overview of the biophysical underpinnings of this test and its application in clinical settings, refer to Kiernan and Lin [30].

2.3 Machine Learning (ML)

This introduction to machine learning forgoes detail in favor of presenting the minimum conceptual foundation that will be necessary to understand the work in later chapters. This is in contrast with those later chapters, which will provide more detailed descriptions of the methods used. The aim of this section is to provide a reader without an ML background with the necessary foundations to grasp the relevant concepts. For a more detailed introduction to ML, consider Goodfellow et al. [18].

The most basic ML method is linear regression, which, like most of the field of ML, is firmly rooted in statistics. Consider a basic x-y scatter plot: if the goal is to predict Y based on X, a line of best fit can be drawn and used to make predictions. This line is described by the standard equation

$$Y = mX + b, \tag{2.1}$$

where m is the slope of the line and b is its y-intercept. If we ignore b, we can roughly calculate the slope of the line as $m = \mathbf{Y}/\mathbf{X}$, where $\mathbf{Y}$ and $\mathbf{X}$ are vectors of all of the dependent and independent variables. (In fact, the correct equation is $m = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, where m could be a matrix if there are multiple dependent or independent variables, but the concept is similar to a simple division.) This method of calculating a line of best fit (likely in multiple dimensions) in order to predict a variable of interest is the most basic machine learning method: linear regression. Various techniques add complexity: data can be transformed before being input to the algorithm, making logarithmic features linear; regularization, such as forcing the multi-dimensional slopes to have similar magnitudes, can aid in generalization to new situations; multiple layers of learning can be used, resulting in "deep learning";

the predicted variable could be a single category, like "good" vs "bad", or a list of non-exclusive labels, like "delicious", "nutritious", "nut-free", and "vegan". However, the basic goal of supervised machine learning remains the same: we want to make predictions based on observed data.

Unsupervised learning is a variation of supervised machine learning: instead of providing known values with the goal of prediction, the entire dataset is thrown at an algorithm which attempts to create order out of chaos. The algorithm does its best to sort the data into groups (or "clusters") based on features it considers to be prominent. For example, consider a dataset of pictures of the digits 0 through 9, with a single digit in each picture. An unsupervised algorithm might sort the data into 10 clusters perfectly corresponding to the ten digits. Or it might instead sort into three clusters: handwritten in pen, handwritten in pencil, and typed. A poor algorithm might instead sort the data into clusters that make no sense to a human interpreter. Unsupervised learning is very effective at clustering data based on features that are prominent in the view of the algorithm, but it is not always clear how the algorithm arrived at its conclusions.

A few specific ML techniques are especially relevant to this project. For missing data, linear regression, neural networks (including autoencoders), and principle component analysis are important, so they are described below. To detect site-specific differences, unsupervised clustering is important, so k-means clustering and hierarchical clusters are described below.

Neural Networks

Neural networks are used for deep supervised learning (among other things) [18]. A neural network is similar to multi-layered linear regression: $\mathbf{X}$ is used to predict $\mathbf{H}$, which is used to predict $\mathbf{Y}$. When building a neural network, we often do not care about the intermediate predictions ($\mathbf{H}$). While multiple layers like this do not improve predictions for a simple linear relationship, when the data is transformed by a non-linear function between layers, neural networks become very powerful prediction engines. (For example, we use the hyperbolic tangent of $\mathbf{H}$ as the input to the final layer.) Known samples of

X and **Y** are provided to the neural network and the weights in the network (which are conceptually similar to the slope of a line of best fit) are modified until the network accurately predicts the known **Y** based on **X**. It can then be used to predict **Y** when **X** is unknown.

Autoencoders

An autoencoder is a type of neural network which aims to reproduce its input [18]. (Chapter 3 includes a block diagram of an autoencoder in Figure 3.2.) Given some values **X**, it attempts to build a network which can reproduce **X** as closely as possible, even after each layer transforms the data in some fashion. One strength of an autoencoder lies in the intermediate layers which are otherwise often ignored (e.g. **H**, as described above). If the input **X** for each sample includes a lot of data (e.g. a long list of all of the movies you have ever watched and the rating you gave each one), the *hidden layer* **H** could be much smaller (e.g. the hidden layer might learn to store a short list of genres and your rating for each, which can be used to approximately predict your movie ratings). This can be useful for data compression (storing **H** instead of **X**), dealing with noise (which might be filtered out by the autoencoder), or predicting missing values (as in this project).

Principle Component Analysis

Principle Component Analysis (PCA) can be viewed as an unsupervised learning method which transforms some input data into a different representation without losing any information, though it is usually used in a way that results in some information loss [26]. One example of transforming data without losing information is to record height and BMI instead of height and weight; either pair of data can be calculated from the other. Another example is transforming from Cartesian (i.e. x-y-z) coordinates to a cylindrical or spherical coordinate system. PCA transforms data into a representation which often has no real-world meaning, but it is useful because the new representation maximizes variance in each dimension; that is, it transforms the input variables into new variables which are ranked in order of how much they vary

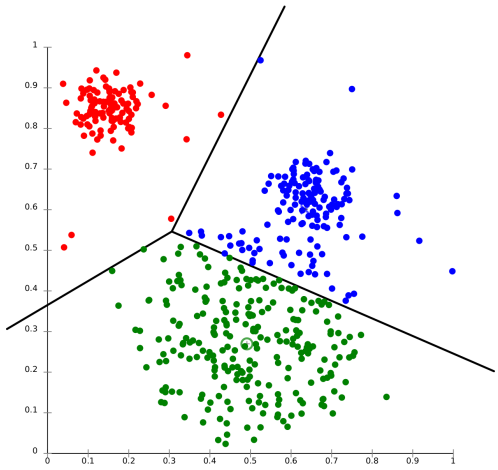


Figure 2.3: The clusters found by k-means are centered around the mean of the points assigned to the cluster. As demonstrated in this figure, this may not be ideal for clusters with different spacial distributions. There are obviously three clusters in this data, but k-means incorrectly labels some points as blue even though they belong to the bottom cluster. This figure by Chire (https://commons.wikimedia.org/wiki/File:KMeans-Gaussian-data.svg) is licensed under CC BY-SA 3.0.

between samples. The low-variance variables can often be discarded without losing much information. This means PCA is capable of data compression and predicting missing values.

k-Means Clustering

Unsupervised learning aims to organize data without using any prior knowledge about the data. In k-means clustering, the samples are sorted into k clusters. As shown in Figure 2.3, each cluster is centered around the mean of the samples that belong to that cluster; each sample is assigned to the cluster with the mean closest to the sample. The means and assignments of samples to clusters are updated iteratively until convergence. The final clusters have some dependence on the initial random choices for the cluster means, so k-means clustering may produce different results for the same inputs. Due to this non-deterministic behavior, the methods for detecting site-specific differences in Chapter 4 use a hierarchical cluster tree instead of k-means, but the

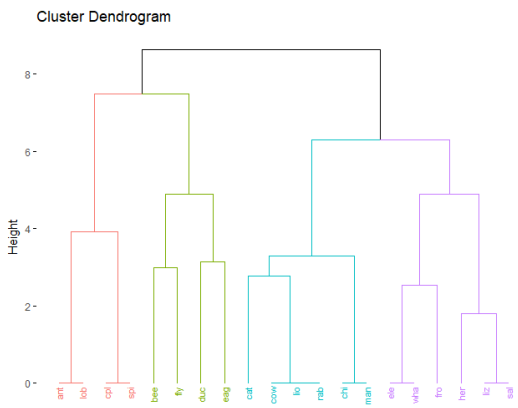


Figure 2.4: A dendrogram, or tree diagram, can be used to cluster samples. In this dendrogram, species are organized by their relationships to one another. When two species are closely related, they are connected by a shorter distance. In this example, different clusters of related species are colored red, green, blue, and purple based on the branches they descend from. Nerve health data can similarly be clustered in a tree diagram according to the similarity between samples. This figure by Jakub al13 (https://commons.wikimedia.org/wiki/File:Dendrogram4.png) is licensed under CC BY-SA 4.0.

simplicity and ubiquity of k-means makes it a reasonable alternative choice.

Agglomerative Hierarchical Cluster Tree

A common way of organizing data is in a dendrogram: a tree diagram (see Figure 2.4). For example, animal species are sorted in a large tree that splits according to kingdom, phylum, class, order, family, genus, and species. Clusters of species can be formed by splitting at any of those levels. Hierarchical tree clustering has two steps. First, the data must be organized into a tree based on a linkage. The linkage, which is a measure of the distance between pairs of samples, will determine the structure of the tree. It is used to assign each sample to a location in the tree. Second, the tree must be split into the desired number of clusters by traveling down the dendrogram to find appropriate places to split the tree. Splitting data into a dendrogram in order to find clusters will be important for detecting site-specific differences in Chapter 4.

2.4 Information Theory

Information theory is a field concerned with understanding how much information is contained in a given event or dataset [53]. For example, it takes one bit of information to communicate the result of each flip of a fair coin: it is either heads or tails. It takes zero bits of information to communicate the result of a flip of an unfair coin that always lands on heads: the result can be assumed without the need to communicate anything. A dataset of 100 independent fair coin flips contains 100 bits of information. It is possible to calculate the amount of information, or entropy, in any dataset with the equation

$$H(\mathbf{X}) = -\sum_i p(x_i) log_2[p(x_i)], \tag{2.2}$$

where $H(\mathbf{X})$, the entropy of the dataset $\mathbf{X}$, is a sum across all elements x_i of $\mathbf{X}$. $p(x_i)$ is the probability of observing event x_i. A fair coin has two states, each with probability 0.5, resulting in the expected entropy of 1 bit.

The joint entropy of two events measures how much information is contained in the two events together:

$$H(\mathbf{X}, \mathbf{Y}) = -\sum_i \sum_j p(x_i, y_j) log_2[p(x_i, y_j)]. \tag{2.3}$$

For two datasets, it is a measure of the total information they contain. Conditional entropy is the additional amount of information in $\mathbf{Y}$ given knowledge of $\mathbf{X}$:

$$I(\mathbf{Y}|\mathbf{X}) = -\sum_i \sum_j p(x_i, y_j) log_2 \frac{p(x_i, y_j)}{p(x_i)}. \tag{2.4}$$

Mutual information is a measure of the information shared between two datasets:

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{X}, \mathbf{Y}) - H(\mathbf{X}|\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}). \tag{2.5}$$

The mutual information I of $\mathbf{X}$ and $\mathbf{Y}$ is equal to the entropy of both datasets minus the entropy of each dataset given the other dataset. Two independent datasets have no mutual information; the combined entropy is equal to the sum of the individual entropies. Two datasets with some shared information (e.g. a dataset of heights compared to a dataset of corresponding weights) have large

mutual information. When the two datasets contain identical information content (e.g. a dataset of height and BMI compared to a dataset of height and weight), the mutual information is maximized and is equal to the entropy of either dataset.

Mutual information is an important foundation for considering whether multi-site data can be considered to come from the same distribution. If the multi-site data is shown to share a common distribution, it is appropriate to combine the data into a single dataset. This topic will be considered in more detail in Chapter 4.

Chapter 3

Missing Data

This chapter describes the first phase of the project: filling missing values in the normative data.

3.1 Introduction

The issue of missing data, and lack of a systematic approach, is pervasive in many scientific fields. A common *ad hoc* method for missing data is listwise deletion, also known as complete case analysis, which involves deleting any cases (or samples) in which data is missing. A more modern approach is to infer missing values with statistical techniques, such as multiple imputation. Reviews of missing data practices in educational psychology in the early 2000s [45, 46] and then a decade later [12, 14] showed improvements in handling missing data, with a trend moving away from more *ad hoc* methods like listwise deletion to those using statistical techniques. The most recent survey, however, still found that only 63% of papers handled missing data at all, and almost 30% still used listwise deletion [14]. Similar issues in how missing data is treated have been discussed more generally in psychology [7], for behavioral neuroscience [48], and for medical research with clinical trials [15, 35]. Repeatedly, it has been found that simplistic approaches underperform compared to more modern missing data techniques [12, 14, 20, 47, 48]. In spite of major advances in the handling of missing data [20, 36], uptake is not pervasive.

The standard method of analyzing NET results with missing values is adaptive n: for each feature, researchers use the mean value of the n samples

which are not missing for that feature, potentially resulting in a different n for each feature. This method of analysis is the default behavior of QTRAC, the primary software package for conducting these tests. While this method of missing data analysis has been adequate for studies thus far, more complex analytical algorithms often rely on matrices of data without missing values. For example, calculations of variation of information (described below) perform very poorly with missing values. Future diagnostic algorithms are also likely to require these values to be present. Therefore, a more systematic approach to missing data is required.

In some cases, it is important to fill the missing values in the original data matrices, but often only means and the covariances of the features are important. For example, confirmatory factor analysis only relies on the covariances, so it is not necessary to fill the missing values [9], though effective missing data techniques are still important to ensure the resulting covariance matrix is accurate. On the other hand, if a researcher's goal is to compare a patient's data to normative controls, it may be important to be able to fill missing values in that individual's test results. Therefore, this study considers both problems: we optimize the reconstruction of missing values in a matrix of test data and we optimize the reconstruction of the covariances of the features in the same test data. This is more likely to preserve the validity of both individual features and the relationships between them.

A common, well-performing, theoretically justified approach to imputation of missing values is multiple imputation (MI). Any process which imputes a single value into a missing field is a single imputation method. While single imputation is simple, it under-reports the variance because the best imputation for any value is the expectation of that value, which removes noise. If single imputation instead uses a model of variance to impute values with a stochastic method, the variance is more likely to be correct at the cost of increased error in the imputed values. MI instead uses a stochastic process to generate the missing values, but repeats the process to impute multiple different copies of the dataset (often 10), each with the missing values imputed by the same stochastic process. The average of these imputed values provides a

Table 3.1: Recent studies with nerve excitability results measured from the median nerve in healthy human participants.

Country	n	n male	Ages	Source
Canada	150	73	18–70	previously unpublished
Japan	84	49	19–86	Bae et al. [1]
Portugal	42	14	22–84	Casanova et al. [11]
Australia	60	28	22–79	Jankelowitz et al. [24]
Ireland	105	54	19–82	McHugh et al. [38]

good measure of the expected value, while the variance between the datasets can be combined with the variance within the averaged dataset to produce a good measure of the expected variance. While MI performs well in practice, it is best used as part of a complete analytical process. Rather than using a MI method to create a dataset with good imputed values and then analyzing that dataset as a second step, the best multiple imputation involves generating multiple datasets, analyzing each one separately, and then combining the results of the analysis using MI techniques so the final results consider the effect of missing data.

There are some scenarios in which MI is not appropriate. First, if data is being imputed into a single sample (e.g. to fill data in a single NET result for comparison to a normative dataset), it does not make sense to calculate the single-sample variance, so a MI approach is not helpful. Second, MI is part of a complete analysis process. If the analysis steps are unknown, it is not possible to carry out MI. Since the NET data is being prepared as a normative dataset for use by other researchers, it is most appropriate to do single imputation. If the downstream analysis is going to take advantage of MI, the normative dataset can be used without pre-filled missing values, but for those researchers who prefer a complete dataset, a version with values filled by single imputation will also be provided.

3.2 Methods

The nerve excitability data came from three geographic locations as described in Table 3.1. All four locations with previously published data were invited to participate, but Australia and Ireland declined. Tests used the standard protocol for NET results: QTRAC software with a TROND protocol [31]. However, the Japanese dataset did not contain threshold electrotonus 20 % (TE20) data, so all TE20 data was removed from the other datasets. Two samples from Japan and one from Portugal were missing participant sex. Since future analysis is likely to segregate data based on sex, it would be inappropriate to impute it, so those values were removed from analysis and are not included in the total count of samples in Table 3.1. One of the Japanese samples did not include any TE and I/V slope data, so it was also removed.

While a NET dataset consists of numerous continuous waveforms, as described previously, this analysis focuses on the standard excitability indices which QTRAC automatically calculates from the plots. (These indices may include the y-value at a specified x-value, the maximum in a specific region, or the slope of a segment.) This reduces the amount of data to analyze while focusing on values known to be of interest. After removal of TE20 data, 31 excitability indices remained, consisting of 30 continuous measures (age, temperature, and 28 other variables) and one categorical measure (male/female).

Due to biological variations or technical difficulties, NET data is sometimes missing. In particular, nine of the 31 features are missing at rates listed in Table 3.2.

Of the 276 samples, 37 of those samples were missing data. Their true values are unknown, but must be filled to allow detection of site-specific differences (and potentially other analysis) on the entire dataset. In order to determine which missing-data algorithm is most effective at filling values in this dataset, the 244 samples without missing data were used to create a complete matrix of NET samples. For analysis of the missing data, all of the normative data was combined into a single dataset to increase the number of samples.

Table 3.2: Percent of observations that are missing for the features that have missing data in Canada (CA), Japan (JP), and Portugal (PT). Rates are reported as *count (percent of dataset)*.

Feature Name	**Count Missing (Percent)**			
	CA	**JP**	**PT**	**All**
Refractoriness at 2 ms (%)	14 (9.3%)	7 (8.3%)	4 (9.5%)	25 (9.1%)
Refractoriness at 2.5ms (%)	6 (4.0%)	1 (1.2%)	1 (2.4%)	8 (2.9%)
TEh(overshoot)	2 (1.3%)	1 (1.2%)	—	3 (1.1%)
Hyperpol. I/V slope	3 (2.0%)	—	—	3 (1.1%)
Resting I/V slope	1 (0.7%)	—	—	1 (0.4%)
Minimum I/V slope	1 (0.7%)	—	—	1 (0.4%)
RRP (ms)	1 (0.7%)	—	1 (2.4%)	2 (0.7%)
TEh(slope 101-140ms)	1 (0.7%)	—	—	1 (0.4%)
Temperature (C)	—	1 (1.2%)	—	1 (0.4%)

To determine which missing data algorithm performed the best, the dataset was resampled 100 different times, using the following methodology. First, a test set with missing data was created by randomly deleting from features in proportions according to Table 3.2 [1]. This resulted in a test set containing approximately 199 samples with complete data and 45 samples with missing data. (The exact numbers of missing and complete samples varied each time the test set was generated, since the distribution was random.) Each algorithm was then executed on the test set to produce a matrix of filled values. The mean squared error (MSE) for the data was calculated for each algorithm by comparing its output to the reference data. Since the missing features are of unequal magnitude and variance, the MSE for each feature was normalized by its variance. This process was repeated to measure each algorithm's performance on 100 different subsets of deleted data. Paired t-tests across the 100 resamples were used to determine significance for each pair of algorithms. This process is diagrammed in Figure 3.1.

When testing for site-specific differences in Chapter 4.1, it will also be necessary to impute missing values in small datasets with around 40 samples. While it is normally appropriate to impute missing data across the combined

[1] To match the observed missing data, "Refractoriness at 2.5 ms (%)" was only deleted if "Refractoriness at 2 ms (%)" was also deleted.

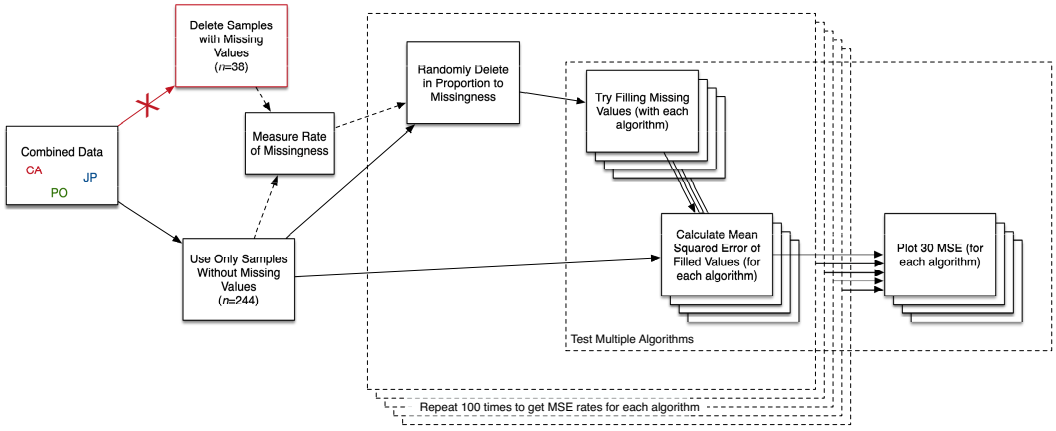


Figure 3.1: The missing data algorithms were tested by randomly resampling complete rows and deleting features in proportion to the observed missing data rates. The mean squared error (MSE) of each algorithm was measured to compare performance.

dataset, it will be necessary to impute smaller datasets (e.g. PO, $n = 42$) when testing for site-specific differences. To consider imputation performance in those datasets, the above process was repeated with smaller sample sizes. Instead of using all 244 samples, a random 40-sample subset was used. Other steps were followed as described above, but note that only temperature was not in this smaller sample because it occurred at rates too infrequent to be deleted.

All code was implemented in Matlab R2018a (9.4.0.813654) and is available online [2]. Actual p-values are reported unless lower than 0.0001.

Methods To Reconstruct Missing Data

The following methods were used to reconstruct missing data. In the following descriptions, m is the number of missing features (up to nine, but sometimes as low as eight since the features with missing rates below 1% are often not missing in smaller samples). In all cases, hyperparameters[2] (if any) were optimized on the test data, so the demonstrated performance represents an upper bound on performance (i.e. it was not validated using separate training and test data). For all algorithms except mean imputation and regression, val-

[2]The number of iterations in the iterating algorithms was one important hyperparameter.

ues were converted to zero mean and unit variance before being input to the algorithm. Block diagrams for these algorithms are in Figure 3.2.

Mean imputation (Mean) In mean substitution, missing values were filled with the corresponding feature's mean value.

Data augmentation (DA) This is a multiple imputation method with code from Folch-Fortuny et al. [16], but it is used here in a single-imputation context. It estimates the parameters describing a feature's distribution and uses them to impute the missing values. This process is iterated for some number of steps or until convergence. Alternating between imputation of missing values and estimation of parameters like this forms a Markov chain. In practice, a chain with length of two was found to converge for the NET data. DA is not recommended for datasets when the number of features approaches the number of samples [16].

Principal components analysis with alternating least squares (PCA) Matrix completion decomposes a data matrix into two lower-dimensional matrices. When the inner dimension of the two lower-dimensional matrices is reduced, the product of those matrices results in a filled approximation of the original data. The specific method of matrix decomposition used in this case was principal component analysis (PCA), since it is a common method for factorizing or decomposing a matrix. PCA on its own cannot handle missing values, but when the PCA solution is found using alternating least squares (ALS), it can fill missing data. Since the resulting matrix is an approximation of the original matrix, all values (not just missing values) will be modified, but only the predicted values for the missing data were used; the final matrix used the original values when available and filled missing values using the output of the ALS PCA solution.

Iterating PCA (iPCA) First, PCA was performed as described above to update the missing values. The matrix with filled values was used for another

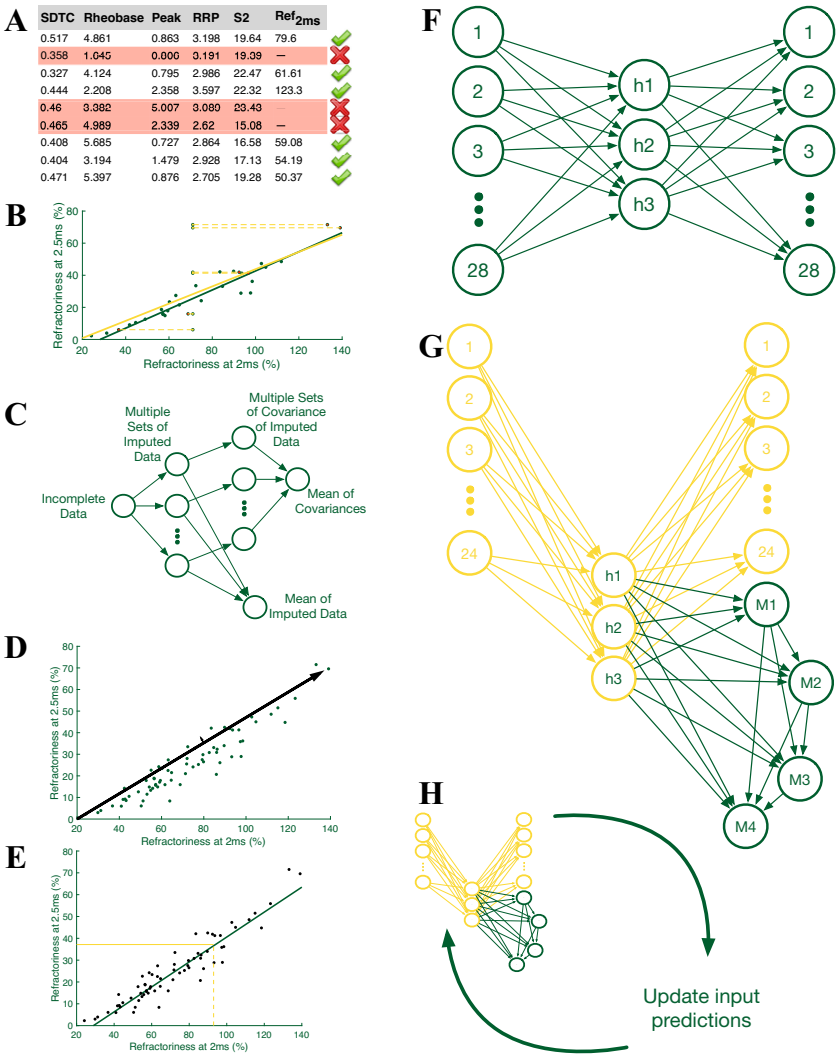


Figure 3.2: Block diagrams for algorithms to fill missing data. (**A**) Complete case analysis deletes any row which contains a missing value. (**B**) Mean imputation uses the feature mean. In this example, Refractoriness at 2ms is sometimes missing (red dots), so the mean is imputed. Sometimes this gives points which are clearly outside the distribution of observed points. (**C**) DA is a multiple imputation algorithm, which allows for better statistical performance on downstream analysis (e.g. calculating a covariance matrix), but the present work only uses it for single imputation (i.e. the mean of the imputed data). (**D**) Principle Components Analysis identifies orthogonal directions of the largest variance. In this example, values would be imputed as close as possible to the large arrow, which is the direction of the largest variance. (**E**) Linear regression imputes values based on regression across other variables. In this example, the strongly correlated features would result in accurate imputation. (**F**) The autoencoder takes the 28 features as inputs and does its best to recreate those same 28 features. (**G**) The cascading autoencoder takes the known features as its input to predict the unknown features. (**H**) The iterating algorithms (cascading autoencoder shown here) perform imputation repeatedly to improve their accuracy.

round of PCA, which provided updated predictions. This process was repeated 20 times.

Linear regression predictor (Regr) First, missing values were filled with the feature mean. Then a linear regression model was trained, and that model was used to predict the missing values. The model was not regularized.

Iterating linear regression predictor (iRegr) First, a linear regression predictor was used as described above to update the missing values. The matrix with filled values was used to create a new linear regression model, which provided updated predictions. This process was repeated 20 times.

Autoencoder (AE) An autoencoder is a neural network in which the inputs are the same as the outputs. The missing inputs were replaced with their feature mean. The autoencoder was optimized with ADADELTA [66].

Iterating Autoencoder (iAE) First an autoencoder was used as described above to update the missing values. The matrix with filled values was used to train a new autoencoder, which provided updated predictions. This process was repeated 5 times.

Cascading Autoencoder (Casc) This started with a true autoencoder with the $31 - m$ inputs and outputs: the complete features. The missing features were sorted from least missing values to most. The first missing feature was then predicted from the k hidden nodes. Then, the next missing feature was predicted from the k hidden notes plus the first missing feature. If the true value of the first feature was known, it was used; otherwise, the prediction was used. This continued for all four missing features, so the final missing feature had $k + m - 1$ weights from the outputs of the hidden layer and the $m - 1$ previous predicted features. Backpropagation from the m missing features was applied to the input weights when the missing features were present, but backpropagation was not carried through the cascading connections between the m filled features. (A version of the cascading autoencoder without any

backpropagation from the missing values was also tested, but it did not perform as well, so it has not been included.)

Iterating Cascading Autoencoder (iCasc) First a cascading autoencoder was used as described above to update the missing values. The matrix with filled values was used to train a new cascading autoencoder, which provided updated predictions. This process was repeated 5 times.

3.3 Results

The performance of missing data algorithms was dependent on the number of samples available for training. When the full dataset (244 samples) was used (Figure 3.3a), DA performed the best, with iRegr nearly as good. Casc, iCasc, and iAE had good performance as well. Mean filling was worse than all others. When a small dataset (40 samples) was used (Figure 3.3b), DA became the worst, as expected, while iPCA, AE, iAE, Casc, and iCasc performed equally well. When a medium-sized dataset was used (100 samples; results not shown), DA, iAE, iRegr, Casc, and iCasc all performed well.

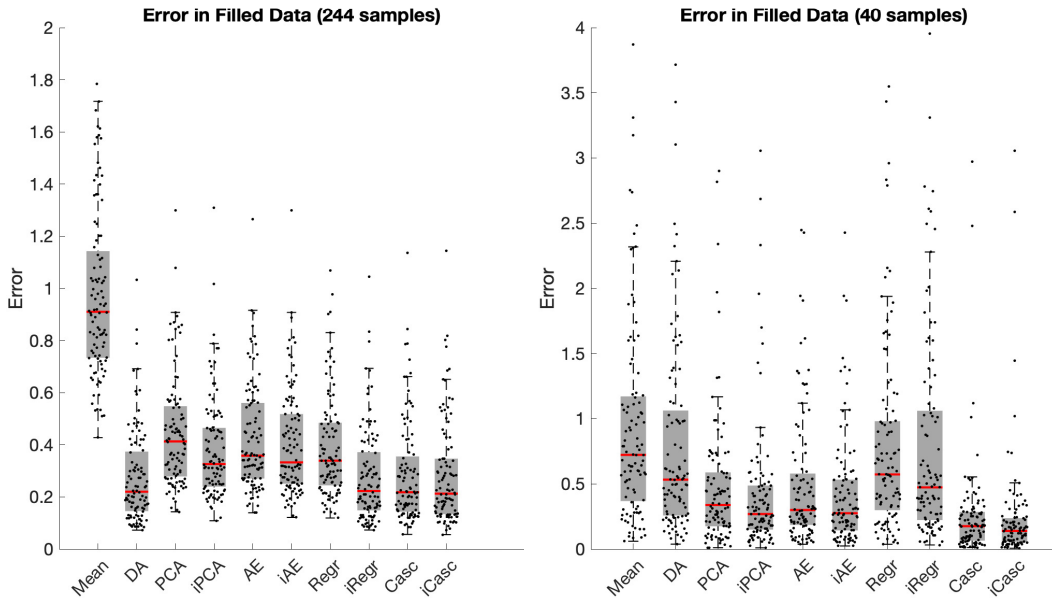
As shown in Figure 3.4, iRegr and DA runtimes were much faster than Casc and iCasc. Runtimes were relatively similar across sample sizes.

Casc, iCasc, AE, and iAE performance was relatively insensitive to the number of hidden nodes, but was best with 6. PCA performance was approximately constant for $4<k<20$, so it was also set to 6. ADADELTA converged fastest with $\rho = 0.99$; ϵ was set to 10^{-7}.

Since iAE, Casc, and iCasc perform quite well across all sample sizes, and since iCasc outperformed the other algorithms in some tests, it was chosen to fill missing data for all further analysis in spite of its long runtime.

3.4 Discussion

Mean substitution, while common in the literature, did not perform as well as the machine learning methods DA and iCasc. Since simpler methods, like mean substitution and CCA, are easy to understand and to implement in



(a) Full Dataset

(b) Small Dataset

Figure 3.3: Error rates for each method when filling the missing values in a (a) full-sized (n=244) and (b) small (n=40) dataset. In (b), one data point each from Mean, DA, and iRegr is outside the bounds of the plot.

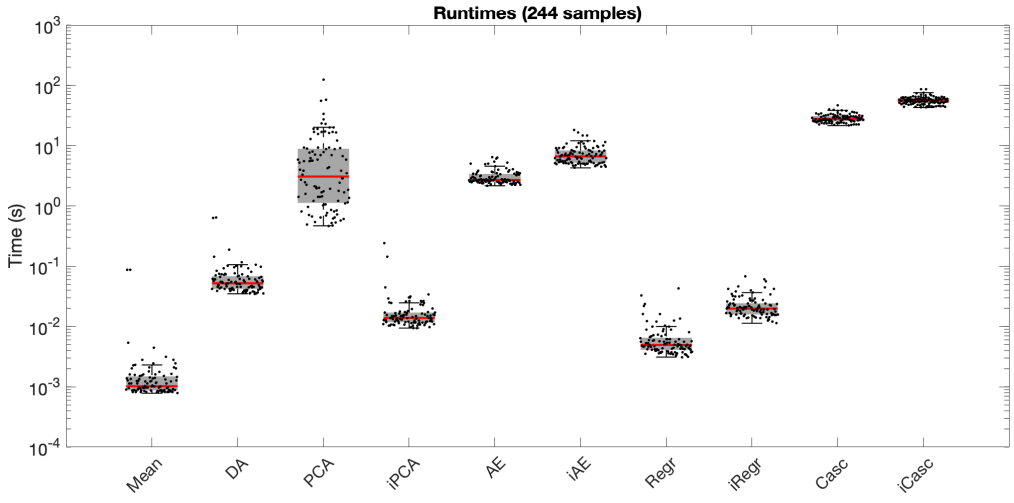


Figure 3.4: Runtimes for each method when filling the missing values in a full-sized (n=244) dataset.

common software packages, it is unsurprising that they are frequently used in spite of their poor performance. While linear regression was performant and is available in Matlab and other software packages, it is more difficult to configure than mean substitution. Additionally, its performance might be worse on other datasets with more missing features, since relatively few values were missing in this dataset. The novel cascading autoencoder performed better than the other techniques across a variety of sample sizes. Future work could consider its application to other datasets, especially those with more missing features or a higher proportion of missing data. The goal in this work was to optimize a method for this specific application, the NET dataset, not necessarily to generalize results to other datasets, so further work would be necessary to determine if these methods, especially the novel cascading autoencoder, are effective with different types of data or rates of missing values.

This experimental setup simulated data missing completely at random (MCAR), since samples were deleted without consideration for their correlation. Many missing data techniques perform well on data that is missing at random (MAR) rather than MCAR [14], so our results are likely to generalize to MAR data, but additional experimentation could verify that assumption. Since the true NET data is likely missing not at random (MNAR), it is possible that the imputation is not estimating the unknown values effectively. Some measures, such as temperature, are almost certainly MCAR (due to experimenter or clinician forgetfulness), but other measures, like refractoriness, are likely due to biological differences which may or may not be captured in other measured values (therefore resulting in MAR or MNAR data). Further work could consider follow-up experimentation to attempt to measure values that were initially missing in order to gauge the effectiveness of these missing data techniques.

Future work could consider optimizations to these algorithms. Since the thirty-five features in this dataset were derived from waveforms with dozens of samples, accuracy could be improved by including that auxiliary data in the missing-data analysis. (For more information on the use of auxiliary variables, see Dong and Peng [14].) Since these features are measurements of waveforms,

polynomial or spline fits of those waveforms could also provide a method of filling this missing data. Since the thirty-five features were originally selected by experts, it could also be effective to get input from those experts regarding which other features or which regions of the original waveforms would be most useful for reconstructing the missing values.

3.5 Conclusions

Since the best performance was achieved by the iterating cascading autoencoder, missing data in the remaining chapters will be filled with the iterating cascading autoencoder.

Chapter 4

Site-Specific Differences

This chapter describes the second phase of the project: determining if site-specific effects preclude the combination of international data. Full realization of the diagnostic potential of the Nerve Excitability Test (NET) requires a substantial database of normative values. The objective of this study was to determine if it is statistically permissible to combine data from multiple sites given the inherent site-specific technical heterogeneity and potential for biological differences associated with race.

Secondary analysis of data collected in previous studies (Japan n=85, 50 male, ages 19-86; Portugal n=42, 14 male, ages 22-84) was used together with new normative data (Canada, n=150, 73 male, ages 18-70). An algorithm to detect site-specific differences, based upon machine-learning clustering and variation of information, was developed to detect site-specific differences that would preclude simple amalgamation of multi-site data into a single dataset. Regression analysis on the pooled dataset was compared to previous site-specific analysis that identified age and sex as covariates for NET results.

The pooled dataset showed low levels of site-specific heterogeneity (94% homogeneity, $p < 0.0001$), indicating pooling is appropriate. Pooling data did not obscure the previously reported relationships between age and NET results (specifically superexcitability and hyperpolarizing I/V slope). There were differences in some NET results between countries that were likely methodological in origin, rather than biological, but these differences were minor. We conclude that it is statistically permissible to combine multi-site data into an

international normative database to improve the diagnostic potential of the NET.

4.1 Introduction

When measurements are gathered by different experimenters, from different locations, or with different types of equipment, there may be errors specific to the experimenter, location, or equipment. For example, if one technician takes blood pressure readings at a location that is, on average, slightly farther down the arm than another technician, her readings will be slightly higher than her colleague's [60]. If she and her colleague both measure 100 different people, they will find they have a similar variance, and her average reading might even be lower than his, but the minor difference between their measurement methodology has introduced a minor error between their sets of 100 measurements. If that difference can be determined quantitatively, it can be subtracted from all of her values (or added to all of his) to make the measurements more comparable. Since both of the technicians still have some variance in the location they place the blood pressure cuff, this correction won't remove all error—and it might even increase the error in an individual case if the correction decreases the blood pressure of a reading that was already measured too low—but overall it will reduce the dataset's error and allow the two sets of data to be more accurately combined. This situation gets more complicated if the first technician is measuring people in Japan while the other is measuring in Portugal. If we observe the means of the two sets of measurements are different, we cannot know if the difference is due to technical or biological differences. In a more realistic test protocol, there may be dozens of measurements, each subtly impacted by differences in equipment or technique.

The Nerve Excitability Test measurements have been collected from different labs around the world. As a result, it is possible that there are site-specific differences in the data. For example, there may be differences due to biology (i.e. race), equipment, or methodology (which may be lab-specific or experimenter-specific). In order to assert that it is appropriate to combine

these measurements into a single dataset, it is necessary to determine if there are any outstanding differences between datasets from different locations. The problem of errors specific to experimenter, location, or equipment has been the subject of considerable research interest in the field of proteomics. In gene microarray datasets, the errors can be much larger than the effects of interest [58, 61]. These effects are called "batch effects" because the error is associated with a batch of measurements. Batch effects are undesired non-biological effects which obscure the desired signal, which is usually biological. Microarray experiments attempt to separate these signals by spreading biological differences between different batches. Here, biological differences are separated into different batches. As a result, if any site-specific differences are present, it might not be possible to determine whether they are true batch effects (i.e. due to non-biological differences such as different equipment, electrode placement, filter settings, etc.) or biological differences. For the purpose of this paper, we are considering any site-specific differences — including biological — to be "batch effects."

Since batch effects are an important problem for gene microarrays, many techniques have been developed to detect and correct for these errors. Many of those methods rely on particular knowledge about the type of data. However, adjusted rand index (ARI) and variation of information (VI) can effectively detect batch effects without making any assumptions about the underlying data [40]. ARI measures the similarity between two partitionings of the same n samples. If the n samples are sorted into k clusters by some clustering algorithm (e.g. k-means), those clusters can be compared to the known labels to measure how effectively the unsupervised clustering was able to predict the batches. However, ARI may take negative values, which complicates interpretation [63], and it has a weak dependence on n [40]. An alternative is a technique based on VI, which is based on the mutual information between two random variables [40].

If batch effects have been detected, they can be corrected with a variety of techniques. Many common batch effect correction techniques are specific to gene microarrays. They use knowledge about the microarrays to understand

and correct the problems, so they are not relevant to NET data. There are techniques that are not specific to gene microarrays (e.g. Vaisipour [61]), but they are not considered in detail here since no batch effect correction will be applied (as described in Methods and Results).

4.2 Methods

Variation of Information (VI) is based on the mutual information between two random variables (i.e. how knowledge about one clustering provides knowledge about the other). Applied to batch effects, VI is a measure of how much information a clustering algorithm provides about the labeled clusters, which suggests the degree to which obvious batches are present. VI is defined as

$$VI(L, C) = H(L) + H(C) - 2I(L, C), \tag{4.1}$$

where $H(L)$ is the entropy of L and $I(X, C)$ is the mutual information between L and C. In the case of batch effects, L is the vector of true labels and C is the corresponding vector of cluster labels.

While VI is an effective measure for detecting batch effects in a fixed number of clusters (i.e. batches), it does not allow comparison when the number of clusters is different. For example, the VI for three batches is of a different magnitude than for four batches, so they cannot be compared. In clustering applications where the number of clusters is not constrained, the maximum value of VI is dependent on the number of data points ($VI_{max} = log_2(n)$) because the number of clusters, K, can increase up to n. However, when K is fixed,

$$VI_{max}(L, C|K) = 2log_2(K), \tag{4.2}$$

where L is the true labels (i.e. the country of origin) and C is the cluster labels. This is clearly dependent on the number of clusters.

While maximum VI is described in Equation 4.2, the maximum for a specific set of cluster labels may be smaller if the clusters are of unequal sizes. If there are no batch effects, an algorithm which evenly distributes the clustered data into K groups is expected to assign a sample to label k with probability

$p(k) = 1/K$, which has an expected entropy of $2log(K)$, as expected. However, if the labels are not neatly divided into equal-sized groups, their entropy will be less than $log(K)$ while $I(L, C)$ remains equal to zero. The resulting VI (by Equation 4.1) will be less than the maximum. In other words, since the individual datasets might not be equal sizes and the clustering algorithm is not constrained to find equal-sized clusters, the expected maximum VI is less than Equation 4.2. In those cases, the expected maximum, $E[VI_{max}]$, is determined empirically by shuffling the labels produced by the clustering algorithm 100 times and calculating the mean VI of those 100 shuffles.

$$E[VI_{max}(L, C|K)] = \frac{1}{100}\sum_{j=1}^{100} VI(L, C_j|K), \tag{4.3}$$

where C_j is the jth shuffle of the cluster labels C. Since shuffled cluster labels by definition exhibit no batch effects, the maximum VI for shuffled cluster labels is the maximum possible VI score, and the expected VI after many different random arrangements is the expected VI for un-batched data. Therefore, $VI(L, C)$ can be compared against $E[VI_{max}(L, C)]$ to determine the magnitude of the batch effects.

VI compares the labels of clustered data to the true labels. To generate cluster labels, an agglomerative hierarchical cluster tree was used (Matlab `linkage` function, default `ward` linkage method). Results were similar when k-means was used (not shown), but results from the linkage cluster are preferred because they are deterministic. (Since k-means clusters are dependent on the initial seed, k-means increases the variance.) VI calculations were based upon methods in Bishop [6] as implemented in Matlab by Chen [13].

Before calculating VI, the missing values were filled with the missing data method that performed the best (a cascading autoencoder; see Section 3.3). The data was imputed separately for each dataset (e.g. each country) before calculating VI to ensure that the imputation method did not increase homogeneity between the datasets by sharing imputation knowledge between datasets.

When calculating VI for a given combined dataset, the data from each individual dataset (e.g. country) was first normalized and randomly sampled

30 times. Normalization consisted of converting each feature to zero mean, unit variance. Each random sample included 80% of the combined dataset. A pre-calculated series of 30 random seeds was used for the 30 trials each time a dataset was tested for batch effects. For each of the 30 trials, the expected maximum VI was calculated as described above (Equation 4.3), giving 30 pairs of results. Comparing these allows two measures to be calculated: a p-value and the dataset homogeneity. A paired t-test gives a p-value indicating whether batching effects are not present:

$$p = ttest(VI(L, C|K), E[VI_{max}(L, C|K)]), \tag{4.4}$$

where C is a matrix of 30 cluster label vectors, so $VI(L, C)$ is a vector of length 30. Note that large p-values indicate that no batching is present; the VI of the data is drawn from the same distribution as a random VI. If instead the p-value is small, some amount of batching is present, but the p-value does not indicate the amount of batching. The homogeneity of the data sources (i.e. the effect size of the batching) is calculated from the ratio of the means:

$$Homogeneity = \frac{VI(L, C|K)}{E[VI_{max}(L, C|K)]}. \tag{4.5}$$

Most of the following figures are ladder plots of the 30 homogeneity scores with mean homogeneity and the p-value displayed below each ladder. The entire analytics process is diagrammed in Figure 4.1.

As an additional test to confirm behavior of the VI homogeneity measure, a dataset of common peroneal (CP) nerve measurements (N=120, 57 male, ages 18–70) were used in some experiments (with missing data methods the same as those used for the normative data). Since CP measurements are different from median nerve measurements, VI for that dataset was expected to be higher. A rat dataset (n=49) was also used for comparisons. Since the rat data was from (1) a different species (2) measured in different locations (3) under the effect of anaesthetics, it was also expected to show significant differences. The rat dataset and the confounding effects of the ketamine-xylazine anaesthetic are described in more detail by Lorenz and Jones [37] and Bell et al. [4], but those details are not relevant to this comparison. Table 4.1 describes these datasets.

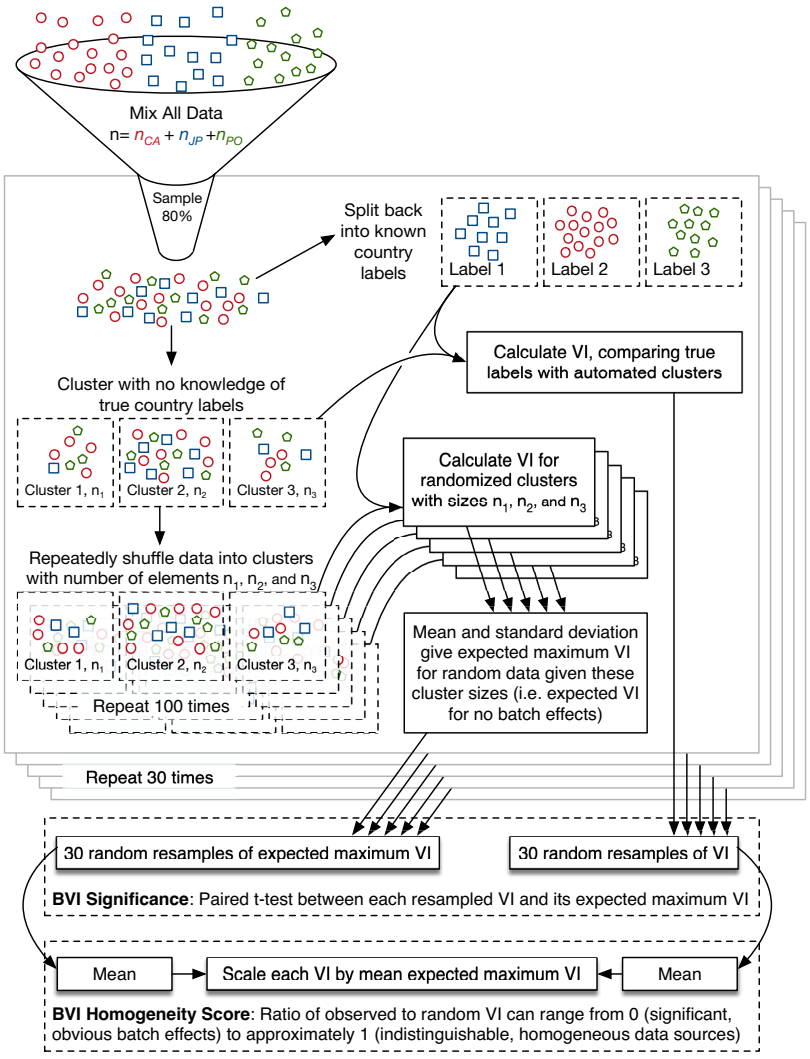


Figure 4.1: This block diagram shows how dataset homogeneity is calculated for the normative dataset. Data from Canada (CA), Japan (JA), and Portugal (PO) is preprocessed separately (not shown) before being resampled (80% of samples) 30 times to calculate mean VI homogeneity. The resampled data is automatically clustered without the use of the country labels. The variance of information (VI) between these clusters and the known country labels gives a measure of the difference in their information content; high values indicate they contain different information and therefore do not demonstrate batch effects. Since the maximum VI depends upon the relative number of samples in each cluster and the number of clusters, the contents of the clusters are shuffled randomly 30 times while preserving the number of samples in each clusters. The VI between these random clusters and the true labels gives the expected VI for truly random clusters. The homogeneity score is the VI of the automatic clusters scaled by the expected random VI, giving a value with a minimum of 0 (when the clusters exactly match the known labels) and approximately 1 (when the clusters are random relative to the known labels). The maximum is not exactly 1 because the scaling factor is the expected random VI rather than the absolute maximum.

Table 4.1: NET data used for comparison with normative median data. Both of these are not the same as normative median data.

Name	n	n male	Ages	Description
Legs	121	57	19–70	common peroneal (leg) nerve
Rats	49	—	—	various nerves in anaesthetized rats

In order to provide intuition about the algorithm's ability to detect batch effects, Figure 4.2 shows six plots demonstrating various amounts of homogeneity. In panel A, data from two independent sources is obviously batched; the blue circles and the green crosses do not overlap, so the algorithm determines that they are 0% homogeneous. In panel B, the two datasets are drawn from the same distribution, and the algorithm correctly determines the data is 100% homogeneous. In both of these cases, the algorithm easily measures homogeneity, just as a human observer would. Panel C is similar to A, but some of the green crosses are drawn from the same distribution as the blue circles, making the data less homogeneous (35%). The green and blue points are still obviously from different sources, but not as obviously. Panel D is similar, but with an unequal number of points in each group. Panel E shows a situation that a human observer would find difficult to differentiate. The two groups are drawn from similar distributions that are difficult to separate. In spite of this challenge, the algorithm is very confident ($p < 0.00001$) that the data is not completely homogeneous (95%). Even with fewer samples (panel F), the algorithm still suggests there are some batch effects. These plots show that VI is able to detect when two independent distributions are overlapping with different variances, suggesting that VI homogeneity is sensitive to batch effects.

All code was implemented in Matlab R2018a (9.4.0.813654) and is available online [2]. Actual p-values are reported unless lower than 0.0001.

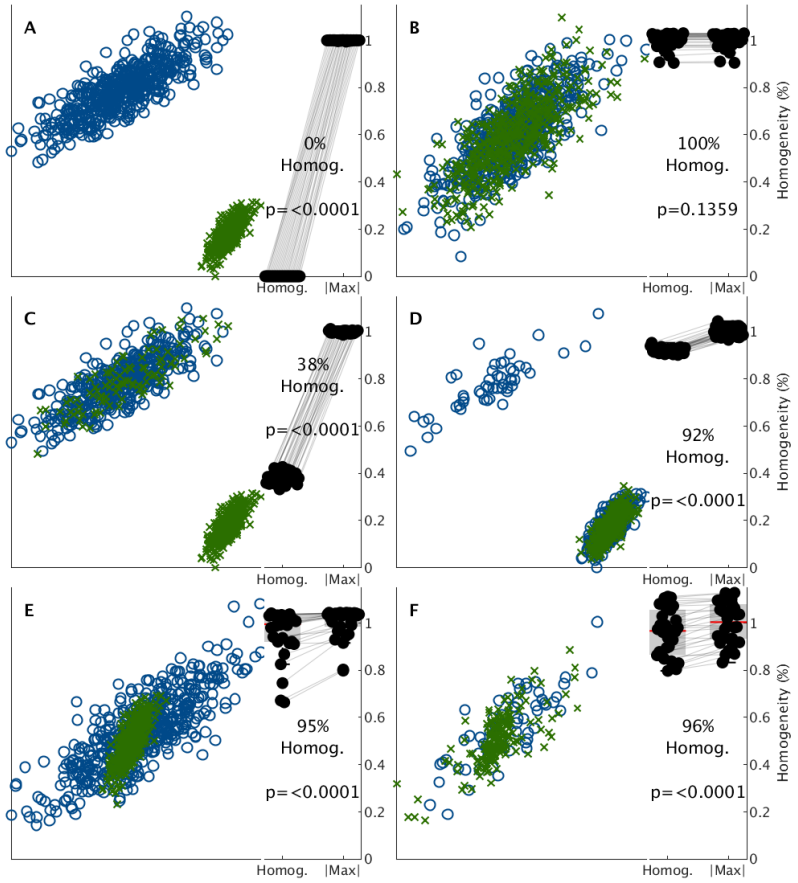


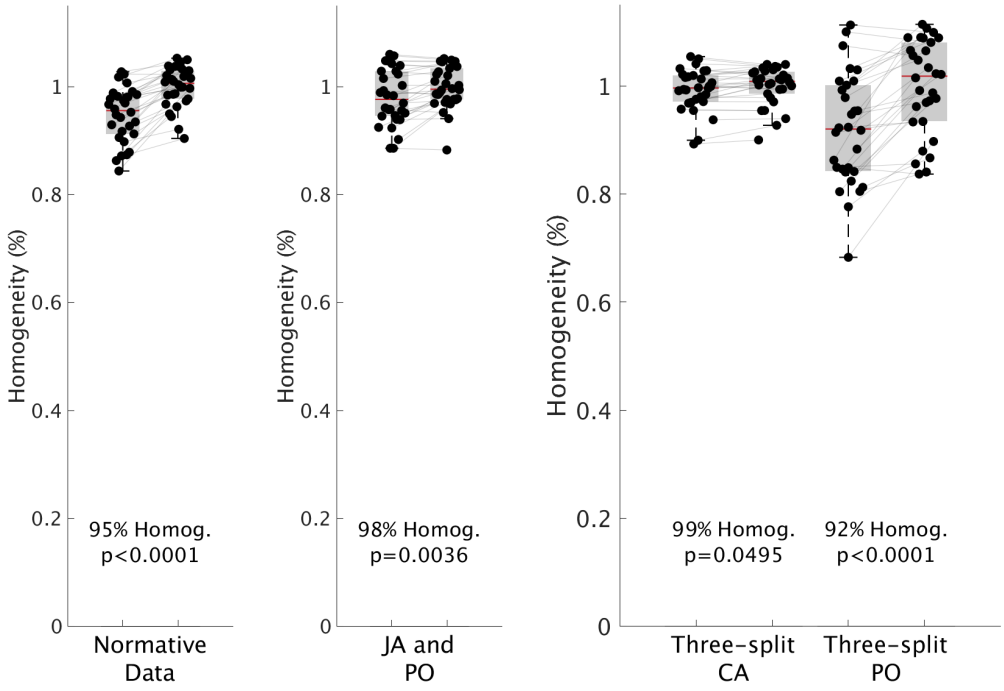
Figure 4.2: Homogeneity of synthetic data. In each example, data is generated from three Gaussian sources. Source 1 has mean $[2, 3]$ and sigma $[11.5; 1.54]$. Source 2 is not overlapping with source 1 and has larger variance, with mean $[-1026]$ and sigma $[2015; 1516.5]$. Source 3 has the same large variance as source 2, but with a mean nearly identical to source 1: $[3, 4]$. In each panel, two datasets, generated from a combination of the Gaussian sources, are plotted on the left. On the right is a ladder plot of the homogeneity. (**A**) Data is generated by two different processes. Blue (circles) is 500 data points generated from source 1. Green (crosses) is 500 data points generated from source 2. The VI measure suggests this data is 0% homogeneous, i.e. that the data is fully batched from two different sources. (**B**) All 500 data points for each dataset have been generated from source 2, resulting in 100% homogeneity. The large p-value indicates the data is *not* from different distributions, as expected. (**C**) This test is similar to example A, except 100 of the points that were blue are now green. This increases the homogeneity of the data, as expected. (**D**) Blue data consists of 50 data points from source 1 and 450 from source 2, while green consists of the same 500 points in example A. Since most of the combined dataset comes from source 2, it appears quite homogeneous, but the outliers generated from source 1 are enough to drop the homogeneity score to 92%. (**E**) The blue data is generated from source 1, as in example A, but the green data is now from source 3. The complete overlap between these sources makes it difficult to distinguish between these datasets, yet the VI method is able to detect the difference, giving a small (5%) batch effect. (**F**) Even with a much smaller sample size, batch effects are detectable. (63 blue data points from source 1 were compared to 63 green data points from source 1 and 126 green data points from source 3.)

4.3 Results

Figure 4.3 shows the homogeneity of the normative data. The homogeneity score of 95% indicates that the data has a small dependence on its country of origin. When each pair of countries is compared (Figure 4.3b), the homogeneity ranges from 97–98 %, indicating no one country is disproportionately responsible for heterogeneity. When the data within each country is randomly split into three separate groups (Figure 4.3c), Canada and Japan show very little heterogeneity (99% Homogeneity), while Portugal's Homogeneity of 92% is much lower than the 100% that is expected for these completely random splits of data. This may be due to the smaller sample size from Portugal. (However, note that the small p-value for PO splits indicates the algorithm found significant differences between PO groups.) These results suggest that the normative data is mostly homogeneous, with minor site-specific differences.

Normative human median nerve data is next compared to other data to determine if the algorithm can detect non-normative data (Figure 4.4). In aggregate, human CP nerve (leg) data is distinguishable from median nerve (arm) data, though individual samples are indistinguishable. As a result, a dataset of leg data is identifiable as different from arm data. Rat data is also dissimilar to human data, especially when rats are under the effect of the anaesthetic ketamine-xylazine, so combining rat and human data also results in low homogeneity. The results from comparing human and rat leg data to human arm data provide strong empirical evidence that the homogeneity measure is effective at identifying biologically plausible differences between NET results.

The homogeneity scores were not sensitive to the method of filling missing data. These results used an iterating cascading autoencoder to fill missing data because it performed best (Chapter 3, but the results are similar when using a combination of iterating linear regression for larger datasets and mean imputation for smaller datasets (results not shown).



(a) (b) (c)

Figure 4.3: The homogeneity of the normative data is 95%. This high homogeneity is confirmed by various splits and recombinations of the normative data.
(**a**) The homogeneity of the Normative Data (i.e. combined CA, JA, and PO data) is compared to the expected VI (assuming no site-specific differences). The data is mostly homogeneous, but has some effects that depend upon the country of origin.
(**b**) Homogeneity for only two of the three countries in the normative dataset shows that none of the countries is disproportionately responsible for the site-specific differences. (Results only shown for one of three possible combinations.)
(**c**) The absence of site-specific differences within a country's dataset shows that the datasets are internally homogeneous. Each country's data is randomly split into three equal sub-groups and compared to random data of the same size, with the normative data for comparison. Since this test is splitting data from within a single country into three groups, it should have a high homogeneity score. The scores for CA (99%) and JA (99%, not shown) indicate that their data is homogeneous. The PO homogeneity score (92%) is quite a bit lower, but note PO has a much smaller sample size.

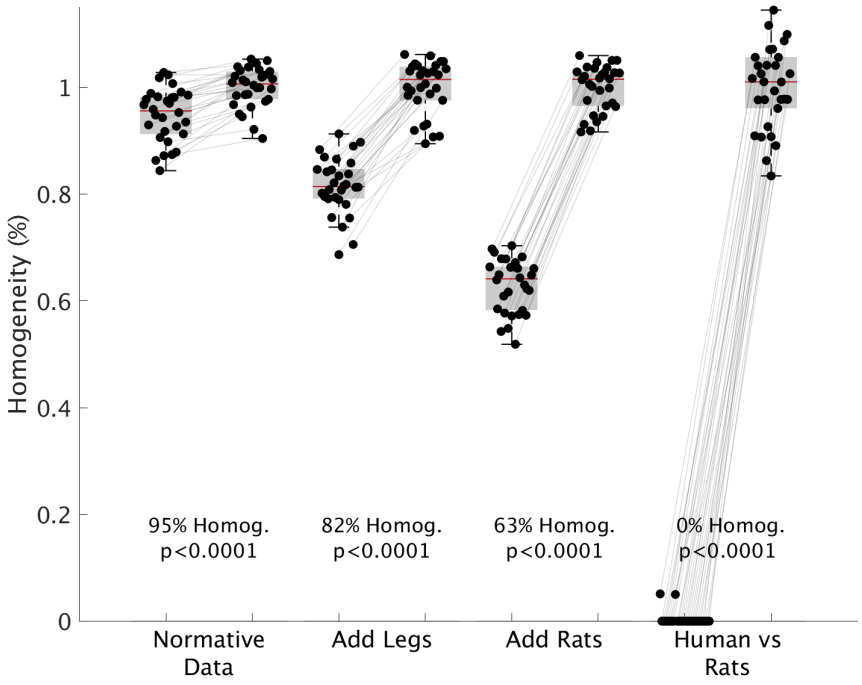


Figure 4.4: Adding non-normative data to normative data results in a decrease in homogeneity. When the leg or rat data is added to the normative dataset (simulating an attempt to use data which is significantly different from normative), homogeneity increases. This is in spite of heterogeneity of the rat data, which consists of measurements from three different muscles (tibialis anterior, soleus, and tail) under the effects of two different anaesthetics (sodium pentobarbital and ketamine xylazine) in two different clinical conditions (healthy and spinal cord injury).

4.4 Discussion

International normative nerve excitability data can be safely combined, allowing larger datasets to be collected and shared across testing locations. While site-specific effects are present, they account for a small proportion of the variance in the data. This is contrary to our initial expectations that site-specific differences would necessitate batch correction measures.

Since Homogeneity for the normative data is close to 100%, we conclude there are not meaningful site-specific differences present in the data. This was contrary to our initial expectation that we would observe differences between the groups. If site-specific differences had been present, it would be necessary to consider whether they are due to experimental differences (e.g. equipment or technique) or due to site-specific biological differences. Since the data was collected without information about participants' race, it would not have been possible to conclusively determine whether the differences were technical in nature, which would allow them to be corrected, or biological in nature, in which case it would be inappropriate to combine the international datasets. However, as will be discussed in more detail in Chapter 5, these effects might be due to differences in SR curves, since they show notable differences between countries. Differences in SR curves can be attributed to experimental data collection differences rather than biological differences, and these differences do not impact the other measures, so they are of lesser clinical importance. As a result, we can conclude that biological differences are not a large factor in the variation in NET results.

To be confident in the homogeneity score for the normative data, it is necessary for the score to be near 100%, but that is not sufficient. The Homogeneity must also change significantly when non-normative data is added. The results when adding leg and rat data show that the homogeneity score is sufficiently sensitive to differences between batches.

The 92% Homogeneity with $p < 0.0001$ when Portuguese data is split into three random sub-groups is likely due to the high heterogeneity of the Portuguese data (e.g. more females than males and a broad temperature

range) and its small sample size. Even though randomly splitting a dataset into three sub-groups should result in homogeneous groups, that expectation is less likely to hold in smaller samples. Future work with this homogeneity measure should provide guidelines for appropriate minimum sample sizes.

4.5 Conclusions

The NET dataset was tested for site-specific differences using machine learning techniques: a linkage cluster and variation of information. These tests indicated that meaningful site-specific differences are not present in the data. Since site-specific testing difference are not a major factor in the Canadian, Japanese, and Portuguese data, it is appropriate to combine them into an international normative database for NET results. This method of combining the datasets can easily be applied to datasets from other countries, allowing international collaboration on an even larger normative database. These analysis methods are also extensible to non-normative datasets and can be used to consider how different non-normative datasets are from the normative dataset, providing diagnostic utility in clinical applications.

Chapter 5

Nerve Health Score

This chapter describes the third phase of the project: development of a nerve health score. The objective of this portion of the project was to combine international nerve excitability test results into a normative dataset of healthy humans and provide a computerized clinical decision support system (CDSS) to identify unhealthy peripheral nerves. This CDSS is the NerveNorms website, which uses international norms as a standard dataset for comparison against healthy human controls [3]. NerveNorms also calculates a nerve health score to determine if new patient measurements are healthy.

5.1 Introduction

Several previous studies have addressed the clinical neurophysiology of nerve excitability tests in normative human controls to provide insight into the effects of age and sex on ionic properties of peripheral motor nerve axons [1, 11, 24, 38]. These results were combined into an international normative dataset of nerve excitability test results in healthy humans (Chapter 4).

Temperature, age, and sex can impact the biophysical properties of peripheral nerves. The effects of temperature on some measures of axonal excitability are large [32, 33, 55]. To mitigate temperature effects, standard recording practice is to maintain temperature above 32°C or to apply temperature correction. Effects of age and sex have also been studied, but with mixed results. The effects reported in four previous studies from Australia (AU), Japan (JP), Ireland (IE), and Portugal (PO) have not been consistent, with some speculation

that racial differences may be responsible [1]. The studies all followed similar designs. Table 5.1 shows the age effects found in previous studies (along with our hypotheses based upon the consensus). These previous studies agreed upon age effects in hyperpolarizing I/V slope, stimulus-response slope, and maximum CMAP. There were mixed results for many other measures, but none of them were diametrically opposing. Table 5.2 similarly shows the effects of sex (along with our hypotheses). There was no agreement on any sex dependence, since two of the four studies did not find any sex-dependent effects.

Variable	**AU**	**JP**	**IE**	**PO**	**Hypothesis**
Max CMAP	ns	—	$\downarrow^{****}$	$\downarrow^{**}$	$\downarrow$
Stim at 50%	$\uparrow^{***}$	ns	ns	ns	mixed, ns
SR slope	$\downarrow$ (?)	—	$\downarrow^{*}$	$\downarrow^{**}$	$\downarrow$
Rheobase	$\uparrow^{**}$	ns	ns	ns	mixed, ns
SDTC	ns	$\uparrow^{*}$	ns	ns	ns
TEd (90–100ms)	ns	$\downarrow^{*}$	$\downarrow^{*}$	ns	mixed, $\downarrow$
TEh (90–100ms)	ns	ns	ns	ns	ns
TEd peak[1]	ns	ns	ns	ns	ns
TEd20 (peak)	—	—	ns	ns	ns
TEd undershoot	—	—	ns	ns	ns
Accom half-time	—	—	ns[3]	$\downarrow^{**}$	mixed, $\downarrow$
Hyper. I/V slope[2]	$\uparrow^{***}$	$\uparrow^{***}$	$\uparrow^{*}$	—	$\uparrow$
Min. I/V slope[2]	$\uparrow^{***}$	$\uparrow^{***}$	$\uparrow^{*}$	ns	mixed, $\uparrow$
Resting I/V slope[2]	ns	$\uparrow^{*}$	ns	ns	mixed, ns
Superexcitable	$\downarrow^{***}$	$\downarrow^{*}$	$\downarrow^{**}$	ns	mixed, $\downarrow$
Late subexcitable	ns	ns	ns	ns	ns
Refractoriness 2.5	$\downarrow^{**}$	—	ns	ns	mixed, ns
RRP	ns	ns	ns	ns	ns

Table 5.1: Age effects observed in previous nerve excitability tests in normative human median nerve. An upward arrow means the absolute value of the measure increased with age. "ns" indicates relationships which were not significant, and a dash was unreported.
[1]JP reported depolarizing TE at 10–30ms instead of the peak value.
[2]Rather than the standard I/V measures, AU reported I/V slope in the hyperpolarizing and depolarizing directions. Japan reported I/V threshold at 50% depolarizing and 100% hyperpolarizing instead of slopes. The numbers in this table are an attempt to translate those non-standard measures.
[3]For IE, after temperature correction, this became significant.
AU: Australia; JP: Japan; IE: Ireland; PO: Portugal.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

Variable	AU	JP	IE	PO	Hypothesis
Max CMAP	ns	—	ns	ns	ns
Stim at 50%	ns	—	$\downarrow^{**4}$	ns	mixed, ns
SR slope	ns	—	ns	ns	ns
Rheobase	ns	$\downarrow^{*}$	$\downarrow^{**4}$	ns	mixed, $\downarrow$
SDTC	ns	ns	ns	ns	ns
TEd (90–100ms)	ns	ns	ns	ns	ns
TEh (90–100ms)	ns	ns	ns	ns	ns
TEd peak[1]	ns	$\uparrow^{**}$	$\uparrow^{*4}$	ns	mixed, ns
TEd20 (peak)	—	—	$\uparrow^{**4}$	ns	mixed, $\uparrow$
TEd undershoot	—	—	$\uparrow^{***}$	ns	mixed, $\uparrow$
Accom half-time	—	—	$\downarrow^{**3}$	ns	mixed, $\downarrow$
Hyper. I/V slope[2]	ns	$\uparrow^{***}$	ns	—	mixed, ns
Min. I/V slope[2]	ns	$\uparrow^{***}$	ns	ns	mixed, ns
Resting I/V slope[2]	ns	ns	ns	ns	ns
Superexcitable	ns	$\uparrow^{*}$	$\uparrow^{**}$	ns	mixed, $\uparrow$
Late subexcitable	ns	$\uparrow^{**}$	$\uparrow^{**}$	ns	mixed, $\uparrow$
Refractoriness 2.5	ns	ns	$\downarrow^{3}$	ns	ns
RRP	ns	ns	$\downarrow^{3}$	ns	ns
Refractoriness 2	—	ns	—	—	ns

Table 5.2: Sex effects observed in previous nerve excitability tests in normative human media nerve. An upward arrow means the absolute value of the measure tends to be higher in females. "ns" indicates relationships which were not significant, and a dash was unreported. All PO measures were $p > 0.2$.
[1]JP reported depolarizing TE at 10–30ms instead of the peak value.
[2]Rather than these standard measures, AU reported I/V slope in the hyperpolarizing and depolarizing directions. Japan reported I/V threshold at 50% depolarizing and 100% hyperpolarizing instead of slopes. The numbers in this table are an attempt to translate those non-standard measures.
[3]Results are shown after temperature correction; these were not significant before correction.
[4]Results are shown before temperature correction; these were not significant after correction.
AU: Australia; JP: Japan; IE: Ireland; PO: Portugal. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

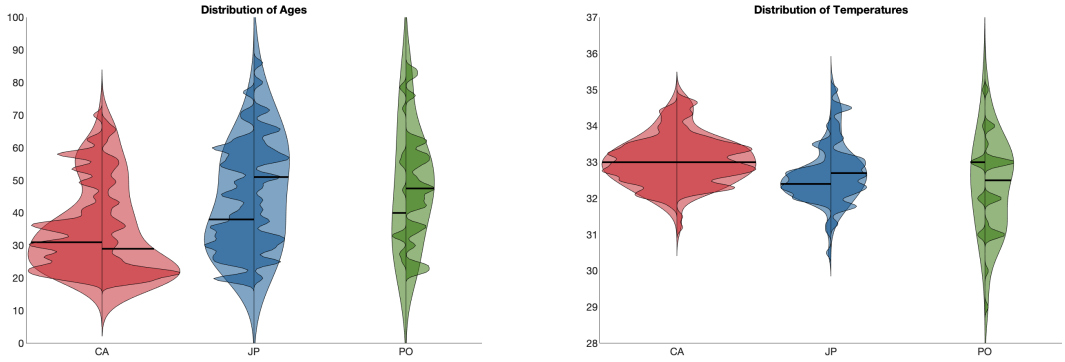


Figure 5.1: Age and Temperature distributions between the three countries are not the same. The left side of each violin plot shows the distribution of males; the right side, females. The area of the plot corresponds to the relative proportion of samples. The plots show the distribution across two different densities.

We have chosen not to repeat the regression analysis of previous studies due to the unequal distribution of covariates (age and sex) and a confounding variable (temperature). As shown in Figure 5.1, age, sex, and temperature are not the same between countries. As a result, would not be clear whether differences in regression coefficients are due to these covariates and confounders.

While NET contains a lot of clinically useful information, its uptake is limited by the complexity of that information. Clinical tests benefit from ease of interpretability. For example, iScore is a recent tool which accurately predicts poor functional outcomes and risk of death for acute ischaemic stroke [49–51]. When presented with five example stroke cases, half of physicians had a 0% accuracy in predicting the probability of the primary outcome, while iScore was correct 90% of the time [52]. Computerized clinical decision support systems often improve clinician practice, especially when study authors are involved in the development of such systems [17]. We aim to provide such a tool for busy clinicians to aid in answering an essential diagnostic question: "Is this nerve healthy?"

5.2 Methods

Data collection for each country's dataset is described in their respective papers. As described in those publications, informed consent was obtained from all participants and relevant ethics guidelines were followed. When necessary, missing data was filled with an iterating cascading autoencoder, since it has been shown to be the most effective for this data (Chapter 3). All code was implemented in Matlab R2018a (9.4.0.813654). Actual p-values are reported unless lower than 0.0001.

In order to increase the diagnostic utility of the normative dataset, a nerve health score was used to measure the health of individual NET results. This nerve health score ranges from 0 to 1 and can be interpreted as the probability that the NET results were drawn from a healthy human, so a score below 0.01 indicates the result is outside of the 99% confidence interval for healthy humans. It was calculated as follows.

1. For a given plot (e.g. RC), for each data point the Gaussian probability was calculated. (Note for absolute SR, this was calculated in both x and y directions.)

2. The geometric mean of those probabilities gave the probability that each specific plot (e.g. RC) was healthy. The geometric mean preserved the range between 0 and 1. For threshold electrotonus, since there were both depolarizing and hyperpolarizing measurements, potentially at more than one current, the overall score was the geometric mean of the means of each of the plots.

3. A probability was also calculated for the excitability variables (excluding age, temperature, sex, and latency) by taking the geometric mean of each of their individual probabilities.

4. Finally, the geometric mean of the scores for the six plots along with the one excitability variable score gave the final probability. If any plots were missing (e.g. charge-duration data from JP), they were excluded from the mean.

To demonstrate the output of this nerve health score, it was calculated for two participants from each country in the normative dataset, along with the mean for each country and two different rats. The participants were selected arbitrarily to demonstrate a range of ages, sexes, and scores. The rats, "Rat Fast Axon" and "Rat Slow Axon", were the means of the two primary groups from a publication demonstrating differences between fast and slow axons in rats [4]; they were chosen to demonstrate the nerve health score on samples that are obviously not healthy humans.

5.3 Results

To demonstrate the output of the nerve health score, which does not correct for age, temperature, or sex, Table 5.3 shows the nerve health scores of normative humans alongside rats. The human scores appear healthy (0.28–0.46), while the rats are not (0.059 and 0.020). Figure 5.2 shows example nerve health scores for 5 plots each for an example human and rat.

Name	Nerve Health Score	Age (years)	Sex
CA-WI20S	0.41	29	Male
CA-AL27H	0.29	44	Female
JP-20-1	0.28	25	Male
JP-70-1	0.46	72	Female
PO-00d97e84	0.41	78	Male
PO-017182a5	0.37	58	Female
CA Mean	0.87	34.622	—
JP Mean	0.73	45.322	—
PO Mean	0.80	46.289	—
Rat Fast Axon	0.059	0.25	Female
Rat Slow Axon	0.020	0.25	Female

Table 5.3: Nerve health scores for arbitrary participants from each country, demonstrating a range of ages and sexes. The country means show a high probability of being drawn from the healthy distribution, while rat data has a low probability of being drawn from a healthy human population.

Figure 5.3 compares two of the normative datasets, JP and CA. Note the large difference in Stimulus Response, but overlapping means in all other plots. The nerve health score for the JP mean relative to CA norms is 0.46. This

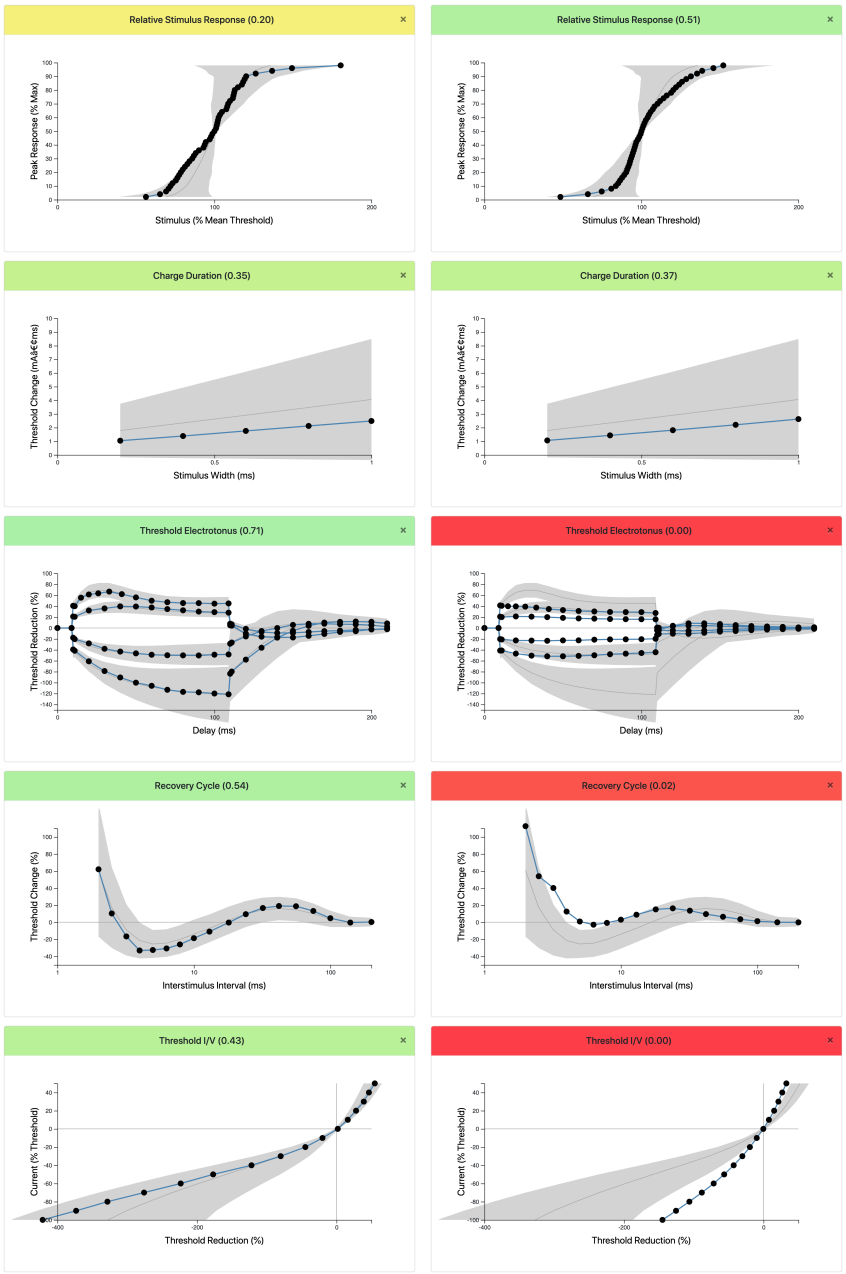


Figure 5.2: Nerve health scores for five plots for an example human (CA-WI20S, left column) and rat (Slow Axon, right column). These images come from the NerveNorms website. Darker red shading in the title bar indicates a smaller (less probable) nerve health score, while green indicates the plot is consistent with a healthy human media nerve. Shaded regions are 99% Gaussian confidence intervals.

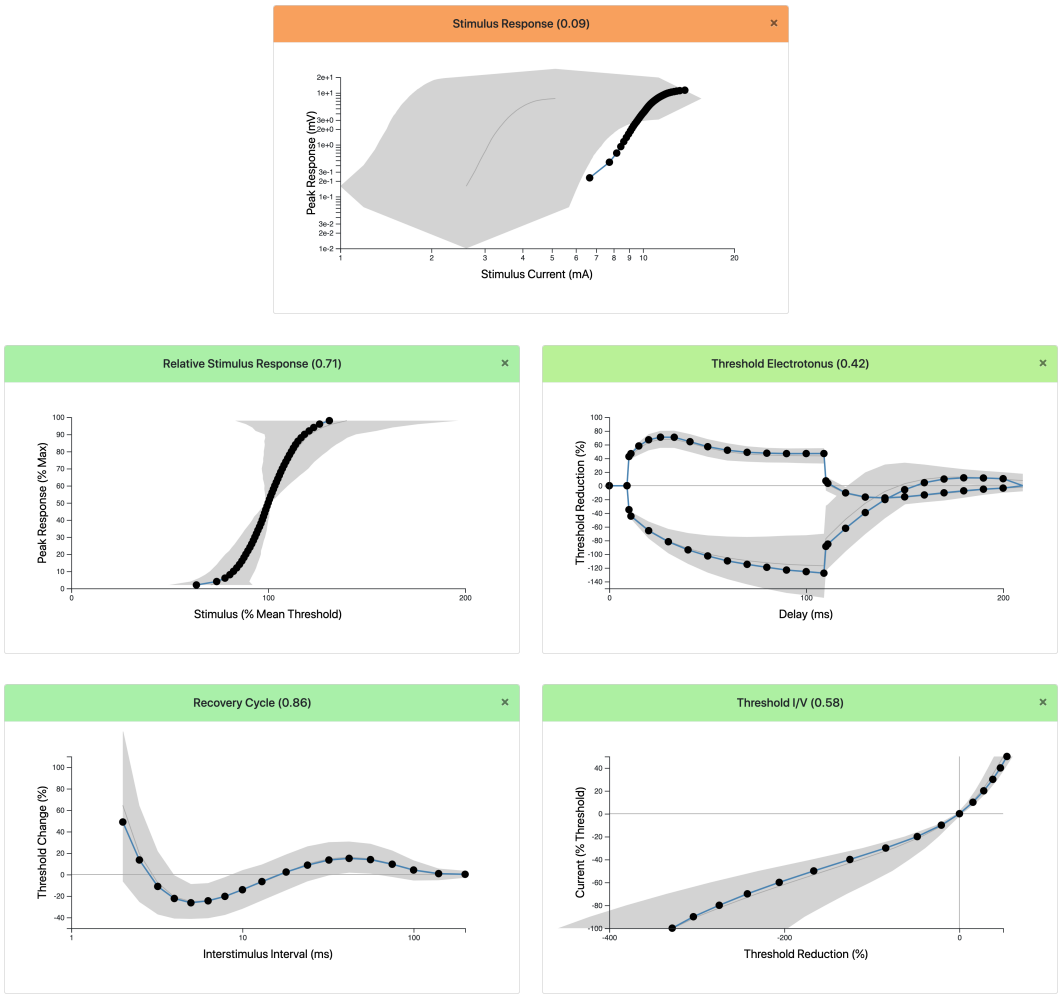


Figure 5.3: The difference between Japanese (JP) and Canadian (CA) normative data, as presented at the NerveNorms website. CA data is in grey (mean and 99% range for healthy data). JP data is black dots and blue lines. Note the large difference in Stimulus Response, but overlapping means in all other plots. (JP data does not include 20% threshold electrotonus and charge-duration.)

comparison was not corrected for age, temperature, or sex.

5.4 Discussion

Skin preparation (technical), skin impedance (biological), and electrode placement (technical) impact NET data, as could distance to nerve (technical due to electrode placement or biological due to wrist circumference). Skin preparation methods were not described in adequate detail in previous studies to determine whether skin preparation could overcome potential biological differences. Different sites may have different methods for placing electrodes. Consider two sites. One site (e.g. JP) places the stimulating electrodes based on anatomical markers, but carefully adjusts the recording electrodes to maximize CMAP. The other site (e.g. CA) carefully places the stimulating electrodes based on the maximally evoked response, but places the recording electrodes based on anatomical markers. The resulting CMAP will be much larger at the first site, but other measures will be unaffected, as shown in Figure 5.3. This shows that the absolute stimulus response is not very important, and site-specific differences in electrode placement are not important for plotting the mean data. It is unclear whether electrode placement could still be a factor in the different variances.

The impact of temperature control is an especially important consideration. Most work to date on human axonal excitability has not maintained temperature precisely. Most data was collected with the aim to keep skin temperature above 32°C, without more precise temperature control. However, Kiernan et al. [32] showed that the effect of temperature is not merely dependent on the present value of the temperature; the rate of change of temperature can also impact axonal excitability. It takes approximately 30 minutes to adequately warm a nerve for excitability testing [33], a restriction which was not followed in previous studies. Instead of controlling temperature, it could be effective to monitor temperature continuously and apply post-hoc correction. However, work thus far has also not monitored temperature precisely. Methods in previous studies ranged from a single temperature measurement at the

beginning or end of the test to PO's temperature measurement before, after, and between each test. However, in all cases, only a single temperature was recorded. Since temperature can vary throughout the test, it would not be reasonable to attempt small-magnitude temperature corrections with only a single temperature recording.

Latency is strongly influenced by temperature, and can serve as a proxy for temperature changes. Future QTRAC versions could monitor latency throughout the test, halting testing if it diverges from its initial value. This would allow clinicians to test at the subject's natural temperature, without the effect of temperature adjustment, unless the subject's temperature changed during the test. Temperature changes could be corrected by external heating or cooling before the test continued. This ongoing monitoring of latency would be similar to the current ongoing monitoring of maximal CMAP, which accounts for changing threshold. Post-hoc analysis could adjust the subject's results based on the single temperature recording. An alternative method would not monitor latency, but could record temperature continuously throughout the test, allowing precise moment-by-moment corrections. However, if the subject's temperature changes rapidly, hysteresis could make this ineffective. In any case, QTRAC recordings would benefit from an increased culture of consideration for temperature effects, while keeping in mind the importance of a solution that is feasible for the high-throughput needs of widespread clinical adoption.

The establishment of this international normative dataset has also allowed for the development of a new computerized clinical decision support system: NerveNorms.Bellstone.ca [3]. NerveNorms allows clinicians to upload test results from a single patient for comparison against the international norms, with automated diagnostics. The system identifies the new data as "Healthy," "Atypical," or "Extremely Atypical" based upon the patient's nerve health score. The website also scores each individual plot, allowing clinicians to quickly and easily investigate atypical NET results.

Future work on NerveNorms could provide additional diagnostic utility, using structural equation analysis or machine learning to identify problems in

Box 1: NerveNorms Examples

The following examples demonstrate some of the ways that nerve excitability tests can be browsed with the NerveNorms.Bellstone.ca website.

- Set the filter to only Canadian data and plot "CA Mean". What is the nerve health score?
- Consider the nerve health score for CA-WI20S. Is this the same as saying he is at the 59th percentile?
- Try filtering by Male, 41–50, Portugal. What problem do you encounter?
- "Rat on Drugs" is a rat anaesthetized by ketamine-xylazine [4]. How does the shape of its Recovery Cycle and Threshold Electrotonus compare to healthy humans? Consider its nerve health score.
- Try uploading a MEM and comparing it to the international norms.

underlying biological factors, such as sodium channel function or myelination. The models can be improved with new data encompassing larger temperature and age ranges. The current version filters norms based on age and sex, but future work could apply automatic temperature correction.

Kawamoto et al. [28] suggest certain factors are important for the success of clinical decision support systems. They found that automated computerized systems are much more effective than those that require referencing charts and tables. Systems that insert recommendations directly into the clinician's workflow are more effective than those that require a separate step. Systems are also more effective when they provide recommendations instead of assessments. NerveNorms.Bellstone.ca automates analysis, but it is not integrated into clinicians' workflow and does not yet provide recommendations. Future work could provide recommendations, directing clinicians to further tests or suggesting diagnoses. Integration into the clinician's workflow would be most effective if QTRAC, the standard recording software, included automated uploads to NerveNorms, allowing in-place calculation of a patient's nerve health

score.

5.5 Conclusions

This study has presented new international norms for the Nerve Excitability Test. Examples of the norms have been presented, showing the similarity between different countries' norms in all plots except the stimulus-response curve, which does not impact other measures. In fact, this difference suggests that the 5% heterogeneity found in Chapter 4 may be due to unimportant technical differences. The release of the combined norms is accompanied by the launch of a clinical decision support system, NerveNorms.Bellstone.ca. NerveNorms allows the upload of a single patient for comparison against the international norms, with the calculation of a nerve health score for automated determination of the patient's nerve health.

Chapter 6

Discussion

In spite of growing interest in the Nerve Excitability Test (NET), there exists no international database of healthy human controls. In part, this was due to prior speculation that racial effects would preclude combination of international data [1]. In order to measure potential site-specific racial or technical differences between datasets, we have developed novel methods for filling missing data and measuring site-specific differences. Both of these methods could also have utility for filling missing values and measuring the homogeneity of other datasets, but that has not yet been tested. The homogeneity measure has shown that international data is 95% homogeneous, supporting the creation of international data. Furthermore, the 5% heterogeneity is likely due to technical differences (specifically skin preparation and placement of recording and stimulating electrodes) rather than true biological differences.

Based on our confidence that it is appropriate to combine the data, we have launched a website, NerveNorms.Bellstone.ca, with the international norms. We have also created an initial nerve health score to provide clinicians with an easy way to determine whether a patient's NET results are healthy. Beyond the potential utility of the novel methods for filling missing data and measuring site-specific differences, this work has implications in the areas of biophysical insight and health services.

6.1 Biophysical Insight

Natural Sciences and Engineering Research Council of Canada (NSERC) supports research that furthers understanding of natural sciences and engineering, including fundamental research related to healthy humans, but not disease-related research [19]. NSERC's focus on biophysical insight was also a major focus for this thesis. The majority of the biophysical insight gleaned from NET comes from the derived excitability variables. In spite of some shortcomings, the current list of excitability variables has been useful in constructing biologically plausible models. Further work could involve the construction of models which can predict underlying biological properties without any plausible mechanism. Such models could provide additional insight into the biophysical changes that occur in peripheral neuropathies. Regardless of the type of model, the derived excitability variables are likely to play a key role.

Most of the past NET studies have primarily focused on the derived excitability variables (a limited selection includes [1, 11, 24, 30, 38]). Some variables are well-established neurophysiological measures (e.g. rheobase [23], SDTC [44], refractory period [5]), but most were developed by Hugh Bostock during the development of the QTRAC software because his formulation of this test protocol was unique. Many of them provide valuable neurobiological insight. For example, TEh(90–100ms) is a proxy measure for I_h (though extended TE provides 70% and 100% S3, which are even better measures of I_h [34]); superexcitability is an indirect measure of myelination, specifically, the capacitance of the internode [29]; and accommodation half-time is sensitive to the activation of slow potassium channels [32]. The strong link between biology and the excitability variables has made them an important part of NET studies.

However, these derived excitability variables could potentially be improved. Many of them are highly correlated (e.g. Rheobase (mA) and Stimulus for 50% max, $\rho = 0.979$; TEd(peak) and TEd(10-20ms), $\rho = 0.960$; Superexcitability at 5 ms and Superexcitability, $\rho = 0.971$), so they could be removed. Others could potentially be added. The ratio of TEh at 109ms and 10ms, or the

ratio of TEh at 109ms to minimum TEh, might be a more effective measure of I_h than TEh(90–100ms). None of the RC measures are sensitive to time delays past 7ms. This reliance on excitability variables is a potential weakness of this thesis and past works. This weakness is evident when the underlying waveforms show something that the variables do not express. For example, in our recent study of fast and slow axons, increased I_h was visually evident in TE (Figure 2.1c), but none of the standard measures were adequate to demonstrate it, so we were forced to use long-term TE (Figure 2.1d). In the case of the detection of site-specific differences in Chapter 4, if site-specific differences are present in ways not measured by the excitability variables, they would be missed by the homogeneity measure. Improvements to these measures could developed by expert users of NET, by machine learning feature selection and dimensionality reduction [42], or some combination of these approaches.

Even though the excitability variables are often correlated with underlying biophysical properties like myelination and potassium channel gating, those relationships are not direct. One approach to this problem is to develop a biologically realistic model of axons, like the Bostock [8], Bostock-Howells [21], and MRG [39] models. These models create a circuit equivalent of the axonal membrane. They have capacitances and conductances which are directly related to biological properties, such as the membrane conductance or the fast potassium conductance. These model parameters can be tweaked to match the model output to a desired NET result, giving putative biological properties for the subject. While these models have the advantage of an understandable and useful equivalent circuit, they are difficult to tune. The Bostock-Howells model is a better qualitative match for the average human, but it cannot express the variation of healthy humans as well as the Bostock model [25, 65]. Future work on these models has the potential to provide incredible insight into the mechanisms of dysfunction in peripheral neuropathies, but at present they do not effectively tie individual NET measurements to the underlying cellular biology.

An alternative to biologically plausible models is structural equation modeling (SEM) or machine learning. SEM is especially suited to the causal analy-

sis of the linear relationships between latent variables (in this case, underlying biophysical properties) and measured variables (in this case, the derived excitability variables). It does not explain *why* the latent and measured variables are correlated, but it can give good estimates of the *how much* the observed variables linearly depend upon the latent variables along with estimated errors [9]. It also allows the model builder to define certain properties of the model, such as indicating which latent and measured variables are expected to be correlated. Its disadvantage is that it assumes linear relationships. Machine learning approaches can be much more powerful and flexible than SEM because they can be non-linear. In particular, generative models could be useful. At present, the biologically plausible models are generative models in the sense that they can generate samples, but they are not classifiers. Furthermore, as discussed above, they are not effective at generating data encompassing the large variation of human norms.

Future work could build generative models based on $p(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is the excitability variables (or the underlying waveforms, potentially after feature extraction). $\mathbf{Y}$ could be a binary healthy/unhealthy class, a disease diagnosis vector, or the target latent biological properties. Since these generative models are explicitly based upon probabilities, they make missing data much easier to handle: the missing variables can simply be integrated out. However, the training of a generative model depends upon some samples which contain both $\mathbf{X}$ and $\mathbf{Y}$. If the goal is to represent the latent biological properties, this $\mathbf{Y}$ is unknown. Future studies could measure some of these biological properties in order to train a generative model (e.g. extracellular K^+), but many of these properties can only be measured invasively and destructively, and others cannot be directly measured at all, making it difficult or impossible to gather this training data. These approaches can be bridged with non-linear SEM [41, 64]. A model between latent and observed variables would allow clinicians to use NET to determine deficiencies in underlying biological properties, even if the exact nature of the relationship is unknown. This could be useful for the development of future nerve health scores and especially clinical decision support system recommendations, as clinicians could be directed toward tests

and treatments directly relevant to the impacted properties.

6.2 Health Services

Canadian Institutes of Health Research (CIHR) defines four pillars of health research: biomedical, clinical, health services, and population health [19]. This project provides normative data which has demonstrable efficacy, but effectiveness has yet to be tested. The normative data has been used in a health service with a simple, yet clinically useful nerve health score. This nerve health score could be further improved with a one-class classifier to support better accuracy and eventually multi-class classification for differential diagnosis.

Normative data is essential for clinical practice. International norms, when possible, are also important because they allow research to be generalized to countries that do not have adequate health care funding. For example, the World Health Organization child growth curves provide valuable guidance to health professionals all around the world [43]. Normative data for NET would save researchers from needing to collect their own independent data from healthy controls. It would also allow clinicians to compare patients' results with known healthy individuals. The normative data collected for this study is not yet adequate ($n = 276$). The addition of the Irish and Australian data would help (resulting in $n = 441$), but further data would allow continued consideration of the effects of age, temperature, sex, race, and technical factors. The effect of biological factors could be further studied by recording participants' race as part of the collection process; this would allow confirmation that biological race is not a differentiating factor. As discussed in Chapter 5, temperature control is still an open question (e.g. is it better to control temperature or to correct it afterward?), so quality control processes should be established prior to further data acquisition. Skin preparation and electrode placement could also result in site specific-differences, so they should be considered in the quality control process. Past human studies have primarily focused on the median nerve, but normative data for other nerves (e.g. common peroneal, as collected locally in Canada) would also be valuable. Once

quality control standards have been resolved, the continued collection of median and other NET normative data will allow for the development of more effective clinical tools.

Clinical decision support systems (CDSS) are most effective when key factors are realized: automatic integration at the time of decision making, involvement of study authors in development, providing recommendations instead of assessments, computer-aided decision making, periodic feedback to clinicians, sharing system recommendations with patients, and documenting deviations from system recommendations [17, 28]. NerveNorms succeeds at some of these factors, but others have yet to be implemented. First of all, the system currently provides automated assessments, but does not have recommendations. Future updates could suggest further tests, especially once differential diagnosis is included. These recommendations could then be provided directly to clinicians and patients. A holistic system could track patient results over time, allowing clinicians to view feedback about outcomes and to enter their reasons for ignoring the recommendations. At present, NerveNorms is a starting point for a health service that could become an effective clinical decision support system.

Clinically relevant work must demonstrate both *efficacy* (performance in controlled trials) and *effectiveness* (real-world utility) [54]. Many previous NET studies have demonstrated efficacy (e.g. the many examples in Kiernan and Lin [30]). The goal of this project was to show the potential effectiveness of NET in a real-world clinical setting by determining whether international data can be used across sites. This work has shown that the expected location-based biological differences cannot be large. In fact, the small differences between sites are more likely technical in nature. The question of the potential impact of site-specific differences in testing methodology has also been answered: such differences are small and relatively inconsequential (e.g. absolute SR). Together, this analysis of biological and technical differences has shown that international data can be combined, demonstrating the potential effectiveness of NET.

The nerve health score is one portion of this work that has focused on

efficacy rather than effectiveness. The nerve health score described in Chapter 5 is capable of differentiating healthy human median nerves from healthy common peroneal nerves and drugged rats, but its effectiveness in a clinical setting has not been demonstrated. The Canadian ALS Neuroimaging Consortium (CALSNIC) has been comparing imaging biomarkers across sites in order to measure the effectiveness of those biomarkers for diagnosis [27]. Thus far, their work has exclusively focused on imaging, so neurophysiologists have not been included. Given the expertise required to interpret NET results, its impact in clinical settings has been limited, but the simplicity of the NerveNorms nerve health score could allow neurophysiologists to participate in CALSNIC's multi-site effectiveness testing.

The current version of the nerve health score is a simple Gaussian model. While this model has the advantage of simplicity, which helps in driving buy-in from clinicians [10, 62], a more advanced model could provide significant advantages. The nerve health score similar to one-class classification: the field of classification in which only a single, known class is available for training, from which the algorithm must learn to differentiate between members of that known class of "inliers" and a potentially infinite number of unknown classes comprising the "outliers" [57]. In this case, of course, the inliers are healthy nerves and all types of peripheral nerve diseases and disorders are outliers. Since it is impossible to identify all types of outliers in advance—some diseases might not have even been identified yet—much less measure a significant number of samples for classification, one-class classification is preferred over multi-class. Multi-class classification could have its place in the future, enabling differential diagnosis once enough diseased samples have been gathered, but one-class classification can provide an effective diagnostic tool with only healthy samples.

Many one-class classification (also known as anomaly detection) algorithms are available [57]. Some of them are sensitive to hyperparameters, which means they would require some example outliers for training. Since that data is not available, it will be important to pick algorithms which are insensitive to hyperparameters, such as k-nearest neighbors Data Description (kNN_{DD}) and Gaus-

sian Data Description ($\text{Gaussian}_{\text{DD}}$) [56]. Preliminary (unpublished) work on these algorithms suggests they perform well on NET data, but a rigorous study of their efficacy has not been undertaken. Example code comparing one-class classification algorithms, using implementations from Tax [59], is available in the repository for this thesis [2].

6.3 Concluding Remarks

The NET can be part of an effective and efficacious clinical decision support system. The creation of an international normative dataset, the development of a nerve health score, and their deployment in the NerveNorms website are the start of such a system. The development of methods for filling missing data in NET and for detecting site-specific differences have provided the necessary foundation to ensure that data from sites around the world can continue to be added to the NerveNorms dataset. Future work can add additonal nerves, species, and disorders, along with advances in modeling, nerve health scores, and clinical recommendations, with the eventual outcome of differential diagnosis of peripheral nerve disorders.

References

[1] Jong Seok Bae, Setsu Sawai, Sonoko Misawa, Kazuaki Kanai, Sagiri Isose, Kazumoto Shibuya, and Satoshi Kuwabara. Effects of age on excitability properties in human motor axons. *Clinical Neurophysiology*, 119(10):2282 – 2286, 2008. ISSN 1388-2457. doi: 10.1016/j.clinph.2008.07.005.

[2] James M. Bell. *Matlab code for my thesis research*. GitHub, Aug 2019. URL `https://github.com/stellentus/healthy-nerves`.

[3] James M. Bell and Matthias Stone. *Source Code for the NerveNorms Website*. GitHub, Aug 2019. URL `https://github.com/stellentus/nerve-norms`.

[4] James M. Bell, Chad Lorenz, and Kelvin E. Jones. Nerve excitability differences in slow and fast motor axons of the rat: more than just Ih. *bioRxiv*, 2019. doi: 10.1101/613984.

[5] MJ Berry and M Meister. Refractoriness and neural precision. *J Neurosci*, 18(6):2200–2211, Mar 1998. ISSN 0270-6474 (Print); 0270-6474 (Linking).

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[7] Todd E Bodner. Missing data: Prevalence and reporting practices. *Psychological Reports*, 99(3):675–680, 2006.

[8] H Bostock, M Baker, and G Reid. Changes in excitability of human motor axons underlying post-ischaemic fasciculations: evidence for two stable states. *The Journal of Physiology*, 441(1):537–557, 2019/07/16 1991. doi: 10.1113/jphysiol.1991.sp018766.

[9] Timothy A. Brown. *Confirmatory Factor Analysis for Applied Research*. Methodology in the Social Sciences. The Guilford Press, second edition, 2015.

[10] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2788613.

[11] I. Casanova, A. Diaz, S. Pinto, and M. de Carvalho. Motor excitability measurements: The influence of gender, body mass index, age and temperature in healthy controls. *Neurophysiologie Clinique/Clinical Neurophysiology*, 44(2):213 – 218, 2014. ISSN 0987-7053. doi: 10.1016/j.neucli.2014.03.002.

[12] Jehanzeb R Cheema. A review of missing data handling methods in education research. *Review of Educational Research*, 84(4):487–508, 2014.

[13] Mo Chen. *Matlab code for machine learning algorithms in book PRML*. GitHub, Jan 2019. URL `https://github.com/PRML/PRMLT`.

[14] Yiran Dong and Chao-Ying Joanne Peng. Principled missing data methods for researchers. *SpringerPlus*, 2(1):222, 2013.

[15] James D Dziura, Lori A Post, Qing Zhao, Zhixuan Fu, and Peter Peduzzi. Strategies for dealing with missing data in clinical trials: from design to analysis. *The Yale Journal of Biology and Medicine*, 86(3):343, 2013.

[16] Abel Folch-Fortuny, Francisco Arteaga, and Alberto Ferrer. PCA model building with missing data: New proposals and a comparative study. *Chemometrics and Intelligent Laboratory Systems*, 146:77 – 88, 2015. ISSN 0169-7439. doi: 10.1016/j.chemolab.2015.05.006.

[17] Amit X. Garg, Neill K. J. Adhikari, Heather McDonald, M. Patricia Rosas-Arellano, P. J. Devereaux, Joseph Beyene, Justina Sam, and R. Brian Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *JAMA*, 293(10):1223–1238, 03 2005. ISSN 0098-7484. doi: 10.1001/jama.293.10.1223.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL `http://www.deeplearningbook.org`.

[19] Science Government of Canada, Innovation, Communications Economic Development Canada, Office of the Deputy Minister, and Marketing Branch. Selecting the appropriate federal granting agency, 12 2016. URL `http://www.science.gc.ca/eic/site/063.nsf/eng/h_FEE7261A.html`.

[20] John W Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009.

[21] James Howells, Louise Trevillion, Hugh Bostock, and David Burke. The voltage dependence of Ih in human myelinated axons. *The Journal of Physiology*, 590(7):1625–1640, 2019/07/16 2012. doi: 10.1113/jphysiol.2011.225573.

[22] Werner Irnich. Georges Weiss' fundamental law of electrostimulation is 100 years old. *Pacing and Clinical Electrophysiology*, 25(2):245–248, 2002. doi: 10.1046/j.1460-9592.2002.00245.x.

[23] Werner Irnich. The terms "chronaxie" and "rheobase" are 100 years old. *Pacing and Clinical Electrophysiology*, 33(4):491–496, 2010. doi: 10.1111/j.1540-8159.2009.02666.x.

[24] S.K. Jankelowitz, P.A. McNulty, and David Burke. Changes in measures of motor axon excitability with age. *Clinical Neurophysiology*, 118(6): 1397 – 1404, 2007. ISSN 1388-2457. doi: 10.1016/j.clinph.2007.02.025.

[25] Karl Jensen, Thu NA Luu, and Kelvin E. Jones. Using axon models to interpret electrodiagnostic nerve tests. *BMC Neuroscience*, 9(1):P43, 2008. doi: 10.1186/1471-2202-9-S1-P43.

[26] Ian T. Jolliffe. *Principal component analysis*. Springer Series in Statistics. Springer, 2002.

[27] Sanjay Kalra, Christian Beaulieu, Michael Benatar, Hannah Briemberg, Annie Dionne, Nicolas Dupre, Dean Eurich, Richard Frayne, Angela Genge, Simon Graham, Christopher Hanstock, Julia Keith, Lawrence Korngut, Christian Shoesmith, Alan Wilman, Yee Hong Yang, Yana Yunusova, and Lorne Zinman. The Canadian ALS Neuroimaging Consortium (CALSNIC) (P1.4-010). *Neurology*, 92(15 Supplement), 2019. ISSN 0028-3878. URL `https://n.neurology.org/content/92/15_Supplement/P1.4-010`.

[28] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*, 330(7494):765, 2005. ISSN 0959-8138. doi: 10.1136/bmj.38398.500764.8F.

[29] Matthew C. Kiernan and Ryuji Kaji. Chapter 4 - physiology and pathophysiology of myelinated nerve fibers. In Gérard Said and Christian Krarup, editors, *Peripheral Nerve Disorders*, volume 115 of *Handbook of Clinical Neurology*, pages 43 – 53. Elsevier, 2013. doi: 10.1016/B978-0-444-52902-2.00004-7.

[30] Matthew C. Kiernan and Cindy Shin Yi Lin. Nerve excitability: A clinical translation. In Michael J. Aminoff, editor, *Aminoff's Electrodiagnosis in Clinical Neurology*, chapter 15, pages 345 – 365. W.B. Saunders, London, sixth edition, 2012. ISBN 978-1-4557-0308-1. doi: 10.1016/B978-1-4557-0308-1.00015-7.

[31] Matthew C. Kiernan, David Burke, Kjeld V. Andersen, and Hugh Bostock. Multiple measures of axonal excitability: A new approach in clinical testing. *Muscle & Nerve*, 23(3):399–409, 2000. doi: 10.1002/(SICI)1097-4598(200003)23:3⟨399::AID-MUS12⟩3.0.CO;2-G.

[32] Matthew C. Kiernan, Katia Cikurel, and Hugh Bostock. Effects of temperature on the excitability properties of human motor axons. *Brain*, 124 (4):816–825, 04 2001. ISSN 0006-8950. doi: 10.1093/brain/124.4.816.

[33] Maria O. Kovalchuk, Hessel Franssen, Féline E.V. Scheijmans, Leonard J. Van Schelven, Leonard H. Van Den Berg, and Boudewijn T.H.M. Sleutjes. Warming nerves for excitability testing. *Muscle & Nerve*, 2019. doi: 10.1002/mus.26621.

[34] C S Lin, I Mogyoros, S Kuwabara, C Cappelen-Smith, and D Burke. Accommodation to depolarizing and hyperpolarizing currents in cutaneous afferents of the human median and sural nerves. *J Physiol*, 529 Pt 2(Pt

2):483–492, Dec 2000. ISSN 0022-3751 (Print); 1469-7793 (Electronic); 0022-3751 (Linking). doi: 10.1111/j.1469-7793.2000.00483.x.

[35] Roderick J Little, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, James D Neaton, Andrea Rotnitzky, Daniel Scharfstein, Weichung J Shih, Jay P Siegel, and Hal Stern. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[36] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.

[37] Chad Lorenz and Kelvin E. Jones. I_H activity is increased in populations of slow versus fast motor axons of the rat. *Frontiers in Human Neuroscience*, 8:766, 2014.

[38] John C. McHugh, Richard B. Reilly, and Sean Connolly. Examining the effects of age, sex, and body mass index on normative median motor nerve excitability measurements. *Clinical Neurophysiology*, 122(10):2081 – 2088, 2011. ISSN 1388-2457. doi: 10.1016/j.clinph.2011.03.020.

[39] Cameron C. McIntyre, Andrew G. Richardson, and Warren M. Grill. Modeling the excitability of mammalian nerve fibers: Influence of afterpotentials on the recovery cycle. *Journal of Neurophysiology*, 87(2): 995–1006, 2002. doi: 10.1152/jn.00353.2001. PMID: 11826063.

[40] Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873 – 895, 2007.

[41] Helfried Moosbrugger, Karin Schermelleh-Engel, Augustin Kelava, and Andreas G. Klein. Testing multiple nonlinear effects in structural equation modeling: A comparison of alternative estimation approaches. In Timothy Theo and Myint Swe Khine, editors, *Structural Equation Modelling in Educational Research: Concepts and Applications*, chapter 6, pages 103–136. Sense Publishers, Rotterdam, 01 2008. ISBN 978-90-8790-787-7.

[42] Lan Huong Nguyen and Susan Holmes. Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, 15(6):1–19, 06 2019. doi: 10.1371/journal.pcbi.1006907.

[43] Chizuru Nishida. Development of a WHO growth reference for school-aged children and adolescents. Technical report, World Health Organisation, 09 2007.

[44] D Noble and RB Stein. The threshold conditions for initiation of action potentials by excitable cells. *The Journal of physiology*, 187(1):129–162, 1966.

[45] Chao-Ying Joanne Peng, Michael Harwell, Show-Mann Liou, Lee H Ehman, et al. Advances in missing data methods and implications for educational research. *Real Data Analysis*, 3178, 2006.

[46] James L Peugh and Craig K Enders. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4):525–556, 2004.

[47] Philip L Roth. Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47(3):537–560, 1994.

[48] Leah H Rubin, Katie Witkiewitz, Justin St Andre, and Steve Reilly. Methods for handling missing data in the behavioral neurosciences: Don?t throw the baby rat out with the bath water. *Journal of Undergraduate Neuroscience Education*, 5(2):A71, 2007.

[49] Gustavo Saposnik, Moira K Kapral, Ying Liu, Ruth Hall, Martin O'Donnell, Stavroula Raptis, Jack V Tu, Muhammad Mamdani, and Peter C Austin. IScore: a risk score to predict death early after hospitalization for an acute ischemic stroke. *Circulation*, 123(7):739–749, Feb 2011. ISSN 1524-4539 (Electronic); 0009-7322 (Linking). doi: 10.1161/CIRCULATIONAHA.110.983353.

[50] Gustavo Saposnik, Stavroula Raptis, Moira K Kapral, Ying Liu, Jack V Tu, Muhammad Mamdani, and Peter C Austin. The iScore predicts poor functional outcomes early after hospitalization for an acute ischemic stroke. *Stroke*, 42(12):3421–3428, Dec 2011. ISSN 1524-4628 (Electronic); 0039-2499 (Linking). doi: 10.1161/STROKEAHA.111.623116.

[51] Gustavo Saposnik, Jiming Fang, Moira K Kapral, Jack V Tu, Muhammad Mamdani, Peter Austin, and S Claiborne Johnston. The iScore predicts effectiveness of thrombolytic therapy for acute ischemic stroke. *Stroke*, 43 (5):1315–1322, May 2012. ISSN 1524-4628 (Electronic); 0039-2499 (Linking). doi: 10.1161/STROKEAHA.111.646265.

[52] Gustavo Saposnik, Robert Cote, Muhammad Mamdani, Stavroula Raptis, Kevin E Thorpe, Jiming Fang, Donald A Redelmeier, and Larry B Goldstein. JURaSSiC: accuracy of clinician vs risk score prediction of ischemic stroke outcomes. *Neurology*, 81(5):448–455, Jul 2013. ISSN 1526-632X (Electronic); 0028-3878 (Linking). doi: 10.1212/WNL.0b013e31829d874e.

[53] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[54] Amit G Singal, Peter D R Higgins, and Akbar K Waljee. A primer on effectiveness and efficacy trials. *Clinical and translational gastroenterology*, 5(1):e45–e45, 01 2014. doi: 10.1038/ctg.2013.13.

[55] Tomlinson Susan, Burke David, Hanna Mike, Koltzenburg Martin, and Bostock Hugh. In vivo assessment of HCN channel current (Ih) in human motor axons. *Muscle & Nerve*, 41(2):247–256, 2010. doi: 10.1002/mus.21482.

[56] Lorne Swersky. A study of unsupervised outlier detection for one-class classification. Master's thesis, University of Alberta, 2018.

[57] Lorne Swersky, Henrique O. Marques, Jörg Sander, Ricardo J. G. B. Campello, and Arthur Zimek. On the evaluation of outlier detection and one-class classification methods. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, Oct 2016. doi: 10.1109/DSAA.2016.8.

[58] W Fraser Symmans, Christos Hatzis, Christos Sotiriou, Fabrice Andre, Florentia Peintinger, Peter Regitnig, Guenter Daxenbichler, Christine Desmedt, Julien Domont, Christian Marth, Suzette Delaloge, Thomas Bauernhofer, Vicente Valero, Daniel J Booser, Gabriel N Hortobagyi, and Lajos Pusztai. Genomic index of sensitivity to endocrine therapy for breast cancer. *J Clin Oncol*, 28(27):4111–4119, Sep 2010. ISSN 1527-7755 (Electronic); 0732-183X (Print); 0732-183X (Linking). doi: 10.1200/JCO.2010.28.4273.

[59] D.M.J. Tax. *Data description toolbox dd_tools 2.0.0: A Matlab toolbox for data description, outlier and novelty detection for PRTools 5.0*, 2013.

[60] Laurie A Tomlinson and Ian B Wilkinson. Does it matter where we measure blood pressure? *British journal of clinical pharmacology*, 74(2): 241–245, 08 2012. doi: 10.1111/j.1365-2125.2012.04203.x.

[61] Saman Vaisipour. *Detecting, correcting, and preventing the batch effects in multi-site data, with a focus on gene expression Microarrays*. PhD thesis, University of Alberta, 01 2007.

[62] Mai-Anh T. Vu, Tülay Adalı, Demba Ba, György Buzsáki, David Carlson, Katherine Heller, Conor Liston, Cynthia Rudin, Vikaas S. Sohal, Alik S. Widge, Helen S. Mayberg, Guillermo Sapiro, and Kafui Dzirasa. A shared vision for machine learning in neuroscience. *Journal of Neuroscience*, 38(7):1601–1607, 2018. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0508-17.2018.

[63] Silke Wagner and Dorothea Wagner. Comparing clusterings - an overview. *Technical Report 2006-04*, 01 2007.

[64] Melanie M. Wall and Yasuo Amemiya. Nonlinear structural equation modeling as a statistical method. In Sik-Yum Lee, editor, *Handbook of Latent Variable and Related Models*, Handbook of Computing and Statistics with Applications, chapter 15, pages 321 – 343. North-Holland, Amsterdam, 2007. doi: 10.1016/B978-044452044-9/50018-5.

[65] Madeline Yee. *Characterizing the variability in nerve excitability test results of human myelinated axons: A computational modelling approach*. Bachelor's thesis, University of Alberta, 2019.

[66] M. D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *ArXiv e-prints*, 2012.